

www.phmsociety.org

ISBN - 978-1-936263-03-5



Proceedings of
**The Annual Conference of the
Prognostics and Health Management
Society 2011**

PHM'11

ISBN - 978-1-936263-03-5

Montreal, Quebec Canada

September 25 - 29, 2011

Edited by:

José R. Celaya,
Sankalita Saha, and
Abhinav Saxena

Table of Contents

Full Papers

A Bayesian Probabilistic Approach to Improved Health Management of Steam Generator Tubes <i>Kaushik Chatterjee and Mohammad Modarres</i>	1
A Combined Anomaly Detection and Failure Prognosis Approach for Estimation of Remaining Useful Life in Energy Storage Devices <i>Marcos E. Orchard, Liang Tang, George J. Vachtsevanos</i>	8
A Mobile Robot Testbed for Prognostics-Enabled Autonomous Decision Making <i>Edward Balaban, Sriram Narasimhan, Matthew Daigle, José R. Celaya, Indranil Roychoudhury, Bhaskar Saha, Sankalita Saha, Kai Goebel</i>	15
A Model-based Prognostics Methodology for Electrolytic Capacitors Based on Electrical Overstress Accelerated Aging <i>José R. Celaya, Chetan Kulkarni, Gautam Biswas, Sankalita Saha, Kai Goebel</i>	31
A Structural Health Monitoring Software Tool for Optimization, Diagnostics and Prognostics <i>Seth S. Kessler, Eric B. Flynn, Christopher T. Dunn, Michael D. Todd</i>	40
A Study on the parameter estimation for crack growth prediction under variable amplitude loading <i>Sang Hyuck Leem, Dawn An, Sangho Ko, Joo-Ho Choi</i>	48
A Testbed for Real-Time Autonomous Vehicle PHM and Contingency Management Applications <i>Liang Tang, Eric Hettler, Bin Zhang, Jonathan DeCastro</i>	56
Adaptive Load-Allocation for Prognosis-Based Risk Management <i>Brian Bole, Liang Tang, Kai Goebel, George Vachtsevanos</i>	67
An Adaptive Particle Filtering-based Framework for Real-time Fault Diagnosis and Failure Prognosis of Environmental Control Systems <i>Ioannis A. Raptis, George J. Vachtsevanos</i>	77
E2GK-pro: An Evidential Evolving Multimodeling Approach for Systems Behavior Prediction <i>Lisa Serir, Emmanuel Ramasso, Nouredine Zerhouni</i>	85
Bayesian fatigue damage and reliability analysis using Laplace approximation and inverse reliability method <i>Xuefei Guan, Jingjing He, Ratneshwar Jha, Yongming Liu</i>	94
Bayesian Software Health Management For Aircraft Guidance, Navigation, and Control <i>Johann Schumann, Timmy Mbaya, Ole Mengshoel</i>	104
Commercialization of Prognostics Systems Leveraging Commercial Off-The-Shelf Instrumentation, Analysis, and Data Base Technologies <i>Preston Johnson</i>	114
Comparison of Fault Detection Techniques for an Ocean Turbine <i>Mustapha Mjit, Pierre-Philippe J. Beaujean, David J. Vendittis</i>	123
Comparison of Parallel and Single Neural Networks in Heart Arrhythmia Detection by Using ECG Signal Analysis <i>Ensieh Sadat Hosseini Rooteh, Youmin Zhang, Zhigang Tian</i>	134
Condition Based Maintenance Optimization for Multi-component Systems Cost Minimization <i>Zhigang Tian, Youmin Zhang, Jialin Cheng</i>	143
Cost Comparison of Maintenance Policies <i>Le Minh Duc, Tan Cher Ming</i>	149
Decision and Fusion for Diagnostics of Mechanical Components <i>Renata Klein, Eduard Rudyk, Eyal Masad</i>	159
Defect source location of a natural defect on a high speed rolling element bearing with Acoustic Emission <i>B Eftekharijad, A. Addali, D Mba</i>	168
Deriving Bayesian Classifiers from Flight Data to Enhance Aircraft Diagnosis Models <i>Daniel L.C. Mack, Gautam Biswas, Xenofon D. Koutsoukos, Dinkar Mylaraswamy, George D. Hadden</i>	175

Design for Fault Analysis Using Multi-partite, Multi-attribute Betweenness Centrality Measures <i>Tsai-Ching Lu, Yilu Zhang, David L. Allen, Mutasim A. Salman</i>	190
Distributed Damage Estimation for Prognostics based on Structural Model Decomposition <i>Matthew Daigle, Anibal Bregon, Indranil Roychoudhury</i>	198
Experimental Polymer Bearing Health Estimation and Test Stand Benchmarking for Wave Energy Converters <i>Michael T. Koopmans, Stephen Meicke, Irem Y. Tumer, Robert Paasch</i>	209
Experiments with Neural Networks as Prognostics Engines for Patient Physiological System Health Management <i>Peter K. Ghavami, Kailash Kapur</i>	222
Exploring the Model Design Space for Battery Health Management <i>Bhaskar Saha, Patrick Quach, Kai Goebel</i>	231
Fault Diagnosis in Automotive Alternator System Utilizing Adaptive Threshold Method <i>Ali Hashemi, Pierluigi Pisu</i>	239
Fault-Tolerant Trajectory Tracking Control of a Quadrotor Helicopter Using Gain-Scheduled PID and Model Reference Adaptive Control <i>Iman Sadeghzadeh, Ankit Mehta, Youmin Zhang, Camille-Alain Rabbath</i>	247
Feature Selection and Categorization to Design Reliable Fault Detection Systems <i>H. Senoussi, B. Chebel-Morello, M. Denai, N. Zerhouni</i>	257
From measurements collection to remaining useful life estimation: defining a diagnostic-prognostic frame for optimal maintenance scheduling of choke valves undergoing erosion <i>Giulio Gola, Bent H. Nystad</i>	267
Gear Health Threshold Setting Based On a Probability of False Alarm <i>Eric Bechhoefer, David He, Paula Dempsey</i>	275
Gearbox Vibration Source Separation by Integration of Time Synchronous Averaged Signals <i>Guicai Zhang, Joshua Isom</i>	282
Health Monitoring of an Auxiliary Power Unit Using a Classification Tree <i>Wlamir O. L. Vianna, Joao P. P. Gomes, Roberto K. H. Galvao, Takashi Yoneyama, Jackson P. Matsuura</i>	293
Identification of Correlated Damage Parameters under Noise and Bias Using Bayesian Inference <i>Dawn An, Joo-Ho Choi, Nam H. Kim</i>	300
Integrating Probabilistic Reasoning and Statistical Quality Control Techniques for Fault Diagnosis in Hybrid Domains <i>Brian Ricks, Craig Harrison, Ole Mengshoel</i>	310
Investigating the Effect of Damage Progression Model Choice on Prognostics Performance <i>Matthew Daigle, Indranil Roychoudhury, Sriram Narasimhan, Sankalita Saha, Bhaskar Saha, Kai Goebel</i>	323
Investigation on the opportunity to introduce prognostic techniques in railways axles maintenance <i>Mattia Vismara</i>	334
Lithium-ion Battery State of Health Estimation Using Ah-V Characterization <i>Daniel Le, Xidong Tang</i>	361
Model-Based Prognostics Under Non-stationary Operating Conditions <i>Matej Gašperin, Pavle Bošković, Dani Juričić</i>	368
Modeling wave propagation in Sandwich Composite Plates for Structural Health Monitoring <i>V. N. Smelyanskiy, V. Hafiychuk, D. G. Luchinsky, J. Miller, C. Banks, R. Tyson</i>	375
Online Estimation of Lithium-Ion Battery State-of-Charge and Capacity with a Multiscale Filtering Technique <i>Chao Hu, Byeng D. Youn, Jaesik Chung, Taejin Kim</i>	385
Optimization of fatigue maintenance strategies based on prognosis results <i>Yibing Xiang, Yongming Liu</i>	398
Physics Based Prognostic Health Management for Thermal Barrier Coating System <i>Amar Kumar, Bhavaya Saxena, Alka Srivastava, Alok Goel</i>	409
Physics based Prognostics of Solder Joints in Avionics <i>Avisekh Banerjee, Ashok K. Koul, Amar Kumar, Nishith Goel</i>	419

Point processes for bearing fault detection under non-stationary operating conditions <i>Pavle Boškovski, Dani Juričić</i>	427
Power Curve Analytic for Wind Turbine Performance Monitoring and Prognostics <i>Onder Uluyol, Girija Parthasarathy, Wendy Foslien, Kyusung Kim</i>	435
Prognostics of Power MOSFETs under Thermal Stress Accelerated Aging using Data-Driven and Model-Based Methodologies <i>José R. Celaya, Abhinav Saxena, Sankalita Saha, Kai Goebel</i>	443
Structural Integrity Assessment Using In-Situ Acoustic Emission Monitoring <i>Masoud Rabiei, Mohammad Modarres, Paul Hoffman</i>	453
Study on MEMS board-level package reliability under high-G impact <i>Jiuzheng Cui, Bo Sun, Qiang Feng, ShengKui Zeng</i>	463
Symbolic Dynamics and Analysis of Time Series Data for Diagnostics of a dc-dc Forward Converter <i>Gregory Bower, Jeffrey Mayer, Karl Reichard</i>	469
Using the Validated FMEA to Update Trouble Shooting Manuals: a Case Study of APU TSM Revision <i>Chunsheng Yang, Sylvain Létourneau, Marvin Zaluski</i>	479
Utilizing Dynamic Fuel Pressure Sensor For Detecting Bearing Spalling and Gear Pump Failure Modes in Cummins Pressure Time (PT) Pumps <i>J. Scott Pflumm, Jeff C. Banks</i>	490
 Poster Papers	
A Discussion of the Prognostics and Health Management Aspects of Embedded Condition Monitoring Systems <i>Roger I. Grosvenor, Paul W. Prickett</i>	502
***(no license)A new method of bearing fault diagnostics in complex rotating machines using multi-sensor mixed hidden Markov models <i>Z. S. Chen, Y. M. Yang, Z. Hu, Z. X. Ge</i>	510
A Prognostic Health Management Based Framework for Fault-Tolerant Control <i>Douglas W. Brown, George J. Vachtsevanos</i>	516
Damage Identification in Frame Structures, Using Damage Index, Based on H2-Norm <i>Mahdi Saffari, Ramin Sedaghati, Ion Stiharu</i>	527
Enhanced Multivariate Based Approach for SHM Using Hilbert Transform <i>Rafik Hajrya, Nazih Mechbal, Michel Vergé</i>	532
Fault Detection in Non Gaussian Problems Using Statistical Analysis and Variable Selection <i>João P. P. Gomes, Bruno P. Leão, Roberto K. H. Galvão, Takashi Yoneyama</i>	540
Fleet-wide health Management Architecture <i>***(no license, other text there)Maxime Monnin, Alexandre Voisin, Jean-Baptiste Léger, Benoit Iung</i>	547
Improving data-driven prognostics by assessing predictability of features <i>Kamran Javed, Rafael Gouriveau, Ryad Zemouri, Noureddine Zerhouni</i>	555
Integrated Robust Fault Detection, Diagnosis and Reconfigurable Control System with Actuator Saturation <i>Jinhua Fan, Youmin Zhang, Zhiqiang Zheng</i>	561
Multiple Fault Diagnostic Strategy for Redundant System <i>Yang Peng, Qiu Jing, Liu GuanJun, Lv Kehong</i>	572
Online Abnormality Diagnosis for real-time Implementation on Turbofan Engines and Test Cells <i>Jérôme Lacaille, Valerio Gerez</i>	579
Proficy Advanced Analytics: a Case Study for Real World PHM Application in Energy <i>Subrat Nanda, Xiaohui Hu</i>	588
 Author Index	 595

A Bayesian Probabilistic Approach to Improved Health Management of Steam Generator Tubes

Kaushik Chatterjee and Mohammad Modarres

Center for Risk and Reliability, University of Maryland, College Park, MD, 20742, USA

kaushikc@umd.edu
modarres@umd.edu

ABSTRACT

Steam generator tube integrity is critical for the safety and operability of pressurized water reactors. Any degradation and rupture of tubes can have catastrophic consequences, e.g., release of radioactivity into the atmosphere. Given the risk significance of steam generator tube ruptures, it is necessary to periodically inspect the tubes using nondestructive evaluation methods to detect and characterize unknown existing defects. To make accurate estimates of defect size and density, it is essential that detection uncertainty and measurement errors associated with nondestructive evaluation methods are characterized properly and accounted for in the evaluation. In this paper we propose a Bayesian approach that updates prior knowledge of defect size and density with nondestructive evaluation data, accounting for detection uncertainty and measurement errors. An example application of the proposed approach is then demonstrated for estimating defect size and density in steam generator tubes using eddy current evaluation data. The proposed Bayesian probabilistic approach helps improve health management of steam generator tubes, thereby enhancing the overall safety and operability of pressurized water reactors.

1. INTRODUCTION

Pressurized water reactors (PWR) use heat produced from nuclear fission in the reactor core to generate electricity. In the process of generating electricity, steam generators (SG) play an important role by keeping the reactor core at a safe temperature and acting as the primary barrier between radioactive and non-radioactive sides of a nuclear power plant. Since SG tubes play such an important role, any degradation and rupture in the tubes can be catastrophic (Chatterjee & Modarres, 2011). According to the US Nuclear Regulatory Commission (2010), there have been 10 steam generator tube rupture (SGTR) occurrences in the US

between 1975 and 2000. One such incident occurred in the North Anna power station in 1987 when the plant reached its 100% capacity (US Nuclear Regulatory Commission, 1988). The cause of tube rupture was found to be fatigue, caused by combination of alternating stresses resulting from flow-induced tube vibration and flaws resulting from denting of tubes at support plates.

Given the risk significance of SGTRs, it is absolutely necessary to periodically inspect the tubes using nondestructive evaluation methods in order to detect and quantify the severity of unknown existing defects.¹ All nondestructive evaluation methods have detection uncertainty and measurement errors associated with them that are a result of test equipment complexity, defect attributes, as well as human error. These uncertainties and errors need to be characterized properly and accounted for while estimating the size and density of defects.

A defect of a given size might be detected only a certain percentage of the time (out of total attempts during nondestructive testing) depending on factors such as, noise level, test probe sensitivity, test equipment repeatability and human error. Hence, a defect has an associated probability of detection, which can be defined as the probability the inspection will detect the defect of true size, a , and is denoted by $POD(a)$ (Kurtz, Heasler, & Anderson, 1992). The data from which POD curves are generated can be categorized into two types: qualitative data, i.e., hit/miss; and quantitative data, i.e., signal response amplitude (\hat{a} vs. a), where \hat{a} is signal response. The hit/miss data type is based on a binary process, i.e., whether a defect is detected or not detected. The POD for this data type is calculated as the ratio of the number of successful detection over the total number of inspections performed for a particular defect size, and is called the averaged POD . Hit/miss data are obtained from test equipments such as Sonic IR, and are very subjective in nature depending on operator experience (Li & Meeker, 2008), which induces uncertainty in the values of the POD . A logistic function is

Chatterjee, K. et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

¹ In this paper defect may indicate a crack, flaw, pit, or any other degradation in a structural component. Size may refer to either through-wall depth or surface length of a defect, unless specified. Density refers to number of defects observed per unit volume.

found to best-fit hit/miss data for modeling POD (Jenson, Mahaut, Calmon & Poidevin, 2010).

The other type of POD data is more continuous in nature and is a measure of the amplitude of signal response recorded by the nondestructive test equipment, e.g., ultrasonic or eddy current. In the signal response data-based POD estimation method, the most important parameters are the inspection threshold (noise level) and the decision threshold. The inspection threshold is chosen to account for the noise indications by test equipment, and responses above this threshold are considered for detection/non-detection decisions. Decision threshold is often based on previous field inspections and knowledge of the noise distribution, laboratory experience, and operator experience. The POD curve for signal response data type is modeled using a cumulative log-normal distribution function (Department of Defense, 1999; Jenson, et al., 2010), by determining the cumulative probability of responses (defect signals) greater than the decision threshold. The selection of decision threshold also determines the probability of false call (or false positive).² Hence, there is lot of uncertainty associated with the values chosen for inspection and decision threshold, which lead to uncertainties in the values of the POD. In some cases, the signal response data is also converted into hit/miss data (Jenson et al., 2010) by using the decision threshold and averaged POD values are estimated, which are then fitted into a logistic function.

The precision and accuracy of nondestructive test equipment, and also the techniques used to analyze and process the test results can contribute to measurement errors. For example, large volume of sensor data (such as ultrasound or digital images) are filtered, smoothed, reduced, and censored into another form by subjectively accounting for only certain features of the data. Also, often measurement models are used to convert the form of a measured or observed data into the corresponding value of the reality of interest (i.e., defect size). Uncertainties associated with data processing, model selection and human error can contribute to measurement errors. Measurement error is defined as the difference between the measured and the true value of a defect size. There are two components of measurement error: systematic (bias) error and random (stochastic) error (Jaech, 1964; Hofmann, 2005). Systematic error or bias is a consistent and permanent deflection in the same direction from the true value (Hofmann, 2005). Systematic error (bias) may indicate overestimation (positive bias) or underestimation (negative bias). In most nondestructive measurements, small defects are oversized and large defects are undersized (Kurtz et al., 1992; Wang & Meeker, 2005). Random error arises due to the scattering

² A nondestructive test equipment response interpreted as having detected a flaw but associated with no known flaw at the inspection location (Department of Defense, 1999).

or random variation in measured values (measurement uncertainty).

In the past, there have been efforts to model defect severity in structural components considering nondestructive evaluation uncertainties. Celeux, Persoz, Wandji, and Perrot (1999) describe a method to model defects in PWR vessels considering the POD and random error in measurements. Yuan, Mao, and Pandey (2009) followed the idea of Celeux et al. (1999), to propose a probabilistic model for pitting corrosion in SG tubes considering the POD and random error of the eddy current measurements. However, both Celeux et al. (1999) and Yuan et al. (2009) did not consider the effect of systematic error or bias in measured defect sizes. Also, the POD has not been adjusted for measurement errors in their models. Further, they did not consider uncertainties in the values of the POD, which can affect the defect severity estimates considerably.

This paper addresses some of the shortcomings of existing literature and develops a Bayesian probabilistic approach for modeling defect severity (size and density) in structural components considering the detection uncertainty (i.e., POD and associated uncertainty) and measurement errors (and associated uncertainty) associated with nondestructive evaluation methods. The paper then presents example application of the proposed approach for estimating defect severity in SG tubes using eddy current evaluation data.

2. PROPOSED BAYESIAN APPROACH

The proposed Bayesian approach updates prior knowledge of defect size and density with nondestructive evaluation data, considering the POD, measurement errors (systematic and random), and associated uncertainties, to infer the posterior distributions of defect size and density. The combined effect of POD, measurement errors, and associated uncertainties on measured defect sizes is captured by a likelihood function. In this section, models for measurement errors and POD function will be first defined; then the defect severity models will be presented, followed by the likelihood functions and Bayesian inference equations.

The analysis of measurement error is based on assessing the deviation of the measured defect size from the actual or true defect size, as shown in Eq. (1):

$$E_m = a^* - a \quad (1)$$

where, E_m is the measurement error, a^* is measured and a is the true defect size. Generally a linear regression relationship of the form shown in Eq. (2) is used to model measurement error (Kurtz et al., 1992; Jaech, 1964).

$$a^* = ma + c + \varepsilon(0, \sigma_a) \quad (2)$$

where, m and c are regression coefficients obtained through a regression analysis of a^* vs. a , and ε is the random error in measurement (scattering of the data), which is assumed to follow a normal distribution with mean zero and standard deviation σ_a (function of defect size). The regression coefficients (m & c) are jointly measure of systematic error or bias in measurements. Distributions of bias parameters represent epistemic uncertainty in the chosen measurement error model. From Eqs. (1) and (2), the measurement error can be expressed as:

$$E_m = \overbrace{(m-1)a + c}^{\text{Bias}} + \overbrace{\varepsilon(0, \sigma_a)}^{\text{Random error}} \quad (3)$$

Measurement error can then be expressed as a function of measured defect size using Eqs. (1) and (3) as:

$$E_m = \left(\frac{m-1}{m} \right) a^* + \frac{c}{m} + \frac{\varepsilon(0, \sigma_a)}{m} \quad (4)$$

The probability density function (PDF) of the measurement error can then be defined using a normal distribution with mean as the bias, B_a , standard deviation as that of random error, $\frac{\sigma_a}{m}$, and measurement error as random variable.

$$g(E_m) = N\left(B_a, \frac{\sigma_a}{m}\right) \quad (5)$$

Assume that true defect size, a , is treated as random variable with the PDF, $f(a|\psi)$, where ψ is the vector of the PDF parameters. Defect size PDF considering measurement error can then be expressed as shown in Eq. (6).

$$f(a|\psi) = \int_{E_m} f((a^* - E_m)|\psi) g(E_m) dE_m \quad (6)$$

All the defects in a structure are not detected during nondestructive testing. The detection of a defect depends on its size and is represented by the POD curve. The POD of a defect of size, a , can be represented by a function as shown in Eq. (7):

$$POD(a|\theta, \sigma_{POD}) = h(a, \theta, a_{th}) + \varepsilon_{POD}(0, \sigma_{POD}) \quad (7)$$

where, $h(a, \theta, a_{th})$ is the POD function, a_{th} is the detection threshold, θ is vector of parameters of the POD function, and ε_{POD} is the random error, which represents uncertainty in the POD data and is assumed to follow a normal distribution with mean zero and standard deviation σ_{POD} (function of true defect size). The POD function is selected based on the type of data, e.g., hit/miss or signal response as discussed in Section 1. Joint distribution of the parameters of the POD function, $k(\theta)$, represents the epistemic uncertainty associated with the choice of the POD function.

The marginal POD independent of random variables, θ and σ_{POD} , can be expressed as shown in Eq. (8), where, $m(\sigma_{POD})$ represents the PDF of random variable, σ_{POD} .

$$POD(a) = \int_{\sigma_{POD}} \int_{\theta} POD(a|\theta, \sigma_{POD}) k(\theta) m(\sigma_{POD}) d\theta d\sigma_{POD} \quad (8)$$

The likelihood function for detecting defect of true size, a , given that the defect is detected ($D = 1$), can then be expressed as shown in Eq. (9) (Celeux et al., 1999):

$$L(a|D = 1) = \frac{f(a|\psi) \times POD(a)}{P_d(\psi)} \quad (9)$$

where, $P_d(\psi)$ is the marginal POD that is a function of defect size distribution parameters only (independent of defect size), and can be expressed as:

$$P_d(\psi) = \Pr(D = 1) = \int_0^{\infty} POD(a) f(a|\psi) da \quad (10)$$

During nondestructive measurements true defect sizes are unknown, while the only known quantities are the measured defect sizes and number of detections. The likelihood function of true defect sizes corresponding to measurements consisting of n_e^* exact defect sizes (using Eq. 9) considering measurement errors can be represented as:

$$L(a_{\text{exact}}|\psi) = \frac{1}{[P_d(\psi)]^{n_e^*}} \prod_{i=1}^{n_e^*} \int_{E_m} POD(a_i^* - E_m) f((a_i^* - E_m)|\psi) g(E_m) dE_m \quad (11)$$

Nondestructive measurements are in most cases interval or left censored, in which case the likelihood function of true defect sizes corresponding to measurements consisting of $n_{\text{int},j}^*$ defects within the j th interval (or in a left censored interval) (Cook, Duckworth, Kaiser, Meeker & Stephenson, 2003), can be expressed as shown in Eq. (12).

$$L_j(a_{\text{int}}|\psi) = \left[\frac{1}{P_d(\psi)} \int_{a_{j-1}^*}^{a_j^*} \int_{E_m} POD(a^* - E_m) f((a^* - E_m)|\psi) g(E_m) dE_m da^* \right]^{n_{\text{int},j}} \quad (12)$$

Therefore, the likelihood function of true defect sizes corresponding to total measurements consisting of m defect size intervals each with certain number of defects ($n_{\text{int},j}^*$ in j th interval), and n_e^* exact defect sizes can then be expressed as shown in Eq. (13).

$$L(a|\psi) = \prod_{j=1}^m \left[\frac{1}{P_d(\psi)} \int_{a_{j-1}^*}^{a_j^*} \int_{E_m} POD(a^* - E_m) f((a^* - E_m)|\psi) g(E_m) dE_m da^* \right]^{n_{\text{int},j}} \times \frac{1}{[P_d(\psi)]^{n_e^*}} \prod_{i=1}^{n_e^*} \int_{E_m} POD(a_i^* - E_m) f((a_i^* - E_m)|\psi) g(E_m) dE_m \quad (13)$$

The posterior defect size distribution parameters can then be estimated using Bayesian inference as:

$$\pi_1(\psi|Data) = \frac{L(Data|\psi)\pi_0(\psi)}{\int_{\psi} L(Data|\psi)\pi_0(\psi)d\psi} \quad (14)$$

where, $\pi_1(\psi|Data)$ is posterior distribution of defect size parameters and $\pi_0(\psi)$ is prior distribution of the parameters. The posterior defect size parameters obtained from Bayesian inference can then be used to estimate the corresponding marginal POD values (Eq. 10).

The likelihood of observing n^* ($= n_e^* + \sum_{j=1}^m n_{int,j}^*$) number of defects given n actual number of defects can be expressed by a binomial function (detection process is binary, i.e., either detection or no detection), as shown by Eq. (15):

$$L(n^*|n) = \binom{n}{n^*} [P_d(\psi)]^{n^*} [1 - P_d(\psi)]^{n-n^*} \quad (15)$$

where, $P_d(\psi)$ is the marginal POD value corresponding to posterior defect size parameters. In Eq. (15), the actual number of defects, n , is unknown whereas n^* and $P_d(\psi)$ are known. The actual number of defects can be estimated using Bayesian inference as shown in Eq. (16):

$$\pi_1(n|n^*) = \frac{L(n^*|n)\pi_0(n)}{\sum_n L(n^*|n)\pi_0(n)} \quad (16)$$

where, $\pi_1(n|n^*)$ is posterior distribution of actual number of defects given the observation, n^* , and $\pi_0(n)$ is the prior distribution of number of defects. The prior distribution of number of defects can be estimated from a Poisson function, which gives the likelihood of observing n total number of defects in a volume V , given prior defect density ρ as shown in Eq. (17). Here Poisson distribution is used because defects are assumed to occur with the same average intensity and independent of each other.

$$\pi_0(n) = e^{-\rho V} \frac{(\rho V)^n}{n!} \quad (17)$$

The posterior distribution of actual number of defects (Eq. 16) can then be used to obtain the posterior defect density. The standard conjugate prior employed for Poisson distribution likelihood (Eq. 17) is a two-parameter gamma distribution (Simonen, Doctor, Schuster, & Heasler, 2003), in which case the posterior has the same functional form as the gamma distribution. Assume that prior distribution of defect density is:

$$\pi_0(\rho) = \text{gamma}(\rho|\alpha_1, \alpha_2) \quad (18)$$

where, α_1 and α_2 are parameters of gamma distribution. Then the posterior distribution of defect density can be expressed as shown in Eq. (19).

$$\pi_1(\rho) = \text{gamma}(\rho|V + \alpha_1, n + \alpha_2) \quad (19)$$

A MATLAB routine was developed to implement this entire Bayesian approach for estimating defect severity in structural components. The proposed Bayesian approach considers systematic (bias) and random error in nondestructive measurements; suitably adjusts measurement errors in POD; considers uncertainty in POD values; incorporates prior knowledge of defect size and density; provides a framework for updating probability distributions of defect model parameters when new data become available; and is applicable to exact, interval, and censored measurements.

3. APPLICATION OF PROPOSED BAYESIAN APPROACH TO EDDY CURRENT DATA

An example application of the proposed Bayesian approach is presented in this section for estimating flaw severity in SG tubes using eddy current measurements of flaw sizes (through-wall depth). In this section, we first model POD and measurement error for eddy current evaluation using available data from literature, and then use the proposed Bayesian approach to estimate the posterior distributions of flaw size and density.

The eddy current measurement error is assessed in this paper by a Bayesian regression approach (Azarkhail & Modarres, 2007) in light of available data from literature (Kurtz, Clark, Bradley, Bowen, Doctor, Ferris & Simonen, 1990). The regression result is illustrated by Figure 1 with the 50% regression line representing the bias corresponding to mean values of the parameters m and c of Eq. (3). The 95% uncertainty bounds of Figure 1 corresponds to the random error with a constant standard deviation, σ . The parameters m , c and σ obtained through Bayesian regression were then used in Eq. (5) to estimate the PDF of measurement error as a function of measured flaw size.

In order to derive the POD model, it was assumed in this paper that eddy current signal response data were converted into equivalent hit/miss. The POD curve can then be expressed by a logistic function of the form as shown in Eq. (20) (Yuan et al., 2009):

$$POD(a|\beta_1, \beta_2, a_{th}) = \begin{cases} 1 - \frac{1+e^{-\beta_1\beta_2}}{1+e^{\beta_1(a-\beta_2-a_{th})}} + \varepsilon_{POD}(0, \sigma_{POD}) & \text{for } a > a_{th} \\ 0 & \text{otherwise} \end{cases} \quad (20)$$

where, a is flaw size, a_{th} is threshold size for detection, β_1 and β_2 are logistic function parameters, and ε_{POD} is the random error, which is assumed to follow a normal distribution with mean zero and standard deviation σ_{POD} . A

flaw of size less than detection threshold will not be detected. Distributions of the POD model parameters β_1 , β_2 , and σ_{POD} were estimated using Bayesian regression approach in light of POD data available from literature (Kurtz et al., 1992). Figure 2 illustrates a sample logistic function curve and associated uncertainties fitted on POD data through Bayesian regression, with $a_{th} = 0$.

Flaws in nuclear reactor vessel and piping are in most cases best fitted with an exponential distribution, with smaller size flaws having higher probability density and larger size flaws having lower probability density. Here we define the PDF of random variable a , i.e., true flow size in SG tubes, assuming exponential distribution as:

$$f(a|\lambda) = \lambda e^{-\lambda a} \quad (21)$$

where, λ is flaw size intensity. Flaw size distribution considering measurement errors can then be expressed as shown in Eq. (22).

$$f(a|\lambda) = \int_{E_m} \lambda e^{-\lambda(a^* - E_m)} g(E_m) dE_m \quad (22)$$

Eddy current measurements for SG tubes (Dvorsek & Cizelj, 1993) used in our paper to demonstrate the application of the proposed Bayesian approach, were left and interval censored. The likelihood function of true flaw sizes corresponding to eddy current measurements was defined using Eq. (13), with measurement error limits set as -1 and 1 (to cover the extremes of bias and random error). The Bayesian posterior inference of the flaw size intensity was carried out using the MATLAB routine (Section 2). Prior distribution for flaw size intensity was generated using available data from literature (Liao & Guentay, 2009). Figure 3 illustrates the posterior and prior flaw size intensity distributions. Flaw size intensity values were sampled from the posterior distribution (Figure 3), and the corresponding marginal POD values, $P_d(\psi)$, were estimated (Eq. 10).

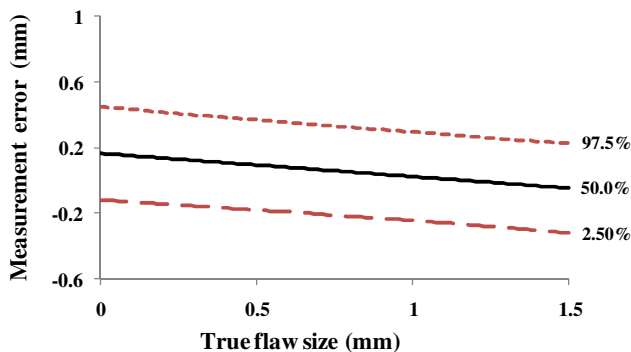


Figure 1. Measurement error and uncertainty bounds (95%)

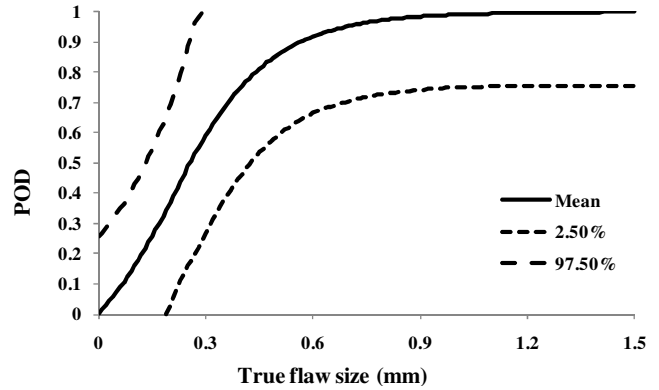


Figure 2. POD curve and uncertainty bounds (95%)

The likelihood function of observed number of flaws was then defined using Eq. (15), and the Bayesian posterior inference of the actual number of flaws (Eq. 16) computed. The prior flaw density distribution used to obtain prior information on number of flaws (Eq. 17) was obtained from the available data in the literature (Liao & Guentay, 2009). Figure 4 illustrates the distribution of actual number of flaws for mean, 2.5% and 97.5% values of posterior flaw size intensity. Posterior distribution of flaw density was then estimated using Eq. (19) for a given volume corresponding to the tube-support plate 9. Figure 5 presents a box and whisker plot showing the distribution of actual number of flaws at tube support plate 9 for different flaw size intervals.

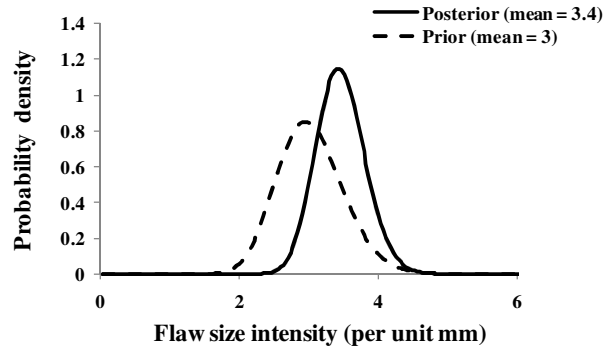


Figure 3. Posterior and prior flaw size intensity

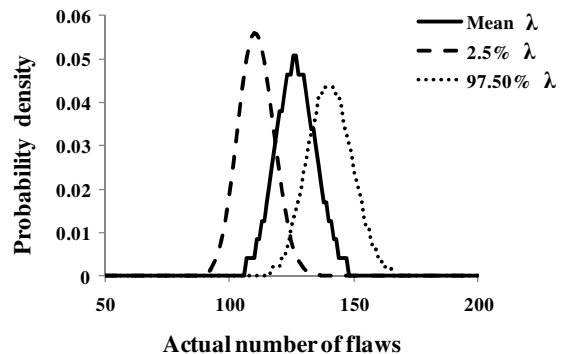


Figure 4. Distributions of actual number of flaws

A comparison between the eddy current measurements and mean of estimated actual number of flaws is presented in Table 1 for different flaw size intervals. It is evident from Table 1 that nondestructive evaluation methods cannot detect and measure all the defects existing in a structure due to associated detection uncertainty and measurement errors. In Table 1, the mean number of flaws estimated using the proposed Bayesian approach (column 3) after considering all uncertainties and prior information, is substantially higher than eddy current measurements (column 2), especially for very small sizes.

As illustrated by the example application, it is critical to consider detection uncertainty and measurement errors associated with nondestructive evaluation methods, in order to estimate the actual defect size and density distributions in critical structures. This is important because the defect size and density distributions estimated during in-service inspections can help in making appropriate and timely replacement/repair decisions, thereby preventing unanticipated failures.

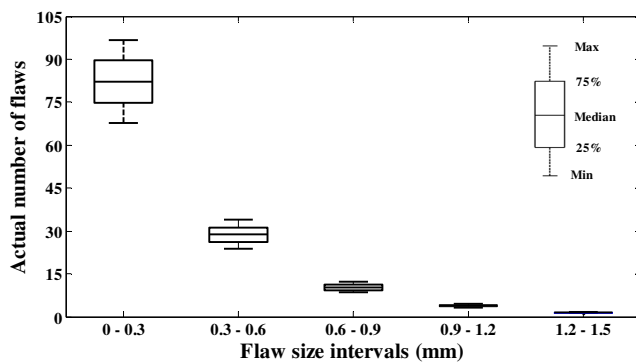


Figure 5. Box and whisker plot of actual number of flaws by size intervals at support plate 9

Flaw size intervals (mm)	Measured # of flaws from eddy current inspection (support plate 9)	Mean actual # of flaws using our Bayesian approach (accounting for all uncertainties)
$a < 0.3$	40	81
$0.3 \leq a < 0.6$	3	29
$0.6 \leq a < 0.9$	3	11
$0.9 \leq a < 1.2$	6	4
$1.2 \leq a < 1.5$	8	1

Table 1. Measured vs. actual number of flaws

4. CONCLUSIONS

It is imperative to assess the health condition of SG tubes periodically during their operating life in order to prevent the occurrence of SGTR failures. Estimating defect size and density in SG tubes require appropriate methods to account for all uncertainties associated with nondestructive evaluation methods. This paper presents a Bayesian approach for estimating defect size and density in structural

components considering detection uncertainty and measurement errors. The proposed Bayesian approach updates prior knowledge of defect size and density with nondestructive evaluation data, considering the POD, measurement errors, and associated uncertainties, to give the posterior distributions of defect size and density. The proposed approach considers both systematic and random error in nondestructive measurements, suitably adjusts measurement errors in POD, considers uncertainties in POD values, and captures the combined effect of POD and measurement errors (including associated uncertainties) on measured defect sizes by a likelihood function. The approach is applicable to exact, interval, and censored measurements; and also provides a framework for updating defect model parameter distribution as and when new information becomes available. An application of this proposed approach is demonstrated for estimating defect size and density in SG tubes using eddy current nondestructive evaluation data. This developed Bayesian probabilistic approach not only fills a critical gap in health management and prognosis of SG tubes, but can also help improve reliability of safety-critical structures in a broad range of application areas, including medical, avionics, and nuclear.

REFERENCES

- Azarkhail, M., & Modarres, M. (2007). A novel Bayesian framework for uncertainty management in physics-based reliability models. *Proceedings of ASME International Mechanical Engineering Congress and Exposition*, November 11-15, Seattle, WA. doi:10.1115/IMECE2007-41333
- Celex, G., Persoz, M., Wandji, J.N., & Perrot, F. (1999). Using Markov Chain Monte Carlo methods to solve full Bayesian modeling of PWR vessel flaw distributions. *Reliability Engineering and System Safety*. vol. 66(3), pp. 243–252. doi:10.1016/S0951-8320(99)00041-1
- Chatterjee, K., & Modarres, M. (2011). A probabilistic physics-of-failure approach to prediction of steam generator tube rupture frequency. *Proceedings of International Topical Meeting on Probabilistic Safety Assessment and Analysis*, March 13-17, Wilmington, NC.
- Cook, D., Duckworth, W.M., Kaiser, M.S., Meeker W.Q., & Stephenson, W.R. (2003). *Principles of maximum likelihood estimation and the analysis of censored data*. Retrieved from Iowa State University website: http://www.public.iastate.edu/~stat415/meeker/ml_estimation_chapter.pdf
- Department of Defense (1999). *Nondestructive evaluation system reliability assessment*. (DoD Publication No. MIL-HDBK-1823). Retrieved from http://www.barringer1.com/mil_files/MIL-HDBK-1823.pdf

- Dvorsek, T., & Cizelj, L. (1993). An analysis of in-service inspection data at tube support plates of KRSKO steam generators. *Proceedings of Regional Meeting: Nuclear Energy in Central Europe*, June 13-16, Portoroz, Slovenia. www.djs.si/proc/bled1997/1191.pdf
- Hofmann, D. (2005). Common sources of errors in measurement systems. In Sydenham, P. & Thorn, R. (Eds.), *Handbook of Measuring System Design* (pp. 289-294). doi:10.1002/0471497398
- Hovey, P., Meeker, W.Q., & Li, M. (2008). Joint estimation of the flaw-size distribution and POD function. *Proceedings of the Review of Progress in Quantitative Nondestructive Evaluation*, July 20-25, Chicago, Illinois. doi:10.1063/1.3114181
- Jaech, J.L. (1964). A Program to Estimate Measurement Error in Nondestructive Evaluation of Reactor Fuel Element Quality. *Technometrics*, vol. 6(3), pp. 293-300.
- Jenson, F., Mahaut, S., Calmon P., & Poidevin, C. (2010). Simulation based POD evaluation of NDI techniques. *Proceedings of 10th European Conference on Non-Destructive Testing*, June 7-11, Moscow, Russia.
- Kurtz, R.J., Clark, R.A., Bradley, E.R., Bowen, W.M., Doctor, P.G., Ferris, R.H., & Simonen, F.A. (1990). *Steam Generator Tube Integrity Program/Steam Generator Group Project*. (NRC Publication No. NUREG/CR-5117), Retrieved from <http://www.nrc.gov/reading-rm/doc-collections/nuregs/contract/cr5117/cr5117.pdf>
- Kurtz, R.J., Heasler, P.G., & Anderson, C.M. (1992). Performance demonstration requirements for eddy current steam generator tube inspection. *Proceedings of 20th Water Reactor Safety Information Meeting*, October 21-23, Bethesda, MD.
- Li, M., & Meeker, W.Q. (2008). A noise interference model for estimating probability of detection for nondestructive evaluations. *Proceedings of the Review of Progress in Quantitative Nondestructive Evaluation*, July 20-25, Chicago, Illinois. doi:10.1063/1.3114172
- Liao, Y., & Guentay, S. (2009). Potential Steam Generator Tube Rupture in the Presence of Severe Accident Thermal Challenge and Tube Flaws Due to Foreign Object Wear. *Nuclear Engineering and Design*, vol. 239(6), pp. 1128-1135. doi:10.1016/j.nucengdes.2009.02.003
- Simonen, F.A., Doctor, S.R., Schuster, G.J. & Heasler, P.G. (2003). *A generalized procedure for generating flaw related inputs for the FAVOR code*, (NRC Publication No. NUREG/CR-6817), Retrieved from <http://pbdupws.nrc.gov/docs/ML0408/ML040830499.pdf>
- US Nuclear Regulatory Commission (1988). *Rapidly propagating fatigue cracks in steam generator tubes*. (NRC Publication No. 88-02). Retrieved from <http://www.nrc.gov/reading-rm/doc-collections/gen-comm/bulletins/1988/bl88002.html>
- US Nuclear Regulatory Commission (2010). *Resolution of generic safety issues: Issue 188: Steam generator tube leaks or ruptures, concurrent with containment bypass from main steam line or feedwater line breaches*. (NRC Publication No. NUREG-0933). Retrieved from <http://www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr0933/sec3/188r1.html>
- Wang, Y., & Meeker, W.Q. (2005). A statistical model to adjust for flaw-size bias in the computation of probability of detection. *Proceedings of the Review of Progress in Quantitative Nondestructive Evaluation*, July 31-August 5, Brunswick, Maine. doi:10.1063/1.2184745
- Yuan, X.X., Mao, D., & Pandey, M.D. (2009). A Bayesian approach to modeling and predicting pitting flaws in steam generator tubes. *Reliability Engineering and System Safety*, vol. 94(11), pp. 1838-1847. doi:10.1016/j.res.2009.06.001



Kaushik Chatterjee received his B.Tech. degree in mechanical engineering from Indian Institute of Technology (IIT), Roorkee (2004); and his M.S. degree in mechanical engineering from University of Maryland College Park, MD (2009). He is currently a PhD candidate in the mechanical engineering department at University of Maryland College Park. His PhD thesis focuses on developing a reliability prognosis and health management approach for steam generator tubes used in modular pressurized water reactors. He is also a graduate research assistant at the Center for Risk and Reliability (CRR) at University of Maryland College Park, and his research interests focus on finite element analysis, reliability modeling of complex systems and components, probabilistic physics of failure (PoF) modeling of failure mechanisms, and Bayesian uncertainty analysis.



Dr. Mohammad Modarres is a Professor of Nuclear Engineering and Reliability Engineering at University of Maryland College Park. His research areas are probabilistic risk assessment, uncertainty analysis, and physics of failure probabilistic modeling of failure mechanisms of mechanical components, systems and structures. He has served as a consultant to several governmental agencies, private organizations and national laboratories in areas related to probabilistic risk assessment, especially applications to complex systems and processes such as nuclear power plants. He has over 200 papers in archival journals and proceedings of conferences and three books in various areas of risk and reliability engineering.

A Combined Anomaly Detection and Failure Prognosis Approach for Estimation of Remaining Useful Life in Energy Storage Devices

Marcos E. Orchard¹, Liang Tang², and George Vachtsevanos³

¹*Electrical Engineering Department, Universidad de Chile, Santiago 8370451, Chile
morchard@ing.uchile.cl*

²*Impact Technologies, LLC, Rochester, NY 14623, USA
liang.tang@impact-tek.com*

³*School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
gfv@ece.gatech.edu*

ABSTRACT

Failure prognosis and uncertainty representation in long-term predictions are topics of paramount importance when trying to ensure safety of the operation of any system. In this sense, the use of particle filter (PF) algorithms -in combination with outer feedback correction loops- has contributed significantly to the development of a robust framework for online estimation of the remaining useful equipment life. This paper explores the advantages of using a combination of PF-based anomaly detection and prognosis approaches to isolate rare events that may affect the understanding about how the fault condition evolves in time. The performance of this framework is thoroughly compared using a set of ad hoc metrics. Actual data illustrating aging of an energy storage device (specifically battery state-of-health (SOH) measurements [A-hr]) are used to test the proposed framework.

1. INTRODUCTION

Particle-filtering (PF) based prognostic algorithms (Orchard, 2009; Orchard and Vachtsevanos, 2009; Orchard *et al.*, 2009) have been established as the de facto state of the art in failure prognosis. PF algorithms allow avoiding the assumption of Gaussian (or log-normal) probability density function (pdf) in nonlinear processes, with unknown model parameters, and simultaneously help to consider non-uniform probabilities of failure for particular regions of the state domain. Particularly, the authors in (Orchard *et al.*, 2008) have proposed a mathematically rigorous method (based on PF, function kernels, and outer correction loops) to represent and manage uncertainty in long-term predictions. However, there are still unsolved issues

Marcos E. Orchard *et al.* This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

regarding the proper representation for the probability of rare events and highly non-monotonic phenomena, since these events are associated to particles located at the tails of the predicted probability density functions.

This paper presents a solution for this problem that is based on a combination of a PF-based anomaly detection modules (which are in charge of detecting rare events within the evolution of the fault condition under analysis) and PF-prognostic schemes to estimate the remaining useful life of a piece of equipment. The paper is structured as follows: Section 2 introduces the basics of particle filtering (PF) and its application to the field of anomaly detection and failure prognostics. Section 3 presents a combined framework using actual failure data measuring battery state-of-health (SOH, [A hr]), where it is of interest to detect capacity regeneration phenomena in an online fashion. Section 4 utilizes performance metrics to assess prognostic results and evaluates the proposed scheme, when compared to the classic PF prognosis framework (Orchard, 2009; Orchard and Vachtsevanos, 2009; Vachtsevanos *et al.*, 2006). Section 5 states the main conclusions.

2. PARTICLE FILTERING, ANOMALY DETECTION AND FAILURE PROGNOSIS

Nonlinear filtering is defined as the process of using noisy observation data to estimate at least the first two moments of a state vector governed by a dynamic nonlinear, non-Gaussian state-space model. From a Bayesian standpoint, a nonlinear filtering procedure intends to generate an estimate of the posterior probability density function $p(x_t | y_{1:t})$ for the state, based on the set of received measurements. Particle Filtering (PF) is an algorithm that intends to solve this estimation problem by efficiently selecting a set of N particles $\{x^{(i)}\}_{i=1 \dots N}$ and weights $\{w_t^{(i)}\}_{i=1 \dots N}$, such that the

state pdf may be approximated (Doucet, 1998; Doucet *et al.*, 2001; Andrieu *et al.*, 2001; Arulampalam *et al.*, 2002) by:

$$\begin{aligned} \tilde{\pi}_t^N(x_t) &= \sum_{i=1}^N w_t^{(i)} \delta(x_t - x_t^{(i)}) \\ w(x_{0:t}) &= \frac{\pi_t(x_{0:t})}{q_t(x_{0:t})} \propto \frac{p(y_t | x_t) p(x_t | x_{0:t-1})}{q_t(x_t | x_{0:t-1})} \end{aligned} \quad (1)$$

where $q_t(x_{0:t})$ is referred to as the importance sampling density function (Arulampalam *et al.*, 2002; Doucet *et al.*, 2001). The choice of this importance density function is critical for the performance of the particle filter scheme. In the particular case of nonlinear state estimation, the value of the particle weights $w_{0:t}^{(i)}$ is computed by setting the importance density function equal to the *a priori* pdf for the state, i.e., $q_t(x_{0:t} | x_{0:t-1}) = p(x_t | x_{t-1})$ (Arulampalam *et al.*, 2002). Although this choice of importance density is appropriate for estimating the most likely probability distribution according to a particular set of measurement data, it does not offer a good estimate of the probability of events associated to high-risk conditions with low likelihood. This paper explores the possibility of using a PF-based detection scheme to isolate those types of events.

2.1 PF-based Anomaly Detection

A PF-based anomaly detection procedure (Orchard and Vachtsevanos, 2009; Verma *et al.*, 2004) aims at the identification of abnormal conditions in the evolution of the system dynamics, under assumptions of non-Gaussian noise structures and nonlinearities in process dynamic models, using a reduced particle population to represent the state pdf. The method also allows fusing and utilizing information present in a feature vector (measurements) to determine not only the operating condition (mode) of a system, but also the causes for deviations from desired behavioral patterns. This compromise between model-based and data-driven techniques is accomplished by the use of a PF-based module built upon the nonlinear dynamic state model (2):

$$\begin{cases} x_d(t+1) = f_b(x_d(t) + n(t)) \\ x_c(t+1) = f_i(x_d(t), x_c(t), \omega(t)) \\ \text{Features}(t) = h_i(x_d(t), x_c(t), v(t)) \end{cases} \quad (2)$$

where f_b , f_i and h_i are non-linear mappings, $x_d(t)$ is a collection of Boolean states associated with the presence of a particular operating condition in the system (normal operation, fault type #1, #2), $x_c(t)$ is a set of continuous-valued states that describe the evolution of the system given those operating conditions, $\omega(t)$ and $v(t)$ are non-Gaussian distributions that characterize the process and feature noise signals respectively. Since the noise signal $n(t)$ is a measure of uncertainty associated with Boolean states, it is

recommendable to define its probability density through a random variable with bounded domain. For simplicity, $n(t)$ may be assumed to be zero-mean i.i.d. uniform white noise.

A particle filtering approach based on model (2) allows statistical characterization of both Boolean and continuous-valued states, as new feature data are received. As a result, at any given instant of time, this framework provides an estimate of the probability masses associated with each fault mode, as well as a pdf estimate for meaningful physical variables in the system. Once this information is available within the anomaly detection module, it is conveniently processed to generate proper fault alarms and to inform about the statistical confidence of the detection routine.

Furthermore, pdf estimates for the system continuous-valued states (computed at the moment of fault detection) may be also used as initial conditions in failure prognostic routines, giving an excellent insight about the inherent uncertainty in the prediction problem. As a result, a swift transition between the two modules (anomaly detection and prognosis) may be performed, and moreover, reliable prognosis can be achieved within a few cycles of operation after the fault is declared.

2.2 PF-based Failure Prognosis

Prognosis, and more generally, the generation of long-term predictions, is a problem that goes beyond the scope of filtering applications since it involves future time horizons. Hence, if PF-based algorithms are to be used for prognosis, a procedure is required that has the capability to project the current particle population into the future in the absence of new observations (Orchard, 2009; Orchard and Vachtsevanos, 2009).

Any prognosis scheme requires the existence of at least one feature providing a measure of the severity of the fault condition under analysis (fault dimension). If many features are available, they can in principle be combined to generate a single signal. In Therefore, it is possible to describe the evolution in time of the fault dimension through the nonlinear state equation (Orchard *et al.*, 2008):

$$\begin{cases} x_1(t+1) = x_1(t) + x_2(t) \cdot F(x(t), t, U) + \omega_1(t) \\ x_2(t+1) = x_2(t) + \omega_2(t) \\ y(t) = x_1(t) + v(t) \end{cases} \quad (3)$$

where $x_1(t)$ is a state representing the fault dimension under analysis, $x_2(t)$ is a state associated with an unknown model parameter, U are external inputs to the system (load profile, etc.), $F(x(t), t, U)$ is a general time-varying nonlinear function, and $\omega_1(t)$, $\omega_2(t)$, $v(t)$ are white noises (not necessarily Gaussian). The nonlinear function $F(x(t), t, U)$ may represent a model, for example a model based on first principles, a neural network, or model based on fuzzy logic.

By using the aforementioned state equation to represent the evolution of the fault dimension in time, one can generate long term predictions using kernel functions to reconstruct the estimate of the state pdf in future time instants (Orchard *et al.*, 2008):

$$\hat{p}(x_{t+k} | \hat{x}_{t+k-1}) \approx \sum_{i=1}^N w_{t+k-1}^{(i)} K(x_{t+k} - E[x_{t+k}^{(i)} | \hat{x}_{t+k-1}^{(i)}]), \quad (4)$$

where $K(\cdot)$ is a kernel density function, which may correspond to the process noise pdf, a Gaussian kernel or a rescaled version of the Epanechnikov kernel.

The resulting predicted state pdf contains critical information about the evolution of the fault dimension over time. One way to represent that information is through the expression of statistics (expectations, 95% confidence intervals), either the End-of-Life (EOL) or the Remaining Useful Life (RUL) of the faulty system. A detailed procedure to obtain the RUL pdf from the predicted path of the state pdf is described and discussed in (Orchard, 2009; Patrick *et al.*, 2007; Zhang *et al.*, 2009). Essentially, the RUL pdf can be computed from the function of probability-of-failure at future time instants. This probability is calculated using both the long-term predictions and empirical knowledge about critical conditions for the system. This empirical knowledge is usually incorporated in the form of thresholds for main fault indicators (also referred to as the hazard zones).

In real applications, hazard zones are expected to be statistically determined on the basis of historical failure data, defining a critical pdf with lower and upper bounds for the fault indicator (H_{lb} and H_{ub} , respectively). Let the hazard zone specify the probability of failure for a fixed value of the fault indicator, and the weights $\{w_{t+k}^{(i)}\}_{i=1 \dots N}$ represent the predicted probability for the set of predicted paths, then the probability of failure at any future time instant (namely the RUL pdf) by applying the law of total probabilities, as shown in Eq. (5).

$$\hat{p}_{TRF}(t) = \sum_{i=1}^N \Pr(\text{Failure} | X = \hat{x}_t^{(i)}, H_{lb}, H_{ub}) \cdot w_t^{(i)} \quad (5)$$

Once the RUL pdf has been computed by combining the weights of predicted trajectories with the hazard zone specifications, prognosis confidence intervals, as well as the RUL expectation can be extracted.

3. A COMBINED ANOMALY DETECTION AND FAILURE PROGNOSIS APPROACH: CASE STUDY DEFINITION

An appropriate case study has been selected to demonstrate the efficacy of a scheme that includes a PF-based anomaly detection module working in combination with a PF-based prognostic algorithm. Consider the case of energy storage

devices, particularly of Li-Ion batteries, where continuous switching between charge and discharge cycles may cause momentary increments in the battery SOH (capacity regeneration). These sudden increments directly affect RUL estimates in a classic PF-based prognostic scheme since the state pdf estimate has to be adjusted according to new measurements (thus modifying long-term predictions), while the observed phenomenon typically disappears after a few cycles of operation. Particularly in the case of Li-Ion batteries, the regeneration phenomena can produce an unexpected short-term increment of the battery SOH of about 10% of the nominal capacity.

The analysis of the aforementioned phenomena will be done using data registering two different operational profiles (charge and discharge) at room temperature. On the one hand, charging is carried out in a constant current (CC) mode at 1.5[A] until the battery voltage reached 4.2[V] and then continued in a constant voltage mode until the charge current dropped to 20[mA]. On the other hand, discharge is carried out at a constant current (CC) level of 2[A] until the battery voltage fell to 2.5[V]. Impedance measurements provide insight into the internal battery parameters that change as aging progresses. Repeated charge and discharge cycles result in aging of the batteries. Impedance measurements were done through an electrochemical impedance spectroscopy (EIS) frequency sweep from 0.1[Hz] to 5[kHz]. The experiments were stopped when the batteries reached end-of-life (EOL) criteria, which was a 40% fade in rated capacity (from 2[A-hr] to 1.2[A-hr]). This dataset can be used both for the prediction of both remaining charge (for a given discharge cycle) and remaining useful life (RUL).

Two main operating conditions are thus distinguished: the *normal* condition reflects the fact that the battery SOH is slowly diminishing as a function of the number of charge/discharge cycles; while the *anomalous* condition indicates an abrupt increment in the battery SOH (regeneration phenomena). To detect the condition of interest, a PF-based anomaly detection module is implemented using nonlinear model (6), where $x_{d,1}$ and $x_{d,2}$ are Boolean states that indicate *normal* and *anomalous* conditions respectively, $x_{c,1}$ is the continuous-valued state that represents the battery SOH, β is a positive time-varying model parameter, $x_{c,2}$ is the added SOH because of the capacity regeneration phenomena, and where $\alpha(t)$ and $v(t)$ have been selected as zero mean Gaussian noises for simplicity. The initial battery SOH in the data set used for this analysis is 2[A-hr.], which determines the initial condition of (6).

Besides detecting the regeneration condition, it is desired to obtain some measure of the statistical confidence of the alarm signal. For this reason, two outputs can be extracted from the anomaly detection module. The first output is the expectation of the Boolean state $x_{d,2}$, which constitutes an

estimate of the probability of regeneration. The second output is the statistical confidence needed to declare the condition via hypothesis testing (H_0 : “no anomaly is being detected” vs. H_1 : “capacity regeneration is being detected”). The latter output needs another pdf to be considered as the baseline. In this case, that pdf could be the filtering estimate of state x_{c1} .

$$\begin{cases} \begin{bmatrix} x_{d,1}(t+1) \\ x_{d,2}(t+1) \end{bmatrix} = f_b \left(\begin{bmatrix} x_{d,1}(t) \\ x_{d,2}(t) \end{bmatrix} + n(t) \right) \\ x_{c1}(t+1) = (1-\beta)x_{c1}(t) + \omega_1(t) \\ x_{c2}(t+1) = 0.95x_{c2}(t) \cdot x_{d,2}(t) + 0.2x_{d,1}(t) + \omega_2(t) \end{cases} \\ y(t) = x_{c1}(t) + x_{c2}(t) \cdot x_{d,2}(t) + v(t) \end{cases} \quad (6)$$

$$f_b(x) = \begin{cases} [1 \ 0]^T, & \text{if } \|x - [1 \ 0]^T\| \leq \|x - [0 \ 1]^T\| \\ [0 \ 1]^T, & \text{else} \end{cases}$$

$$\begin{bmatrix} x_{d,1}(0) & x_{d,2}(0) & x_{c1}(0) & x_{c2}(0) \end{bmatrix}^T = [1 \ 0 \ 2 \ 0]^T$$

Moreover, since this is a PF-based module, one way to generate an on-line indicator of statistical confidence for the detection procedure is to consider the sum of the weights of all particles i such that $x_c^{(i)}(T) \geq z_{1-\alpha, \mu, \sigma^2}$, where α is the desired test confidence and T is the detection time, which is essentially equivalent to an estimate of $(1 - \text{type II error})$, or equivalently the probability of detection. If additional information is required, it is possible to compute the value of the Fisher’s Discriminant Ratio, as in (7).

$$F_{index}(T) = \frac{\left| \mu - \sum_{i=1}^N w_T^{(i)} \cdot x_c^{(i)}(T) \right|^2}{\left(\sigma^2 + \sum_{i=1}^N w_T^{(i)} \cdot \left(x_c^{(i)}(T) - \sum_{j=1}^N w_T^{(j)} \cdot x_c^{(j)}(T) \right)^2 \right)} \quad (7)$$

It must be noted that, in this approach, no particular specification about the detection threshold has to be made prior to the actual experiment. Customer specifications are translated into acceptable margins for the *type I* and *type II* errors in the detection routine. The algorithm itself will indicate when the *type II error* (false negatives) has decreased to the desired level.

Once the regeneration phenomena have been adequately isolated, it is the task of the PF-based prognosis framework to come up with a pdf estimate of the remaining useful life of the Li-Ion battery. For this purpose, instead of a physics-based model we will employ here a population-growth-based model (Patrick *et al.*, 2007; Orchard *et al.*, 2008, Zhang *et al.*, 2009) that has been trained using online SOH measurements (fault dimension in [A-hr]), where $x_1(t)$ is a

state representing the fault dimension, $x_2(t)$ is a state associated with an unknown model parameter, $x_3(t)$ is a state associated with the capacity regeneration phenomena, a , b , C and m are constants associated to the duration and intensity of the battery load cycle (external input U), and $0 \leq \alpha \leq 1$ is a parameter that characterizes the regeneration.

$$\begin{cases} x_1(t+1) = x_1(t) + C \cdot x_2(t) \cdot (a - b \cdot t + t^2)^m + \omega_1(t) \\ x_2(t+1) = x_2(t) + \omega_2(t) \\ x_3(t+1) = \alpha \cdot x_3(t) + \omega_3(t) \end{cases}, \quad (8) \\ y(t) = x_1(t) + x_3(t) + v(t)$$

The objective of a prognostic routine applied to the system defined by (8), and particularly for the ones based on PF algorithms, is to estimate (preferably in an online fashion) the current battery SOH, isolating the effect of the regeneration phenomena, and to use that information to estimate the amount of cycles remaining until this quantity falls below the threshold of 1.2[A-hr].

The analysis will focus on the quality of the estimate for the state components x_1 and x_3 , after each capacity regeneration phenomena and on the accuracy exhibited by the corresponding End-of-Life (EOL) pdf estimate. Performance comparison is done with respect to a classic (SIR) PF-based prognostic framework (Orchard *et al.*, 2009), given same initial conditions. It should be noted that the implementation chosen here considers a correction loop that simultaneously updates the variance of kernel associated to the white noise $\omega_2(t)$ according to the short-term prediction (Orchard, Tobar and Vachtsevanos, 2010).

The implementation of the aforementioned scheme has been performed using MATLAB® environment. A complete description of the results obtained, and a comparison with classic PF-based routines, follows in Section 4.

4. ASSESSMENT AND EVALUATION OF THE PROPOSED FRAMEWORK USING PERFORMANCE METRICS

Estimates obtained from a Particle Filtering algorithm are based on the realization of a stochastic process and measurement data. Assessment or comparison between different strategies should consider performance statistics rather than a performance assessment based on a single experiment or realization. For that reason, all results presented in this paper consider the statistical mean of 30 realizations for the particle filter algorithm and a single measurement data set (no major differences were found when considering more realizations).

In addition, the assessment and evaluation of prognostic algorithms require appropriate performance metrics capable of incorporating concepts such as “accuracy” and “precision” of the RUL pdf estimate (Vachtsevanos *et al.*, 2006). “Accuracy” is defined as the difference between the actual failure time and the estimate of its expectation, while

“precision” is an expression of the spread (e.g., standard deviation). These indicators should also consider the fact that both the RUL and $E_t\{RUL\}$ (estimate, at time t , of the expectation of the equipment RUL) are random variables. Moreover, it is desirable that all indicators assume that, at any time t , it is possible to compute an estimate of the 95% confidence interval (CI_t) for the EOL.

In particular this paper uses three indicators to evaluate prognostic results, which are presented and detailed in (Orchard, Tobar and Vachtsevanos, 2009): (1) RUL precision index ($RUL-OPI$), (2) RUL accuracy-precision index, and (3) RUL online steadiness index ($RUL-OSI$). $RUL-OPI$ considers the relative length of the 95% confidence interval computed at time t (CI_t), when compared to the RUL estimate. It is expected that the more data the algorithm processes, the more precise the prognostic becomes:

$$RUL-OPI(t) = e^{-\left(\frac{\sup(CI_t) - \inf(CI_t)}{E_t\{EOL\} - t}\right)} \quad (9)$$

$$0 < RUL-OPI(t) \leq 1, \forall t \in [1, E_t\{EOL\}], t \in \mathbb{N}.$$

The RUL accuracy-precision index, measures the error in EOL estimates relative to the width of its 95% confidence interval (CI_t). It also penalizes late predictions, i.e., whenever $E_t\{EOL\}$ (the expected EOL) is bigger than $GroundTruth\{EOL\}$ (actual failure happens before the expected time). This indicator can be computed only after the end of the simulation. Finally, the $RUL-OSI$ considers the variance of the EOL conditional expectation, computed with measurement data available at time t . Good prognostic results are associated to small values for the $RUL-OSI$. All performance metrics will be evaluated at all time instants.

In the case study presented in this paper (RUL/EOL estimation of a Li-Ion battery) the time is measured in cycles of operation. A cycle of operation consists of two different operational states applied to the battery at room temperature (charge and discharge).

It is essential to note that algorithm assessment only considered RUL estimates generated until the 120th cycle of operation, which corresponded to about 75% of the actual useful life of the battery (actual EOL of the experiment is 159 cycles), since it is of more interest to evaluate the algorithm’s performance when the size of the prediction window is large enough to allow for corrective actions. Moreover, given that PF-based prognostic algorithms tend to improve their performance as the amount of available data increases (Orchard and Vachtsevanos, 2009), the closer the system is to the actual EOL, the more accurate the resulting EOL estimate. This needs to be kept in mind when analyzing results presented both in Figure 1 and Figure 2.

Figure 1 (a) shows online tracking for the battery SOH (coarse trace) using a classic PF-based prognostic approach until the 120th cycle of operation, the hazard zone around

1.2 [A-hr] (marked as a horizontal band), and the 95% confidence interval of EOL (coarse vertical dashed lines) computed at the 120th cycle. Figure 1 (b) only shows the EOL pdf estimate computed at the end of the 120th cycle of operation. The result of the classic PF-based prognostic approach is accurate to two cycles (the expected value of the EOL pdf is 161 cycles, while the ground truth data for the EOL is 159 cycles). However, the state estimate, in this case, does not exhibit the same level of accuracy when describing capacity regeneration phenomena registered at the 19th, 30th, and 47th cycles of operation; see Table 1. In fact, regeneration phenomena momentarily affect the algorithm performance, in particular in terms of steadiness of the solution (Orchard, Tobar and Vachtsevanos, 2009) as the analysis based on performance metrics will corroborate shortly.

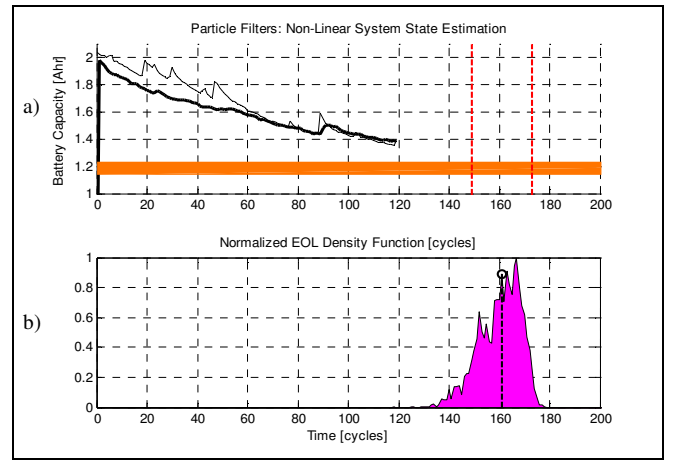


Figure 1. Case study. (a) Measurement data (fine trace), PF-based estimate (coarse trace), and 95% confidence interval (vertical read dashed lines). (b) EOL pdf estimate using classic PF-based prognosis framework and its expectation

Table 1. Estimates for system output $y(t)$

Cycle	Measured data	$E_t\{y(t)\}$ Classic PF-based routine	$E_t\{y(t)\}$ PF-based Anomaly detection and prognosis scheme
19 th	1.98	1.77	1.80
30 th	1.92	1.70	1.78
47 th	1.82	1.62	1.69

Figure 2 (a) shows online tracking for the battery SOH (coarse trace) using the proposed scheme that combines a PF-based anomaly detection module to identify regeneration phenomena and a PF-based prognosis framework for the estimation of the battery EOL. This figure also illustrates the hazard zone around 1.2 [A-hr], and the 95% confidence interval of EOL computed at the 120th cycle. Figure 2 (b)

shows the EOL pdf estimate computed at the end of the 120th cycle of operation.

Figure 2 shows that the proposed scheme is equally capable of providing an accurate estimate of the battery RUL with an expected value of the EOL pdf (computed at the 120th cycle of operation) of 158 cycles, while the ground truth data for the EOL is 159 cycles). However, it is more interesting to note that the information provided by the anomaly detection module noticeably improves the state estimate at early stages of the test, particularly between the 20th and the 60th cycle of operation (see Table 1), allowing a better description of the regeneration phenomena that affect the Li-Ion battery. This demonstrates how the existence of particles in areas of low likelihood can help to improve the state estimate when rare, unlikely events or highly non-monotonic phenomena occur.

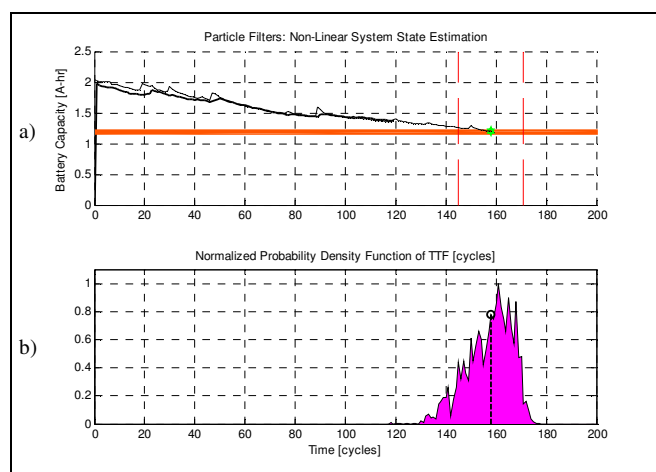


Figure 2. (a) Measurement data (fine trace), PF-based estimate (coarse trace), and 95% confidence interval. (b) EOL pdf estimate, using a combination of PF-based anomaly detection and prognosis approaches, and its expectation

Even more compelling, similar conclusions can be drawn when using prognostic performance metrics to assess the performance of the classic PF and the proposed scheme that includes an anomaly detection module; Figure 3 summarizes tracking and prediction with all tracking estimates generated until the 120th cycle.

Figure 3 (a) shows the evaluation of RUL-OPI as measurement data are included in a sequential manner into the prediction algorithm. One of the main characteristics of this indicator is that it penalizes the width of the 95th% confidence interval as the system approaches EOL. The value of this indicator is comparable for both algorithms (around 0.5 near the end of the experiment).

However, both the accuracy-precision and the RUL-OSI indices indicate noticeable advantages of the combination of PF-based anomaly detection and prognosis routines when compared to its classic version as illustrated in Figure 3 (b)

and Figure 3 (c). The evaluation of the accuracy-precision index clearly shows improved performance in the case of the proposed framework, which translates into better estimates for the EOL conditional expectation. Similar conclusions can be obtained from Figure 3 (c), where the steadiness RUL-OSI index shows that the impact of detecting regenerating phenomena (and adjusting state estimates accordingly) is limited to bounded periods of time.

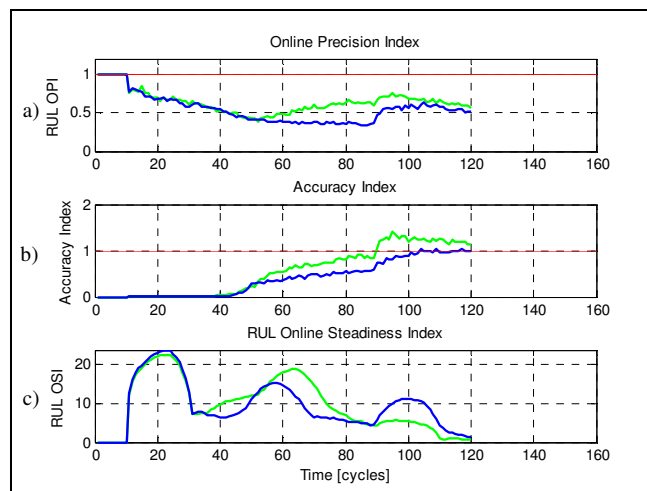


Figure 3. Performance metric evaluation in case study.

Comparison between classic PF (green trace) and a combined PF-based anomaly detection/prognostic module (blue trace)

Previous research work (Orchard, 2009; Orchard and Vachtsevanos, 2009) has already shown better results when using classic PF-based prognostic framework, compared to other approaches. For this reason, this performance analysis did not consider other methods such as the extended Kalman filter in its formulation.

5. CONCLUSION

This paper presents a case study where a combined version of PF-based anomaly detection and classic PF-based prognosis algorithms is applied to estimate the remaining useful life of an energy storage device (Li-Ion battery). A comparison based on prognosis performance metrics indicates that the proposed anomaly detection/prognostic approach is more suitable than classic PF methods to represent highly non-monotonic phenomena such as capacity regeneration phenomena between charging periods, in terms of accuracy of the state estimate and steadiness of the RUL estimate. We surmise that the information provided by the anomaly detection module, in an online fashion, allow providing a more conservative estimate of the RUL of the faulty piece of equipment. We surmise that it also helps to incorporate the probability of rare and costly events in the evolution of the fault condition in time.

ACKNOWLEDGEMENT

The authors would like to acknowledge the support of NASA Aviation Safety Program/IVHM Project NRA NNA08BC20C. Thanks also go to Conicyt for Dr. Orchard's financial support via Fondecyt #1110070.

REFERENCES

- Andrieu, C., A. Doucet, E. Puskaya, (2001). "Sequential Monte Carlo Methods for Optimal Filtering," in *Sequential Monte Carlo Methods in Practice*, A. Doucet, N. de Freitas, and N. Gordon, Eds. NY: Springer-Verlag.
- Arulampalam, M.S., S. Maskell, N. Gordon, T. Clapp, (2002). "A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174 – 188.
- Doucet, A., (1998). "On sequential Monte Carlo methods for Bayesian Filtering," Technical Report, Engineering Department, Univ. Cambridge, UK.
- Doucet, A., N. de Freitas, N. Gordon, (2001). "An introduction to Sequential Monte Carlo methods," in *Sequential Monte Carlo Methods in Practice*, A. Doucet, N. de Freitas, and N. Gordon, Eds. NY: Springer-Verlag.
- Orchard, M., G. Kacprzyński, K. Goebel, B. Saha, G. Vachtsevanos, (2008). "Advances in Uncertainty Representation and Management for Particle Filtering Applied to Prognostics," 2008 *International Conference on Prognostics and Health Management PHM 2008*, Denver, CO, USA.
- Orchard, M., (2009). *On-line Fault Diagnosis and Failure Prognosis Using Particle Filters. Theoretical Framework and Case Studies*, Publisher: VDM Verlag Dr. Müller Aktiengesellschaft & Co. KG, Saarbrücken, Germany, 108 pages. Atlanta: The Georgia Institute of Technology, Diss., 2007.
- Orchard, M. G. Vachtsevanos, (2009). "A Particle Filtering Approach for On-Line Fault Diagnosis and Failure Prognosis," *Transactions of the Institute of Measurement and Control*, vol. 31, no. 3-4, pp. 221-246.
- Orchard, M., F. Tobar, G. Vachtsevanos, (2009). "Outer Feedback Correction Loops in Particle Filtering-based Prognostic Algorithms: Statistical Performance Comparison," *Studies in Informatics and Control*, vol.18, Issue 4, pp. 295-304.
- Orchard, M., L. Tang, K. Goebel, G. Vachtsevanos, (2009). "A Novel RSPF Approach to Prediction of High-Risk, Low-Probability Failure Events," First Annual Conference of the Prognostics and Health Management Society, San Diego, CA, USA.
- Patrick, R., M. Orchard, B. Zhang, M. Koelemay, G. Kacprzyński, A. Ferri, G. Vachtsevanos, (2007). "An Integrated Approach to Helicopter Planetary Gear Fault Diagnosis and Failure Prognosis," 42nd Annual Systems Readiness Technology Conference, AUTOTESTCON 2007, Baltimore, USA.
- Verma, V., G. Gordon, R. Simmons, S. Thrun, (2004). "Particle Filters for Rover Fault Diagnosis," *IEEE Robotics & Automation Magazine*, pp. 56 – 64.
- Vachtsevanos, G., F.L. Lewis, M.J. Roemer, A. Hess, B. Wu, (2006). *Intelligent Fault Diagnosis and Prognosis for Engineering Systems*, Hoboken, NJ, John Wiley and Sons.
- Zhang, B., T. Khawaja, R. Patrick, M. Orchard, A. Saxena, G. Vachtsevanos, (2009). "A Novel Blind Deconvolution De-Noising Scheme in Failure Prognosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 58, no. 2, pp. 303-310.

A Mobile Robot Testbed for Prognostics-Enabled Autonomous Decision Making

Edward Balaban¹, Sriram Narasimhan², Matthew Daigle³, José Celaya⁴, Indranil Roychoudhury⁵, Bhaskar Saha⁶, Sankalita Saha⁷, and Kai Goebel⁸

^{1,8}*NASA Ames Research Center, Moffett Field, CA, 94035, USA*
edward.balaban@nasa.gov
kai.goebel@nasa.gov

^{2,3}*University of California, Santa Cruz, NASA Ames Research Center, Moffett Field, CA, 94035, USA*
sriram.narasimhan-1@nasa.gov
matthew.j.daigle@nasa.gov

^{4,5}*SGT Inc., NASA Ames Research Center, Moffett Field, CA, 94035, USA*
jose.r.celaya@nasa.gov
indranil.roychoudhury@nasa.gov

^{6,7}*MCT Inc., NASA Ames Research Center, Moffett Field, CA, 94035, USA*
bhaskar.saha@nasa.gov
sankalita.saha-1@nasa.gov

ABSTRACT

The ability to utilize prognostic system health information in operational decision making, especially when fused with information about future operational, environmental, and mission requirements, is becoming desirable for both manned and unmanned aerospace vehicles. A vehicle capable of evaluating its own health state and making (or assisting the crew in making) decisions with respect to its system health evolution over time will be able to go further and accomplish more mission objectives than a vehicle fully dependent on human control. This paper describes the development of a hardware testbed for integration and testing of prognostics-enabled decision making technologies. Although the testbed is based on a planetary rover platform (K11), the algorithms being developed on it are expected to be applicable to a variety of aerospace vehicle types, from unmanned aerial vehicles and deep space probes to manned aircraft and spacecraft. A variety of injectable fault modes is being investigated for electrical, mechanical, and power subsystems of the testbed. A software simulator of the K11 has been developed, for both nominal and off-nominal operating modes, which allows prototyping and validation of algorithms prior to their deployment on hardware. The simulator can also aid in the

decision-making process. The testbed is designed to have interfaces that allow reasoning software to be integrated and tested quickly, making it possible to evaluate and compare algorithms of various types and from different sources. Currently, algorithms developed (or being developed) at NASA Ames - a diagnostic system, a prognostic system, a decision-making module, a planner, and an executive - are being used to complete the software architecture and validate design of the testbed.

1. INTRODUCTION

Over the last several years, testbeds have been constructed at NASA and elsewhere for the purpose of diagnostic and prognostic research on components important to aerospace vehicles: electronics, actuators, batteries, and others. For examples, please refer to (Poll, et al., 2007), (Smith, et al., 2009), (Balaban, Saxena, Narasimhan, Roychoudhury, Goebel, & Koopmans, 2010). However, there still remained a need for a testbed that supported development of algorithms performing reasoning on both the component and system levels, and optimizing decision-making with system health information taken into account. Such a testbed would also, ideally, be inexpensive to operate and not require lengthy experiment setup times. The main categories of tasks to be performed on the testbed were defined as the following: (1) development of system-level prognostics-enabled decision making (PDM) algorithms; (2) maturation

Balaban et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

and standardization of interfaces between various reasoning algorithms; (3) performance comparison among the algorithms from different organizations; and (4) generation of publicly available datasets for enabling further research in PDM.

Adding decision optimization capability to a diagnostic/prognostic health management system will allow to not only determine if a vehicle component is failing and how long would it take for it to fail completely, but also to use that information to take (or suggest) actions that can optimize vehicle maintenance, ensure mission safety, or extend mission duration. Depending on the prediction time horizon for the fault, the character of these actions can vary. If a fault is expected to develop into a complete failure in a matter of seconds, a rapid controller reconfiguration, for example, may be required. If the fault progression takes minutes then, perhaps, reconfiguration of the vehicle (such as switching from the low gain antenna to the high gain antenna) can help remedy the situation. Finally, if the remaining useful life (RUL) for the faulty component is measured in hours, days, weeks, or longer, a new mission plan or adjustments to the logistics chain may be warranted (often in conjunction with lower level actions).

While eventually an aerial test vehicle, such as an unmanned fixed wing airplane or a helicopter, would allow testing of the aforementioned technologies on complex scenarios and with motion in three-dimensional space, operating such a testbed is often expensive. A single flight hour often requires many days of preparation. Safety requirements for an aerial test vehicle, even without a human onboard, are also usually quite stringent. In contrast, a rover whose movement is restricted to two dimensions can operate at low speeds in a controlled environment, making experiments easier and safer to set up. The experiments can still involve motion, complex subsystems interactions, and elaborate mission plans, but the possibility of a dangerous situation occurring is reduced significantly. For technologies in early phases of development in particular, a land vehicle platform could provide a suitable initial test environment for the majority of development goals at a fraction of the cost of an aerial vehicle, usually with a clear transition path to the latter.

Guided by the above reasons and requirements, an effort was started to develop such a platform on the basis of the K11, a rover originally slated to serve as a robotic technologies test vehicle in the Antarctic (Lachat, Krebs, Thueer, & Siegwart, 2006). The rover equipment (such as its batteries) was updated and its sensor suite was expanded. A key distinction from other planetary rover development efforts should be stressed, however. The focus of this research is not to develop next-generation planetary rover hardware, but rather to use the K11 rover platform to create a realistic environment for testing novel PDM algorithms. These algorithms would then be used as blueprints by other

organizations in order to create PDM functionality for their particular applications.

Fault modes in components that are common to various types of vehicles (such as electric motors, batteries, or control electronics) were identified and injection methods for some of them were developed – with as much realism as practical. A software simulator, meant for allowing rapid validation of autonomy algorithms and for providing optimization guidance during hardware-in-the-loop experiments, was developed as well. While for the time being algorithms developed at NASA Ames are being used to populate the autonomy architecture on the K11, algorithms from other sources could be tested and evaluated in the future.

The next section of the paper, Section 2 focuses on the testbed hardware, while Section 3 summarizes work on the simulator to date, including experimental validation of the models. Section 4 describes the current reasoning software suite being deployed on the testbed and Section 5 provides a summary of the accomplishments and outlines potential future work.

2. TESTBED

The following section consists of three main parts: the first part describes the hardware of the testbed, including its sensor suite; the second focuses on the testbed (core) software; and the third one describes the methods used for fault injection. It should be noted that there is a distinction made in this work between core software and reasoning (including PDM) software. Examples in the former category include the operating system, the middleware providing communication between components, the data acquisition software, the low-level drivers – essentially the elements that enable the K11 to perform all of its functions under direct human control. The reasoning package, on the other hand, is the software that lessens or completely removes the dependence on a human operator. PDM software is what constitutes the test article for this testbed and its elements will be swapped in and out depending on the test plan. The current set of PDM elements is described in Section 3.

2.1. Hardware

The K11 is a four-wheeled rover (Figure 5) that was initially developed by the Intelligent Robotics Group (IRG) at NASA Ames to be an Antarctic heavy explorer. It had a design capacity to transport 100 kilograms of payload across ice and frozen tundra (Lachat, Krebs, Thueer, & Siegwart, 2006).

The rover was also previously used in experiments to test power consumption models and in a gearing optimization study. It has been tested on various types of terrain, including snow. The lightweight chassis was designed and built by BlueBotics SA. It consists of an H-structure and a



Figure 1: The K11 rover

joint around the roll axis to ensure that the wheels stay in contact with the ground on uneven terrain. The mass of the rover without the payload is roughly 140 kg. Its dimensions are approximately 1.4m x 1.1m x 0.63m. Each wheel on the K11 is driven by an independent 250 Watt graphite-brush motor from Maxon Motors equipped with an optical encoder. The wheels are connected to the motors through a bearing and gearhead system (gearhead ratio $r = 308$). Motors are controlled by Maxon Epos 70/10 single-axis digital motion controllers, capable of operating in velocity, position, homing, and current modes.

After considering various alternatives, LiFePO₄ (lithium iron phosphate) batteries, commonly used in modern electric vehicles, were selected to power the rover. LiFePO₄ batteries have a high charge density, perform well in high temperatures, and are not prone to combustion or explosion. Furthermore, they can withstand a high number (approximately 500) of charge/discharge cycles before needing to be replaced. There are four 12.8V 3.3 Ah LiFePO₄ batteries on the K11, connected in series. Each battery contains 4 cells.

The philosophy in developing the sensor suite on the K11 (summarized in Table 1) was to employ only those sensors or data acquisition hardware that are commonly available on a variety of vehicles or can be added at a reasonable cost, while also providing sufficient data for a PDM system. Each component is utilized to the maximum extent possible. For instance, the motor controllers are not only used for their primary purpose of operating the motors and giving feedback on their velocity and current consumption, but are also used to support external sensors. The unused controller analog input channels are called upon to read battery voltage and current readings. In a similar vein, a decision was made to utilize a modern off-the-shelf smartphone for part of the instrumentation suite instead of, for example, a dedicated GPS receiver and a gyroscope. The smartphone also provides a still/video camera, a compass, and data

Measurement Type	Manufacturer	Location/comments
GPS (longitude and latitude)	Motorola	On the smartphone
Gyroscope (roll, pitch, yaw)	Motorola	On the smartphone
Motor temperature	Omega	On each motor (to be implemented)
Battery temperature	Omega	On each battery pack (to be implemented)
Position encoder	Maxon	On each drive motor
Battery voltage	custom	On a custom PCB board measuring individual battery pack voltages
Total current	custom	On a custom PCB board measuring individual battery pack voltages
Individual motor current	Maxon	Part of motor controller

Table 1: Measurements available on the K11

processing and storage resources. It has a built-in wireless capability for communicating with other on-board components and directly with the ground station (as a backup to the main communication link through the on-board computer). The current phone used on the K11 is a Google Nexus S.

The bulk of the computational resources needed to operate the rover are provided by the onboard computer (an Intel Core 2 Duo laptop). Its responsibilities include executing the motor control software, performing data acquisition, as well as running all of the reasoning algorithms. A second laptop computer currently serves as a ground control station.

2.2. Software

Several of the core software elements on K11 are adopted, or being adopted, from the Service-Oriented Robotic Architecture (SORA) developed by the Intelligent Robotics Group (Fluckiger, To, & Utz, 2008). This includes the navigation software; the middleware, based on Common Object Request Broker Architecture (CORBA) (Object Management Group, 2004) and Adaptive Communication Environment (ACE) (Schmidt, 1994); and the telemetry software, the Robot Application Programming Interface Delegate (RAPID) (NASA Ames Research Center, 2011).

The smartphone (running Google Android 2.2 operating system) hosts a data acquisition module written in Java. That module collects data from the phone's sensors (described in the previous section) and sends it over a User Datagram Protocol (UDP) socket to the onboard computer. The central data acquisition software running on the computer receives the phone data, merges it with data received from other sources (e.g., voltage sensors, current

sensors, controller state, etc) and records it into a unified data file, which can then be transmitted to the ground control station. The central data acquisition software on the K11 is based on LabView from National Instruments.

The ground station graphical user interface (GUI) software is also written in LabView. It allows the operator to take manual control of the rover (via on-screen commands or a joystick) and to set data recording preferences. The operator can switch the GUI from interacting with the K11 hardware to interacting with the K11 simulator. One of the goals in developing the simulator (described further in Section 3) is to make the difference in interacting with it versus the rover hardware as minimal as possible. The operating system currently used on both the onboard computer and the ground station is Microsoft Windows XP. It will be replaced by a UNIX-based operating system in the near future.

2.3. Fault Modes

A number of fault modes have been identified so far for implementation on the testbed (Table 2). The criteria for their selection include relevance to a variety of aerospace vehicles (not just rovers), feasibility of implementation, and progression time from fault to failure. The last criterion is important because if the progression time is too brief (e.g. microseconds), then likely no useful action can be taken in the prognostic context to predict the remaining useful life of the component and remedy the situation. On the other hand, if the fault-to-failure progression time is measured in years, then running experiments on those fault modes may become impractical. Faults in both of the above categories could still be handled by diagnostic systems, however. Out of the fault modes described in Table 2, a few were selected for the initial phase of the project. The methods for their injection on the K11 are covered in more detail next. The methods for modeling progression of these faults in the simulator are described in Section 3.

2.3.1. Mechanical Jam and Motor Windings Deterioration

The first fault mode selected for implementation is a mechanical jam on the motor axle which leads to increased current, overheating of motor windings, deterioration of their insulation, and eventual failure of the motor due to a short in the motor windings. To maintain realism, a performance region for the motor is chosen (using manufacturer's specifications) where a healthy motor would have no problems keeping up with either speed or load requirements. In the presence of increased friction, however, the amount of current needed to satisfy the same speed and load demands is higher, leading to overheating. Unless speed and/or load are reduced or duty cycle (the proportion of time the motor is on versus duration of cool-down intervals) is adjusted, the heat build-up will eventually

Fault Mode	Injection method	Subsystem
battery capacity degradation	accelerated aging	power
battery charge tracking	normal operations	power
parasitic electric load	programmable	power distribution
motor failure	software	electro-mechanical
increased motor friction	mechanical brake	electro-mechanical
bearing spalls	machined spalls	electro-mechanical
sensor bias/drift/failure	software	sensors
motor driver faults	MOSFET replacement in the controller with an aged component	power distribution

Table 2: Potential fault modes

destroy the insulation of the motor windings and lead to motor failure. This fault mode was first implemented in the simulator and its model verified using experimental data collected on smaller-sized motors that were run to failure under similar conditions (please see section 3.2, Motor Modeling). A hardware fault injection using a mechanical brake on one of the rover motors will be implemented next. The rover motor will not be run to complete failure initially; instead the simulator model parameters and prognostic algorithms will be validated in experiments stopping short of creating permanent damage. Eventually, experiments that will take motors all the way to failure will be performed.

2.3.2. Parasitic Load

A parasitic electrical load will be injected on the main power distribution line via a remotely controlled rheostat. The rheostat can be set for resistance from 0 to 100 Ohms and can dissipate up to 600 Watts of power. The rheostat will simulate a situation where, for example, an accessory motor is continuously engaged due to a failed limit microswitch.

2.3.3. Electronics Faults

The systems on the K11 provide several opportunities for fault injection in electronics subsystems. Power electronics in the motor drivers allow fault injection in power-switching devices such as Metal-Oxide-Semiconductor Field-Effect Transistors (MOSFETs), Insulated Gate Bipolar Transistors (IGBTs) and electrolytic capacitors used for voltage filtering. These devices have a key role in providing current

to the motors, but are known for relatively high failure rates. Fault injection will also be implemented on the power switches of the motor winding H-bridges, where current will be routed to degraded power transistors during rover operation. In addition, some of the symptoms of power transistors failing will be replicated programmatically by varying the gate voltage. The premise of the fault injection in the H-bridge transistor is that it will diminish the performance of a motor winding, reducing torque and altering motor performance characteristics, making control difficult.

Efforts on accelerated aging of IGBTs and power MOSFETs are presented in (Celaya, Saxena, Wysocki, Saha, & Goebel, 2010). Accelerated aging methodologies for electrolytic capacitors under nominal loading and environmental conditions are presented in (Kulkarni, Biswas, Koutsoukos, Celaya, & Goebel, 2010); methodologies for accelerated aging via electrical overstress are presented in (Kulkarni, Biswas, Celaya, & Goebel, 2011). MOSFETs, IGBTs, and electrolytic capacitors at various levels of degradation will be used to inject component-level electronic faults, with some of the faults expected to have a cascading effect on other electronic and/or mechanical subsystems.

2.3.4. Battery Capacity Degradation

As the rover batteries go through charge/discharge cycles, their capacity to hold charge will diminish. The degradation rate will depend on several factors such as imposed loads, environmental conditions, and charge procedures. For example, Li-Ion chemistry batteries undergo higher rates of capacity fade with higher current draw and operational temperatures. Even at rest, this type of battery has chemical processes occurring that have long-term effects - for instance, latent self-discharge and transient recovery during relaxation. The depth-of-discharge (DoD) and even the storage temperature have major influences on the overall life of the battery as well. There is no specific mechanism required for injecting this fault – the batteries will age naturally in the course of rover operations. Some experiments will, however, utilize battery cells aged to a desired point in their life cycle on the battery aging test stand (Saha & Goebel, 2009)

2.3.5. Remaining Battery Charge Tracking

While not being, in the strict sense, a fault, tracking the remaining battery charge will be one of the main tasks of the prognostic system. End of charge is an end-of-life criterion,

so the remaining charge estimate is expected to be a factor in most of actions undertaken by PDM software. Most battery-powered devices have some form of battery state-of-charge (SOC) monitoring onboard. This is mostly based on Coulomb counting, i.e. integrating the current drawn over time, divided by the rated capacity of the battery. The definition used in this work is the following:

$$SoC = 1 - \frac{\int_{t=0}^{t|V=V_{cutoff}} I(t)dt}{\text{Capacity of Current Cycle}} \times 100\%$$

It should be noted that both the numerator and denominator of the fraction are predictions, not the actual measurements: battery voltage prediction for the former and capacity prediction for the latter. Further details are discussed in (Saha and Goebel 2009).

3. TESTBED SIMULATOR

As mentioned previously, a simulator has been developed to aid in the design of PDM algorithms for the testbed. It captures both nominal and faulty behavior, with the controlled ability to inject faults. In this way, it serves as a virtual testbed through which algorithms can be initially tested and validated. Faults in the simulator are modeled as undesired changes in system parameters or configuration. In addition to serving as a virtual testbed, the simulator will also be utilized in guiding the decision making process. A graphical user interface was developed for interacting with the simulator, and is shown in Figure 2. In this section, the models used by the simulator are reviewed and some model validation results are presented.

3.1. Rover Dynamics Modeling

The rover consists of a symmetric rigid frame with four independently-driven wheels. Generalized rover coordinates are shown in Figure 3. The F subscript stands for “front”, the B subscript for “back”, the L subscript for “left”, and the R subscript for “right”. The rover pose is given by (x, y, θ) . The independent dynamic variables describing the motion include the body longitudinal velocity v , the body rotational velocity ω , and the wheel rotational velocities ω_{FL} , ω_{FR} , ω_{BL} , and ω_{BR} . Note that the body velocities and wheel velocities are independent due to the presence of slip. Velocity in the lateral direction is negligible (Madow, 2007).

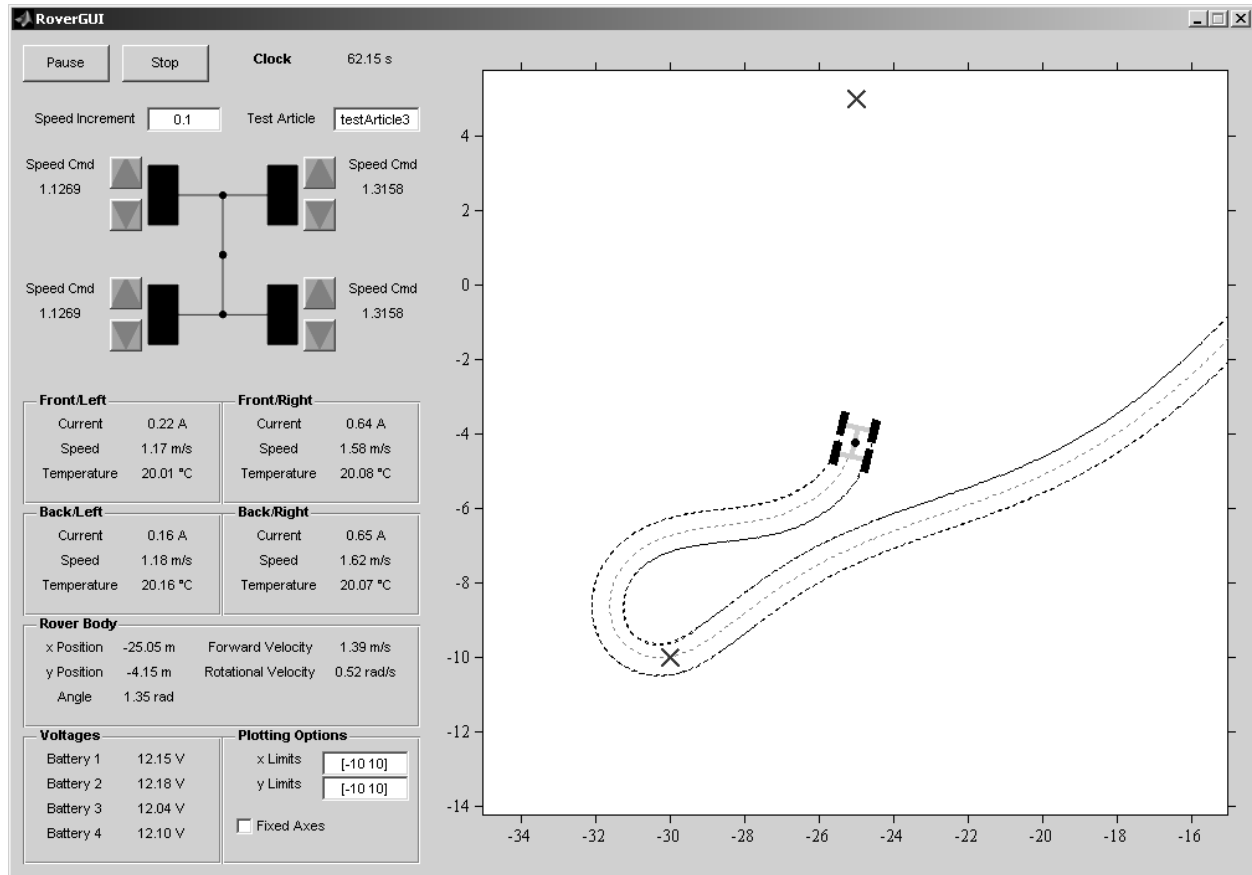


Figure 2: Rover simulation GUI

Since the rover exhibits both longitudinal and rotational velocity, it will experience forces opposing both of these movements. The rover forces are shown in Figure 4. Each motor produces a torque that drives its wheel. When the longitudinal velocity of the rover is equal to the rotational velocity of the wheel times its radius, then there is no slip and no force. Otherwise, some amount of slip will be present and the difference in the relative velocities of the wheel and the ground produce a ground force F_{gl} that pushes the wheel along the ground. These forces are transmitted to the rover body, moving it in the longitudinal direction. The F_{gl} forces produce torques on the rover body, producing a rotation. The rotation is opposed by additional friction forces F_{gr} . The friction forces are defined as:

$$F_{gl} = \mu_g(v_w - v)$$

$$F_{gr} = \mu_r\omega$$

Note that μ_g and μ_r are not in the same units. The F_{gr} forces, opposing the rotation, act at a right angle from the diagonal going from the robot center to the wheel, and in the direction that opposes the rotation. The forward component of this force affects the forward velocity of the rover, just as the component of a F_{gl} force perpendicular to the diagonal

affects the rotational velocity. The angle γ is of interest here, given by

$$\gamma = \arctan l/(2b)$$

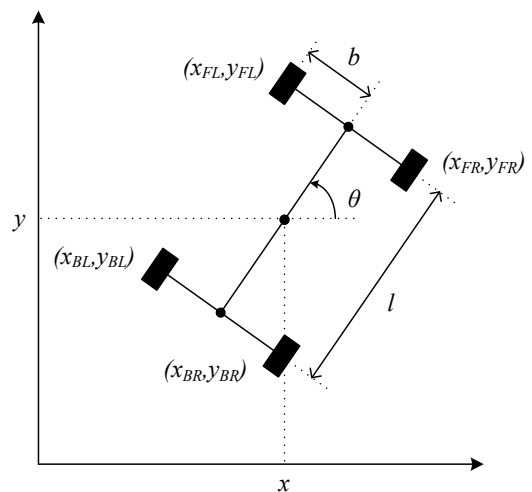


Figure 3: Generalized rover coordinates

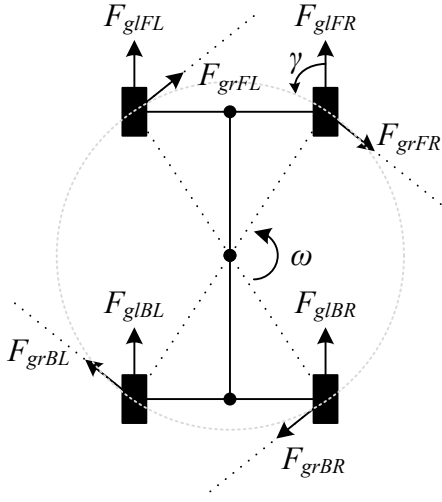


Figure 4: Rover forces

For a given wheel w , the rotational velocity is described by

$$\dot{\omega}_w = \frac{1}{J_w} (\tau_{mw} - \tau_{fw} - r_w F_{glw} + r_w F_{gr} \cos \gamma),$$

where J_w is the wheel inertia, τ_{mw} is the motor torque, and τ_{fw} is the friction torque:

$$\tau_{fw} = \mu_w \omega_w.$$

The forward velocity is described by

$$\dot{v} = \frac{1}{m} (F_{glFL} + F_{glFR} + F_{glBL} + F_{glBR}),$$

assuming that μ_r are the same for all wheels so that the contributions from the F_{gr} forces cancel out. The rotational velocity is described by

$$\dot{\omega} = \frac{1}{J} (d \cos \gamma F_{glFR} + d \cos \gamma F_{glBR} - d \cos \gamma F_{glFL} - d \cos \gamma F_{glBL} - 4d F_{gr})$$

We note that the F_{gl} forces are at distance d from the rover center with the perpendicular component at angle γ . The $\cos \gamma$ factor projects the force onto the tangent of the rotation.

3.2. Motor Modeling

The wheel motors are DC motors with PID control. The DC motor model is given by

$$i_m = \frac{1}{L} (V_w - i_m R - k_w \omega_w)$$

where V_w is the motor voltage, L is the winding inductance, R is the winding resistance, and k_w is an energy transformation term. The motor torque given by

$$\tau_m = k_\tau i_m$$

where k_τ is an energy transformation term.

Increased motor/wheel friction for wheel w is captured by an increase in μ_w . A change in motor resistance is captured by a change in R . The motors windings are designed to withstand temperatures up to a certain point, after which, the insulation breaks down, the windings short, and the motor fails. It is therefore important to model the temperature behavior of the motor.

The motor thermocouple is located on the motor surface. The surface loses heat to the environment and is heated indirectly by the windings, which, in turn, are heated up by the current passing through them. The temperature of the windings is given by

$$\dot{T}_w = \frac{1}{J_w} (i^2 R - h_w (T_w - T_m)),$$

where J_w is the thermal inertia of the windings, h_w is a heat transfer coefficient, and T_m is the motor surface temperature (Balaban, et al., 2009). It is assumed that heat is lost only to the motor surface, and that winding resistance R is approximately constant for the temperature range considered. The surface temperature is given by

$$\dot{T}_m = \frac{1}{J_s} (h_w (T_w - T_m) - h_a (T_m - T_a))$$

where J_s is the thermal inertia of the motor surface, h_a is a heat transfer coefficient, and T_a is the ambient temperature. Heat is transferred from the windings to the surface and lost to the environment.

This model was validated for DC motors using experimental data collected on the Flyable Electro-Mechanical Actuator (FLEA) testbed (Balaban, Saxena, Narasimhan, Roychoudhury, Goebel, & Koopmans, 2010). The unknown parameters J_w, J_s, h_w, h_a , and R were identified to match data acquired from a scenario where the motor was overloaded and, as a result, heated up considerably. The motor current and surface temperatures were measured. A comparison of predicted vs. measured temperature is shown in Figure 5.

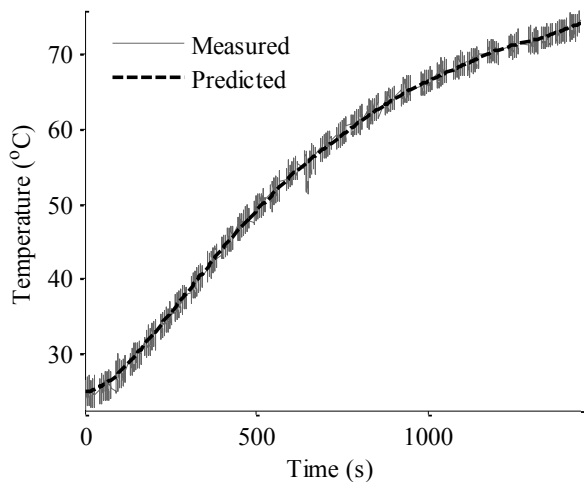


Figure 5: Comparison of measured and model-predicted motor surface temperature for a DC motor

3.3. Sensor Fault Modeling

Sensor faults are captured with bias, drift, gain, and scaling terms. Ranges for typical fault magnitude values have been identified through a literature search and discussions with manufacturers (Balaban, Saxena, Bansal, Goebel, & Curran, 2009). Faults in common sensors such as current, voltage, temperature, and position will be modeled.

3.4. Battery Modeling

The key challenge in modeling a battery is estimating its open-circuit voltage, E_0 . The theoretical open-circuit voltage of a battery is traditionally assessed when all reactants are at 25°C and at 1M concentration (or 1 atm pressure). However, this voltage cannot be measured directly during battery use due to the influence of internal passive components such as the electrolyte, the separator, and the terminal leads. The measured voltage will be lower; the factors contributing to the voltage drop are characterized in the following paragraphs.

The first factor considered is the *ohmic drop*. The term refers to the diffusion process through which Li-ions migrate to the cathode via the electrolytic medium. The internal resistance to this ionic diffusion process can also be referred to as the IR drop. For a given load current, this drop usually decreases with time due to the increase in internal temperature, which results in increased ion mobility.

The next factor is *self-discharge*, which is caused by the residual ionic and electronic flow through a cell even when there is no external current being drawn. The resulting drop in voltage has been modeled to represent the activation polarization of the battery. All chemical reactions have a certain activation barrier that must be overcome in order for

the reaction to proceed and the energy needed to overcome this barrier leads to the activation polarization voltage drop. The dynamics of this process are described by the Butler–Volmer equation, which, in this work, is approximated by a logarithmic function.

Concentration polarization is the voltage loss due to spatial variations in reactant concentration at the electrodes. This occurs primarily when the reactants are consumed faster by the electrochemical reaction than they can diffuse into the porous electrode. The phenomenon can also occur due to variations in bulk flow composition. The consumption of Li-ions causes a drop in their concentration along the cell, which causes a drop in the local potential near the cathode. The magnitude of concentration polarization is usually low during the initial part of the discharge cycle, but grows rapidly towards the end of it or when the load current increases.

Finally, the degradation of battery capacity with aging, as encapsulated by the cycle life parameter, can be modeled by the concept of *Coulombic efficiency*, η_c , defined as the fraction of the prior charge capacity that is available during the following discharge cycle (Huggins, 2008). As mentioned previously, this depends upon a number of factors, particularly on current and depth of discharge in each cycle. The temperature at which the batteries are stored and operated under also has a significant effect on the Coulombic efficiency. For further details on battery modeling, please refer to (Saha and Goebel 2009).

3.5. Electronics Fault Modeling

The field of electronics prognostics is relatively new compared to prognostics for mechanical systems. As a result, research efforts to develop physics-based degradation models that take into account loading and operational conditions are in their early stages. There are several well-known electronics reliability models that deal with failure rates under specific stress factors and corresponding failure mechanisms. However, such models do not take into account usage time, thus making them less suitable for prediction of remaining useful life.

Empirical degradation models of IGBTs, based on the turn-off tail of the drain current, have recently been used for prediction of their future health state (Saha B., Celaya, Wysocki, & Goebel, 2009). Their collector-emitter voltage has been used as precursor of failure as well (Patil, 2009).

In the case of power MOSFETs, the on-state drain to source resistance has been identified as a precursor to failure for the die-attach failure mechanism (Celaya, Saxena, Wysocki, Saha, & Goebel, 2010). For gate-related failure, empirical degradation models based on the exponential function have also been developed (Saha S., Celaya, Vashchenko, Mahiuddin, & Goebel, 2011).

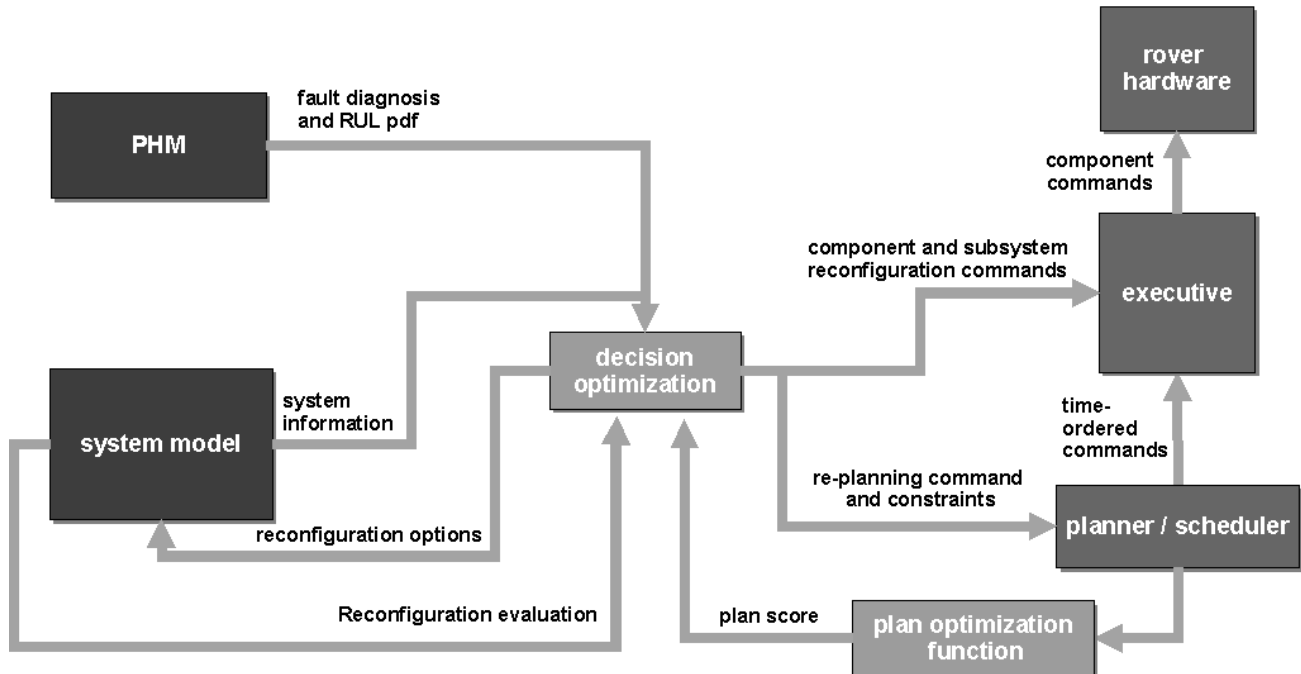


Figure 6: Autonomy architecture

For electrolytic capacitors, an empirical degradation model, also based on an exponential function is presented in (Celaya, Saxena, Vaschenko, Saha, & Goebel, 2011). This model is built based on accelerated degradation data. Details on its physical foundations are presented in (Kulkarni, Biswas, Celaya, & Goebel, 2011).

4. PROGNOSTICS-ENABLED DECISION MAKING ALGORITHM DEVELOPMENT

This section describes the reasoning architecture currently being deployed on the K11 and is meant to mainly provide an overview of the types of algorithms that can be plugged into it for testing and comparison. The current set of algorithms is expected to evolve as this research progresses and other organizations become involved. Figure 6 outlines the architecture and depicts the information flow among its components. The Prognostic Health Management (PHM) element on the figure combines diagnostic and prognostic reasoning engines. If a system fault occurs, the diagnostic engine is tasked with detecting it and identifying what it is, followed by invocation of an appropriate prognostic algorithm to track fault progression. Once a prognosis of the remaining useful life is made, the information is passed to the decision optimization module, which identifies the best way to respond. The K11 simulator, with its nominal and fault-progression models, is used to guide the decision optimization process in some of the cases. The response chosen may involve reconfiguration of low-level controllers or requesting the planner to come up with a new mission plan. Planner output is used to generate action schedules and

then, through the executive module, time-ordered commands for individual components.

4.1. Diagnostics

Diagnosis can be defined as the process of detecting, isolating, and identifying faults in the system. A fault is defined as an undesired change in the system that causes the system to deviate from its nominal operation regime. Diagnostic approaches can be broadly divided into two types: model-based and data-driven (Gertler, 1998). Model-based methods rely on a system model built from *a priori* knowledge about the system. Data-driven methods, on the other hand, do not require such models but instead require large, diverse sets of exemplar failure data, which are often not available. The decision of whether to adopt a model-based or a data-driven diagnostic approach depends on the sensor suite properties and the fault modes of interest, among other factors.

Currently a model-based approach is adopted for providing a diagnostic system for the rover, as the sensors (Table 1) and fault modes (Table 2) lend themselves to physics-based modeling. Once sensors measuring more complex dynamics (e.g. accelerometers) are added to the system, data-driven diagnosis methods may be required. Additionally, model-based and data-driven algorithms can be synergistically combined to improve upon either approach implemented individually (Narasimhan, Roychoudhury, Balaban, & Saxena, 2010).

Typically, model-based methods require a nominal system model, as well as a model that captures the relations between faults and symptoms. The overall goal is to use the model to generate estimates of nominal system behavior, then compare them with observed system behavior. If any deviation from nominal, i.e., a symptom, is observed, the fault-symptom model is used to isolate the fault modes that explain the symptoms.

A model-based diagnosis approach is generally split into three tasks: fault detection, fault isolation, and fault identification, with the following event flow:

- Fault detection involves determining if the system behavior has deviated from nominal due to the occurrence of one or more faults. The fault detector takes as inputs the measurement readings, y , and the expected nominal measurement values, \hat{y} , generated by the nominal system observer. The detector indicates a fault if the residual, $r = y - \hat{y}$, is statistically significant.
- Once a fault is detected, the fault isolation module generates a set of fault hypotheses, F , and, at every time step, reasons about what faults are consistent with the sequence of observed measurements in order to reduce F . The goal of fault isolation is to reduce F to as small a set as possible. If only single faults are assumed then, ideally, the goal of fault isolation is to reduce F to a singleton set.
- Once the fault (or faults) are isolated, fault identification is invoked. It involves quantitatively evaluating the magnitude of each fault, $f \in F$.

Once the fault magnitude is identified, prognostic algorithms can be invoked to predict how the damage grows over time and estimate the remaining useful life of the affected component and the overall system.

4.2. Prognostics

For the purposes of this research, prognostics is defined as the process which predicts the time when a system variable or vector indicating system health no longer falls within the limits set forth by the system specifications (End-of-Life or EOL). The prediction is based on proposed future usage. In some cases the trajectory of the aforementioned variable or vector through time is predicted as well. Similarly to diagnostic methods, prognostics methods are generally classified as either data-driven or model-based: (Schwabacher, 2005); (Saha and Goebel 2009); (Daigle & Goebel, 2011).

Generally, the inputs to a prognostic algorithm include information on the fault provided by the diagnostic algorithm (e.g. fault type, time and magnitude). Output of a prognostic algorithm could be then presented to a PDM algorithm in one of the following ways:

- as an estimate $\phi^q_{t_j, t_i, L}$ of the variable of interest (e.g. accumulated damage or remaining life) at a specific time t_j given the information up to time t_i for a component q , where $t_j > t_i$. L is the anticipated average load up to t_j .
- As a discrete point trajectory $\Phi^q_{i, L}(j)$ given information up to the point i , where $L = \{l_1, l_2, \dots, l_{EoP}\}$ are the anticipated load values for each point on the

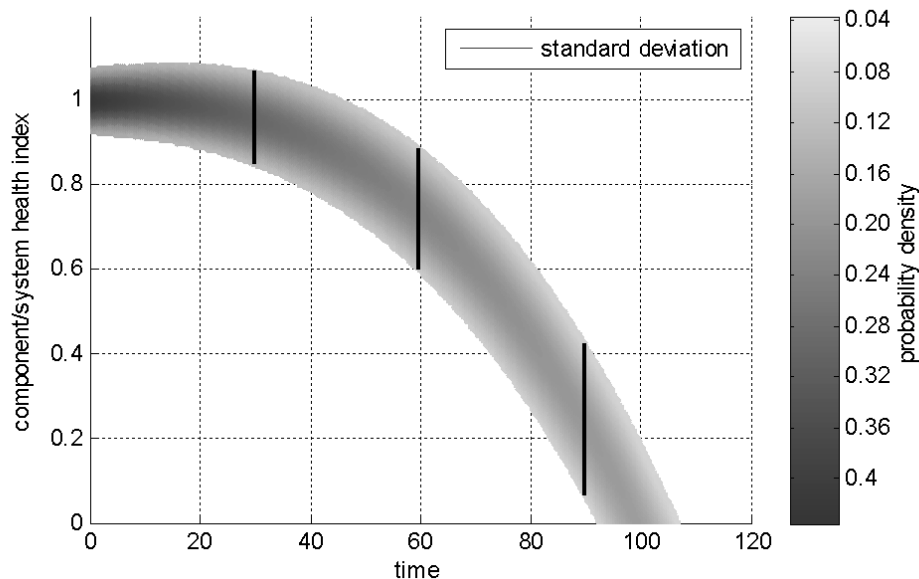


Figure 7. Prognostic prediction example

prediction trajectory, EOP is the end of prediction index, and $I < j < EOP$.

- c. As a continuous function $\Phi_{t_i, L}^q(t_j)$ given information up to a time t_i and an anticipated load function $L(t)$

The estimate produced in all of the above cases may be expressed as a probability density function (pdf) or as moments derived from the probability distribution. Each of the three options assumes time-variability of the prognostic function, which is one of the main factors that make PDM an interesting and challenging research problem. The function may change from one time of prediction to the next as more information about the behavior of the system becomes available.

Figure 7 shows the important features of an example prediction curve produced at a specific time t . The points on the time axis are relative to the moment of prediction. The health index values are normalized to be in the $[0, 1]$ interval. Probability density of health index values for each point in time is illustrated using a grayscale map (shown on the right side of the figure). The solid black bars are drawn to show one standard deviation of probability distributions at different time points into the future. End-of-Life is a time value corresponding to the health index chosen to indicate that the component or a system can no longer provide useful performance. In this example EOL corresponds to health index of 0, however the threshold can be defined as any other value in the $[0, 1]$ interval.

The prediction step requires knowledge of the future usage of the system. For the rover, this involves the expected future trajectory and environmental inputs, such as the terrain and the ambient temperature. The physics models developed for the simulator can be utilized in both the estimation and prediction phases. Damage progression processes that are difficult to model may require use of data-driven prognostics methods. In the remainder of this section, more details are provided on the prognostic methods currently investigated for the fault modes of interest.

4.2.1. Mechanical Jam/Windings Insulation Deterioration

The thermal build-up model as described in the simulator section will be used to predict when the interior of a motor would reach the temperature at which insulation of the windings is likely to melt, thus disabling the motor. A machine-learning prediction method will also be utilized for comparison. The method is based on the Gaussian Process Regression (GPR) principles and was previously tested on another testbed developed at NASA Ames, the FLEA (Balaban, Saxena, Narasimhan, Roychoudhury, & Goebel, Experimental Validation of a Prognostic Health Management System for Electro-Mechanical Actuators,

2011). The FLEA was used to inject and collect data progression of the same type of fault in the motors of electro-mechanical actuators. Several motors were run to complete failure and GPR demonstrated a high accuracy in predicting their remaining useful life.

GPR does not need explicit fault growth models and can be made computationally less expensive by sampling techniques. Further, it provides variance bounds around the predicted trajectory to represent the associated confidence (Rasmussen & Williams, 2006). Domain knowledge available from the process is encoded by the covariance function that defines the relationship between data points in a time series. In the present implementation, a Neural Network type covariance function is used.

Sensor data is processed in real-time to extract relevant features, which are used by the GPR algorithm for training during the initial period. The longer is the training period, the better are the chances for the algorithm to learn the true fault growth characteristics. However, to strike a balance between the length of the training period and the risk of producing an insufficient prediction horizon, a limit for the training period is set. Once this limit is reached, the algorithm starts predicting fault growth trajectories. EOL is subsequently determined by where these trajectories intersect the predetermined fault level threshold. As time progresses, the GPR model is updated with new observations and, subsequently, the predictions are updated as well. Best fitting hyper-parameters for the covariance function are determined via a maximum-likelihood optimization. The uncertainty created by this process is handled by drawing a large ($p \sim 50$) number of samples from the observed data at each prediction instance (t_p) and training p different Gaussian Process models on these distinct data sets. Their prediction results are then averaged.

4.2.2. Battery Capacity Deterioration and Charge Tracking

There are several methods widely in use for batteries that relate capacity and SOC to the number of cycles a battery has undergone and its open circuit voltage. Most such methods, however, are reliability based, i.e. they assume certain usage profiles are maintained throughout the cycle life of the battery. Such assumptions (for instance that the battery undergoes full discharge followed by full charge repeatedly until its end-of-life) are not always realistic. This is especially true for a platform such as the K11, where individual missions may have different goals with different load profiles. Under such circumstances, it is advantageous to model the internal processes of the battery and let a Bayesian inference framework, such as the Particle Filter (Arulampalam, Maskell, Gordon, & Clapp, 2002) manage the uncertainty in the model (Saha & Goebel, 2009)

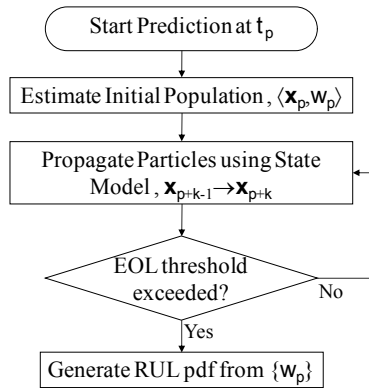


Figure 8: Prediction flowchart

In the remaining battery charge prediction application, the main state variable is the battery voltage, E , and in the objective is to predict when E reaches the low-voltage cutoff. During the prognostic process a tracking routine is run until a long-term prediction is required, at which point the state model is used to propagate the last updated particles until the end-of-discharge (EOD) voltage cutoff is reached. The RUL pdf is computed from the weighted RUL predictions of the particles in the last tracking step. Figure 8 shows the flow diagram of the prediction process.

In the case of the overall battery life cycle, the main variable that needs to be tracked is the capacity of the battery itself, and the goal is to predict when the battery capacity will fade by more than 20% from when it was new. At that point the battery is said to have reached its EOL. For a description of the main physical phenomena behind these processes, please refer to Section 3.2. A battery lifecycle model and a Particle Filter algorithm utilizing it are presented in (Saha & Goebel, 2009) and (Saha, et al., 2011).

4.2.3. Electronics Faults

Previously researched data-driven and model-based (direct physics and empirical/Bayesian) techniques are being utilized for addressing electronics faults. The particle filter approach has been used in conjunction with an empirical degradation model for IGBTs experiencing failures related to thermal overstress (Saha B. , Celaya, Wysocki, & Goebel, 2009). For power MOSFETs, a data-driven prognostics approach based on Gaussian Process Regression has recently been implemented for die-attach degradation (Celaya, Saxena, Vaschenko, Saha, & Goebel, 2011). For electrolytic capacitors, a remaining useful life prediction based on a Kalman filter has been developed using a

degradation model based on an empirical exponential function (Celaya, Kulkarni, Biswas, & Goebel, 2011). It should be noted that the aforementioned efforts make the assumption of usage levels and operational conditions staying constant in the future. New accelerated aging experiments aimed at producing degradation models without these limitations are currently underway.

4.3. Decision Making

As stated previously, one of the main objectives of the K11 testbed is to investigate PDM algorithms in order to enhance an aerospace vehicle's capability to achieve its high-level goals – be it under a faulty condition, degraded operation of a subsystem, or an anticipated catastrophic failure. There has been an increasing amount of research conducted over the last several years in prognostic methodologies for various types of components or systems. The effort described in this paper aims to bring more attention to the “management” aspect of prognostic health management, i.e. what could be done after a fault is detected and the trajectory of its progression is predicted.

Several factors are being used to select the appropriate system level (or levels) on which to respond to an off-nominal condition. These factors include the severity of a fault, its criticality, and predicted time-to-failure interval. A faulty electronic component in an electric motor driver could prompt the decision-making system to trigger a controller reconfiguration - so as to ensure the dynamic stability of the system and a certain level of retained performance. At a different level, a control effort reallocation can be triggered by a supervisory mid-level controller in order to reduce the torque required from a faulty drive motor and compensate for the reduction with the other motors. Reallocating the load could, potentially, extend the remaining useful life of the affected component long enough to ensure achievement of the mission objectives. At the highest level, the rover mission can be re-planned based on prognostic health information so as to achieve maximum possible utility and safety. The above examples call on different system components in their response; there are, however, commonalities for all of them. There is always an objective (or a set of objectives) to be met and a series of actions to be selected by the decision making process in order to meet those objectives. Therefore, the decision making process is, essentially, an optimization process which tries to achieve specified objectives by considering system performance and health constraints.

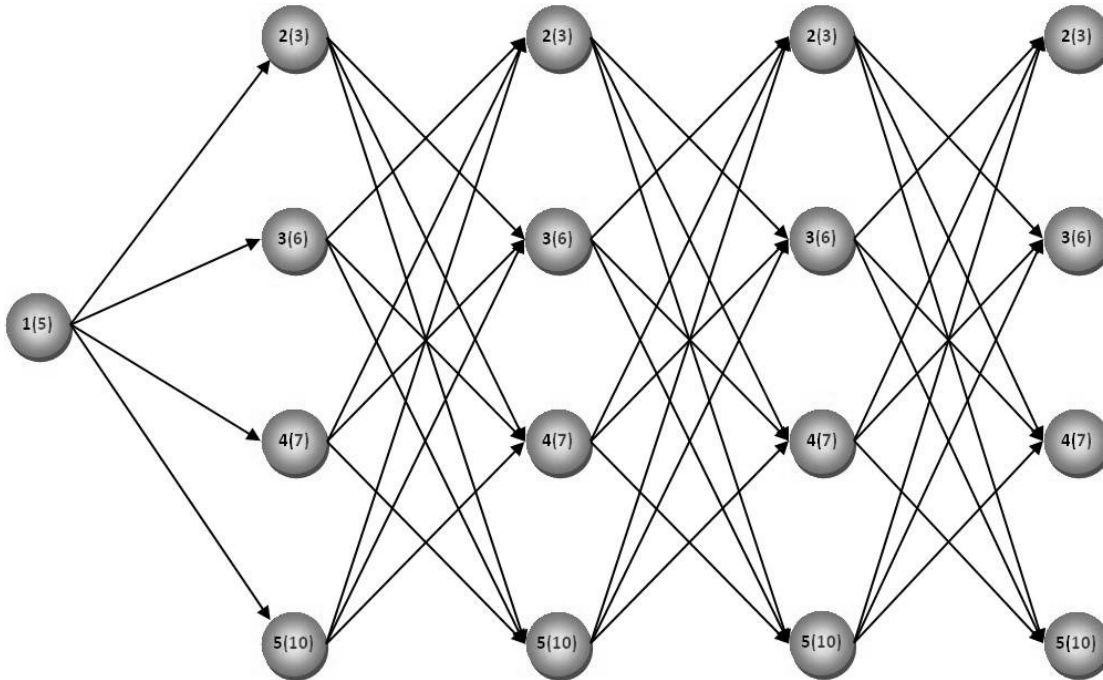


Figure 9. Decision optimization solution space

The scope for the decision-making module in the current implementation is defined as the following: getting vehicle health information from the prognostic health reasoner and the simulator, the decision-making module evaluates the best course of action to take (e.g., controller reconfiguration or mission replanning), while stopping short of performing the actual reconfiguration or re-planning. Instead, the decision-making module adjusts goals and constraints for other software components. To use the planner as an example, the module could set a constraint on rover speed or on the total mission duration, then request the planner to come up with a detailed new plan. In the future, however, it may be necessary to consider whether making PDM-specific modifications to the planner or the adaptive controller, for instance, would improve performance.

The rover is assumed to be an autonomous planetary vehicle, operating with only high level control from human operators (initial sets of goals and constraints). The following definitions and assumptions are used in the current phase of the work:

- The initial sets of goals $G = \{g_1, g_2, \dots, g_N\}$ and constraint variables $K = \{k_1, k_2, \dots, k_M\}$ are provided as inputs.
- The initial constraint ranges are: $\forall k_j \in K, \exists [a_j^0, a_j^1], k_j \in [a_j^0, a_j^1]$. The constraint ranges are adjusted given information from the diagnostic, prognostic, and decision optimization routines.
- Some elements of G may be eliminated as a result of the optimization process. The size of K (M) will remain constant.
- Goal rewards: $\forall g_i \in G, \exists r_i \in [0, r_{max}]$. r_{max} is the maximum possible value of goal reward.
- Goal locations: $\forall g_i \in G, \exists l(x_1, x_2, x_3, \theta_1, \theta_2, \theta_3), l \in L$. The preceding location definition is general for a three-dimensional space. In the case of a rover it simplifies to $l(x_1, x_2, \theta_1)$.
- Constraint ranges: $\forall k_j \in K, \exists [a_j^0, a_j^1], k_j \in [a_j^0, a_j^1]$
- Transition cost: $\forall (l_i, l_j) \exists \bar{c}$, where $\bar{c} = \{c_e, c_h\}$, c_e is the energy cost and c_h is the health (life) cost for the system. The former is calculated based on the distance between the goal locations, proposed velocity, and the load index (terrain difficulty). The later is estimated using the health prognostic function, which takes the distance, velocity, and the load index as its inputs.
- Mission starts with energy amount E_0 available
- $E(t)$ and $H(t)$ are the energy and system health 'amounts' at time t , respectively. $E(t)=0$ or $H(t)=0$ constitutes an end-of-life (EOL) condition
- The objective is to accumulate the maximum possible reward before energy and health budgets are exceeded

Two PDM approaches are currently implemented and verified in a simulation involving a small number of goals: Dynamic Programming (DP) and a variant of Probability Collectives (PC) (Wolpert, 2006). A (more computationally expensive) exhaustive search method is used for verification of the DP and PC algorithms. A simple five-goal example is presented next for illustration purposes.

Each of the nodes (goals) is associated with a reward value in the parenthesis (

Figure 9). The vehicle starts out at goal 1, but can choose to traverse the rest of the goal in any order. For some of the degradation modes system health cost may correlate to the energy cost (which, in turn, could be proportional to the sum of loads on the system, whether they are mechanical or electrical). Motor windings insulation deterioration due to increased friction in the system could be one such example. One the other hand, deterioration of an electronic component in one of the subsystems may be determined primarily by the amount of current flowing through that component, ambient temperature, and time.

The implemented DP algorithm uses forward propagation, evaluating the best solution for transitioning from stage to stage, while assuming optimality of the previously made decisions. An ‘efficiency index’ is used for guiding the stage-to-stage decision process:

$$e_{ij} = r_j / (\sum \bar{c}_{ij})$$

If either health or energy values become less or equal to zero (or all the nodes are visited), the forward propagation phase is stopped. After the forward traverse is completed, an optimal path p_{opt} is ‘backed-out’ by traversing the stages in the opposite (right-to-left) direction, starting with the node associated with the highest accumulated reward.

The algorithm based on Probabilities Collectives principles is structured in the following manner:

- P is defined as the enumerated set of all possible paths p
- An initial probability distribution $f(p)$ for P is assigned
- ‘Related’ paths are defined as those that share $n \in [2, N]$ initial nodes, with n incremented progressively every m iterations of the algorithm
- P is sampled using $f(p)$, obtaining a sample path p_i . The cumulative path reward is evaluated by ‘walking’ the sampled path and taking into account energy and health budgets. If the reward is equal or greater than the current maximum, the probability of the sampled path **and** paths that are ‘related’ to it are increased and P is re-normalized.

Uncertainty in transition costs and node rewards is incorporated by associating them with probability distributions as well. These distributions are then sampled when ‘walking’ a path during its evaluation.

Experimenting with DP- and PC-based algorithms showed that both would work well on relatively uncomplicated problems, such as the one described in this section. Limitations of the two approaches started to become evident as well, however. A DP implementation will become more challenging if will multi-objective problems are posed (i.e. optimization over component(s) RUL in addition to cumulative mission reward is desired), unless the multiple optimization variables lend themselves to being aggregated into a single ‘composite’ variable. The PC-based method, on the other hand, will likely have a lower limit on the size of the goal set it can practically process, at least in its current form. Nevertheless, PC appears to be well-suited for the problem of optimizing system parameters and constraints for maximum RUL, which is being pursued next.

4.4. Task Planning and Execution

Once the high level goals and constraints are determined by the prognostics-enabled decision making module, the detailed task planning for the rover will be generated using NASA’s Extensible Universal Remote Operations Architecture (EUROPA) (Frank & Jonsson, 2003). EUROPA provides the capability of solving task planning, scheduling, and constraint-programming problems.. In a complex system, such as a rover, scheduling specific tasks to be executed is often a non-trivial problem. There are resources that are shared by different processes that may not necessarily be available at all times, so EUROPA supports generation of a schedule of activities. Plans and schedules generated by EUROPA (either nominal or those generated in response to a fault) will be passed for automated execution via Plan Execution Interchange Language (PLEXIL) (Dalal, et al., 2007).

5. CONCLUSIONS

The work described in this paper is aimed at providing an inexpensive, safe platform for development, validation, evaluation, and comparison of prognostics-enabled decision making algorithms. Technologies resulting from this research are planned to be transferred for further maturation on unmanned aerial vehicles and other complex systems. At present, the K11 testbed already constitutes a promising platform for PDM research. A list of fault modes of interest has been identified and a number of them have already been implemented in software and/or hardware. A software simulator has been developed that incorporates models of both nominal and off-nominal behavior, with some of the models verified using experimental data. The software architecture for the testbed has been defined in such a way as to allow quick replacement of autonomy elements

depending on testing objectives and customers. The first set of reasoning algorithms, developed at NASA Ames, is being deployed.

Plans for the near future include addition of further injectable fault modes, field experiments of greater complexity, simulator model refinement, and extension of PDM methods to handle more complex problems, including constraints adjustment for optimal RUL. Data collected on the testbed is planned for distribution to other researchers in the field.

ACKNOWLEDGEMENT

Student interns George Gorospe, Zachary Ballard, Sterling Clarke, Tabitha Smith and Kevin Rooney have contributed tremendously to this effort through their creativity and hard work. The team also gratefully acknowledges all the advice and assistance from the Intelligent Robotics Group (Terry Fong, Liam Pedersen, Vinh To, Susan Lee, Hans Utz, Lorenzo Fluckiger, Maria Bulaat, Vytas SunSpiral, and others). Jeremy Frank and Michael Dalal of the Planning & Scheduling Group have also been very generous with their time, providing expertise on planning, scheduling, and automated execution technologies. Many ideas in this work have been borne out of discussions with colleagues at Impact Technologies (Liang Tang, Eric Hettler, Bin Zhang, and Jonathan DeCastro) who are actively collaborating with NASA Ames on prognostics-enabled automated contingency management, testbeds, and other topics. The funding for this research is provided by NASA ARMD System-wide Safety & Assurance Technology (SSAT) project.

NOMENCLATURE

ARMD	Aeronautics Research Mission Directorate
CORBA	Common Object Request Broker Architecture
DDS	Data Distribution Service
DM	Decision Making
EOD	End Of Discharge
EOL	End Of (Useful) Life
EUROPA	Extensible Universal Remote Operations Architecture
GPR	Gaussian Progress Regression
GPS	Global Positioning System
IGBT	Insulated Gate Bi-polar Transistor
LiFePO4	Lithium Iron Phosphate
MOSFET	Metal Oxide Semiconductor Field Effect Transistor
PDM	Prognostics-enabled Decision Making
PHM	Prognostics and Health Management
PLEXIL	Plan Execution Interchange Language
RUL	Remaining Useful Life
SOC	State Of Charge

REFERENCES

- Arulampalam, M., Maskell, S., Gordon, N., & Clapp, T. (2002). A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking. *IEEE Transactions on Signal Processing*, 50 (2), 174-189.
- Balaban, E., Saxena, A., Bansal, P., Goebel, K., & Curran, S. (2009). Modeling, Detection, and Disambiguation of Sensor Faults for Aerospace Applications. *IEEE Sensors Journal*, 9 (12), 1907 - 1917.
- Balaban, E., Saxena, A., Goebel, K., Byington, C., Watson, M., Bharadwaj, S., et al. (2009). Experimental Data Collection and Modeling for Nominal and Fault Conditions on Electro-Mechanical Actuators. *Annual Conference of the Prognostics and Health Management Society*. San Diego, CA.
- Balaban, E., Saxena, A., Narasimhan, S., Roychoudhury, I., & Goebel, K. (2011). Experimental Validation of a Prognostic Health Management System for Electro-Mechanical Actuators. *AIAA Infotech@Aerospace*.
- Balaban, E., Saxena, A., Narasimhan, S., Roychoudhury, I., Goebel, K., & Koopmans, M. (2010). Airborne Electro-Mechanical Actuator Test Stand for Development of Prognostic Health Management Systems. *Annual Conference of the Prognostics and Health Management Society*. San Diego, CA.
- Celaya, J., Kulkarni, C., Biswas, G., & Goebel, K. (2011). Towards Prognostics of Electrolytic Capacitors. *AIAA Infotech@Aerospace*. St. Louis, MO.
- Celaya, J., Saxena, A., Vaschenko, V., Saha, S., & Goebel, K. (2011). Prognostics of power MOSFETs. *23rd International Symposium on Power Semiconductor Devices and ICS*. San Diego, CA.
- Celaya, J., Saxena, A., Wysocki, P., Saha, S., & Goebel, K. (2010). Towards Prognostics of Power MOSFETs: Accelerated Aging and Precursors of Failure. *Annual Conference of the Prognostics and Health Management Society*. Portland, OR.
- Daigle, M., & Goebel, K. (2011). Multiple Damage Progression Paths in Model-based Prognostics. *IEEE Aerospace Conference*. Big Sky, Montana.
- Dalal, M., Estlin, T., Fry, C., Iatauro, M., Harris, R., Jonsson, A., et al. (2007). *Plan Execution Interchange Language (PLEXIL)*. Moffett Field: NASA Ames Research Center.
- Fluckiger, L., To, V., & Utz, H. (2008). Service oriented robotic architecture supporting a lunar analog test. *International Symposium on Artificial Intelligence, Robotics, and Automation in Space*. Los Angeles, CA.
- Frank, J., & Jonsson, A. (2003). Constraint-Based Attribute and Interval Planning. *Journal of Constraints, Special Issue on Constraints and Planning*, 8 (4), 335-338.
- Gertler, J. (1998). *Fault Detection and Diagnosis in Engineering Systems*. New York: Marcel Dekker Inc.
- Huggins, R. (2008). *Advanced Batteries: Materials Science Aspects* (1st Edition ed.). Springer.
- Kulkarni, C., Biswas, G., Celaya, J., & Goebel, K. (2011). Prognostic Techniques for Capacitor Degradation and

- Health Monitoring. *Maintenance & Reliability Conference*. Knoxville, TN.
- Kulkarni, C., Biswas, G., Koutsoukos, X., Celaya, J., & Goebel, K. (2010). Aging Methodologies and Prognostic Health Management for Electrolytic Capacitors. *Annual Conference of the PHM Society*. Portland, OR.
- Lachat, D., Krebs, A., Thueer, T., & Siegwart, R. (2006). Antarctica Rover Design and Optimization for Limited Power Consumption. *MECHATRONICS - 4th IFAC-Symposium on Mechatronic Systems*.
- Madow, A. M.-C. (2007). Experimental kinematics for wheeled skid-steer mobile robots. *Intelligent Robots and Systems, 2007 (IROS 2007)*. *IEEE/RSJ International Conference on*, (pp. 1222-1227).
- Narasimhan, S., Roychoudhury, I., Balaban, E., & Saxena, A. (2010). Combining Model-based and Feature-driven Diagnosis Approaches - A Case Study on Electromechanic Actuators. *21st International Workshop on Principles of Diagnosis (DX 10)*. Portland, Oregon.
- NASA Ames Research Center. (2011). *The Robot Application Programming Interface Delegate Project*. Retrieved from <http://rapid.nasa.gov/>
- Object Management Group. (2004). *CORBA/IIOP specification*. Framingham, MA: OMG.
- Patil, N. C. (2009). Precursor parameter identification for insulated gate bipolar transistor (IGBT) prognostics. *IEEE Transactions on Reliability*, 58(2), 271-276.
- Poll, S., Patterson-Hine, A., Camisa, J., Nishikawa, D., Spirkovska, L., Garcia, D., et al. (2007). Evaluation, Selection, and Application of Model-Based Diagnosis Tools and Approaches. *AIAA Infotech@Aerospace Conference and Exhibit*. Rohnert Park, CA.
- Rasmussen, C., & Williams, C. (2006). *Gaussian Processes for Machine Learning*. Boston, MA: The MIT Press.
- Saha, B., & Goebel, K. (2009). Modeling Li-ion Battery Capacity Depletion in a Particle Filtering Framework. *Proceedings of Annual Conference of the PHM Society*. San Diego, CA.
- Saha, B., Celaya, J., Wysocki, P., & Goebel, K. (2009). Towards Prognostics for Electronics Components. *IEEE Aerospace*, (pp. 1-7). Big Sky, MT.
- Saha, B., Koshimoto, E., Quach, C., Hogge, E., Strom, T., Hill, B., et al. (2011). Predicting Battery Life for Electric UAVs. *AIAA Infotech@Aerospace*.
- Saha, S., Celaya, J., Vashchenko, V., Mahiuddin, S., & Goebel, K. (2011). Accelerated Aging with Electrical Overstress and Prognostics for Power MOSFETs. *IEEE EnergyTech, submitted*.
- Schmidt, D. (1994). The ADAPTIVE Communication Environment: Object-Oriented network programming components for developng client/server applications. *Proceedings of the 12 th Annual Sun Users Group Conference* (pp. 214–225). San Francisco, CA: SUG.
- Schwabacher, M. (2005). A Survey of Data-drive Prognostics. *AIAA Infotech @ Aerospace Conference*.
- Smith, M., Byington, C., Watson, M., Bharadwaj, S., Swerdon, G., Goebel, K., et al. (2009). Experimental and Analytical Development of a Health Management System for Electro-Mechanical Actuators. *IEEE Aerospace Conference*. Big Sky, MT.
- Wolpert, D. (2006). Information Theory - The Bridge Connecting Bounded Rational Game Theory and Statistical Physics. (D. Braha, A. Minai, & Y. Bar-Yam, Eds.) *Complex Engineered Systems*, 14, 262-290.

A Model-based Prognostics Methodology for Electrolytic Capacitors Based on Electrical Overstress Accelerated Aging

José R. Celaya¹, Chetan Kulkarni², Gautam Biswas³, Sankalita Saha⁴, Kai Goebel⁵

¹ *SGT Inc. NASA Ames Research Center, Moffett Field, CA, 94035, USA*
jose.r.celaya@nasa.gov

^{2,3} *Vanderbilt University, Nashville, TN, 37235, USA*
chetan.kulkarni@vanderbilt.edu
biswas@eecsmail.vuse.vanderbilt.edu

⁴ *MCT. NASA Ames Research Center, Moffett Field, CA, 94035, USA*
sankalita.saha@nasa.gov

⁵ *NASA Ames Research Center, Moffett Field, CA, 94035, USA*
kai.goebel@nasa.gov

ABSTRACT

A remaining useful life prediction methodology for electrolytic capacitors is presented. This methodology is based on the Kalman filter framework and an empirical degradation model. Electrolytic capacitors are used in several applications ranging from power supplies on critical avionics equipment to power drivers for electro-mechanical actuators. These devices are known for their comparatively low reliability and given their criticality in electronics subsystems they are a good candidate for component level prognostics and health management. Prognostics provides a way to assess remaining useful life of a capacitor based on its current state of health and its anticipated future usage and operational conditions. We present here also, experimental results of an accelerated aging test under electrical stresses. The data obtained in this test form the basis for a remaining life prediction algorithm where a model of the degradation process is suggested. This preliminary remaining life prediction algorithm serves as a demonstration of how prognostics methodologies could be used for electrolytic capacitors. In addition, the use degradation progression data from accelerated aging, provides an avenue for validation of applications of the Kalman filter based prognostics methods typically used for remaining useful life predictions in other applications.

1. INTRODUCTION

This paper proposes the use of a model based prognostics approach for electrolytic capacitors. Electrolytic capacitors

Celaya et.al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

have become critical components in electronics systems in aeronautics and other domains. This type of capacitors is known for its low reliability and frequent breakdown in critical systems like power supplies of avionics equipment and electrical drivers of electro-mechanical actuators of control surfaces. The field of prognostics for electronics components is concerned with the prediction of remaining useful life (RUL) of components and systems. In particular, it focuses on condition-based health assessment by estimating the current state of health. Furthermore, it leverages the knowledge of the device physics and degradation physics to predict remaining useful life as a function of current state of health and anticipated operational and environmental conditions.

1.1 Motivation

The development of prognostics methodologies for the electronics field has become more important as more electrical systems are being used to replace traditional systems in several applications in fields like aeronautics, maritime, and automotive. The development of prognostics methods for electronics presents several challenges due to great variety of components used in a system, a continuous development of new electronics technologies, and a general lack of understanding of how electronics fail. Traditional reliability techniques in electronics tend to focus on understanding the time to failure for a batch of components of the same type. Just until recently, there has been a push to understand, in more depth, how a fault progresses as a function of usage, namely, loading and environmental conditions. Furthermore, just until recently, it was believed that there were no precursor of failure indications for electronics systems. That is now understood to be incorrect, since electronics systems, similar to mechanical systems, undergo a measurable wear process from which one

can derive features that can be used to provide early warnings to failure. These failures can be detected before they happen and one can potentially predict the remaining useful life as a function of future usage and environmental conditions.

Avionics systems in on-board autonomous aircraft perform critical functions greatly escalating the ramification of an in-flight malfunction (Bhatti & Ochieng, 2007; Kulkarni et al., 2009). These systems combine physical processes, computational hardware and software; and present unique challenges for fault diagnosis. A systematic analysis of these conditions is very important for analysis of aircraft safety and also to avoid catastrophic failures during flight.

Power supplies are critical components of modern avionics systems. Degradations and faults of the DC-DC converter unit propagate to the GPS (global positioning system) and navigation subsystems affecting the overall operation. Capacitors and MOSFETs (metal oxide field effect transistor) are the two major components, which cause degradations and failures in DC-DC converters (Kulkarni, Biswas, Bharadwaj, & Kim, 2010). Some of the more prevalent fault effects, such as a ripple voltage surge at the power supply output can cause glitches in the GPS position and velocity output, and this in turn, if not corrected can propagate and distort the navigation solution.

Capacitors are used as filtering elements on power electronics systems. Electrical power drivers for motors require capacitors to filter the rail voltage for the H-bridges that provide bidirectional current flow to the windings of electrical motors. These capacitors help to ensure that the heavy dynamic loads generated by the motors do not perturb the upstream power distribution system. Electrical motors are an essential element in electro-mechanical actuators systems that are being used to replace hydro-mechanical actuation in control surfaces of future generation aircrafts.

1.2 Methodology

The process followed in the proposed prognostics methodology is presented in the block-diagram in Figure 1. This prognostics methodology is based on results from an accelerated life test on real electrolytic capacitors. This test applies electrical overstress to commercial-off-the-shelf capacitors in order to observe and record the degradation process and identify performance conditions in the neighborhood of the failure criteria in a considerably reduced time frame.

Electro-impedance spectroscopy is used periodically during the test to characterize the frequency response of the capacitor. These measurements along a reduced order model based on passive electrical elements are used to identify the capacitance and parasitic resistance element.

We present here an empirical degradation model that is based on the observed degradation process during the accelerated life test. A model structure is suggested based on the observed degradation curves. Model parameters are estimated

using nonlinear least-squares regression. A Bayesian framework is employed to estimate (track) the state of health of the capacitor based on measurement updates of key capacitor parameters. The Kalman filter algorithm is used to track the state of health and the degradation model is used to make predictions of remaining useful life once no further measurements are available. A discussion and physical interpretation of the degradation model is presented. An analysis of the frequency response guides the selection of the precursor of failure variable used in the RUL prediction framework. A first order capacitance and equivalent series resistance (ESR) model is employed and the capacitance value is used in the development of the algorithm.

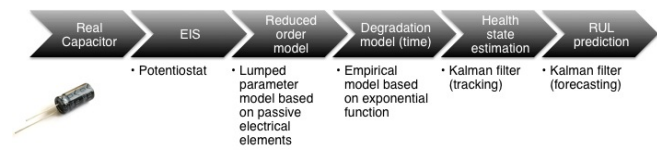


Figure 1. Model-based prognostics methodology for electrolytic capacitor.

1.3 Previous work

In earlier work (Kulkarni, Biswas, Koutsoukos, Goebel, & Celaya, 2010b), we studied the degradation of capacitors under nominal operation. There, work capacitors were used in a DC-DC converter and their degradation was monitored over an extended period of time. The capacitors were characterized every 100-120 hours of operation to capture degradation data for ESR and capacitance. The data collected over the period of about 4500 hours of operation were then mapped against an Arrhenius inspired ESR degradation model (Kulkarni, Biswas, Koutsoukos, Goebel, & Celaya, 2010a).

In following experimental work, we studied accelerated degradation in capacitors (Kulkarni, Biswas, Koutsoukos, Celaya, & Goebel, 2010). In that experiment the capacitors were subjected to high charging/discharging cycles at a constant frequency and their degradation progress was monitored. A preliminary approach to remaining useful life prediction of electrolytic capacitors was presented in (Celaya et al., 2011). This paper here builds upon the work presented in the preliminary remaining useful life prediction in (Celaya et al., 2011).

1.4 Other related work and current art in capacitor prognostics

The output filter capacitor has been identified as one of the elements of a switched mode power supply that fails more frequently and has a critical impact on performance (Goodman et al., 2007; Judkins et al., 2007; Orsagh et al., 2005). A prognostics and health management approach for power supplies of avionics systems is presented in (Orsagh et al., 2005). Re-

sults from accelerated aging of the complete supply were presented and discussed in terms of output capacitor and power MOSFET failures; but there is no modeling of the degradation process or RUL prediction for the power supply. Other approaches for prognostics for switched mode power supplies are presented in Goodman et al. (2007) and Judkins et al. (2007). The output ripple voltage and leakage current are presented as a function of time and degradation of the capacitor, but no details were presented regarding the modeling of the degradation process and there were no technical details on fault detection and RUL prediction algorithms.

A health management approach for multilayer ceramic capacitors is presented in Nie et al. (2007). This approach focuses on the temperature-humidity bias accelerated test to replicate failures. A method based on Mahalanobis distance is used to detect abnormalities in the test data; there is no prediction of RUL. A data driven prognostics algorithm for multilayer ceramic capacitors is presented in Gu et al. (2008). This method uses data from accelerated aging test to detect potential failures and to make an estimation of time of failure.

2. ACCELERATED AGING EXPERIMENTS

Accelerated life test methods are often used in prognostics research as a way to assess the effects of the degradation process through time. It also allows for the identification and study of different failure mechanisms and their relationships with different observable signals and parameters. In the following section we present the accelerated aging methodology and an analysis of the degradation pattern induced by the aging. The work presented here is based on an accelerated electrical overstress. In the following subsections, we first present a brief description of the aging setup followed by an analysis of the observed degradation. The precursor to failure is also identified along with the physical processes that contribute to the degradation.

2.1 Accelerated aging system description

Since the objective of this experiment is studying the effects of high voltage on degradation of the capacitors, the capacitors were subjected to high voltage stress through an external supply source using a specially developed hardware. The capacitors are not operated within DC-DC converters; only the capacitors were subjected to the stress.

The voltage overstress is applied to the capacitors as a square wave form in order to subject the capacitor to continuous charge and discharge cycles.

At the beginning of the accelerated aging, the capacitors charge and discharge simultaneously; as time progresses and the capacitors degrade, the charge and discharge times vary for each capacitor. Even though all the capacitors under test are subjected to similar operating conditions, their ESR and capacitance values change differently. We therefore monitor charging and discharging of each capacitor under test and

measure the input and output voltages of the capacitor. Figure 2 shows the block diagram for the electrical overstress experiment. Additional details on the accelerated aging system are presented in (Kulkarni, Biswas, Koutsoukos, Celaya, & Goebel, 2010).

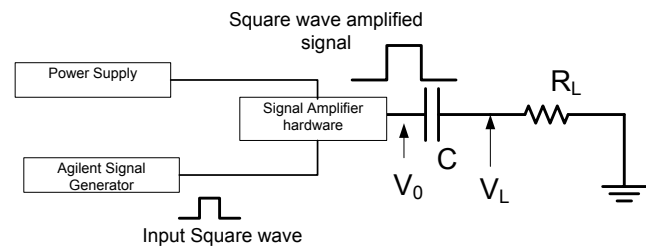


Figure 2. Block diagram of the experimental setup.

For this experiment six capacitors in a set were considered for the EOS experimental setup. Electrolytic capacitors of $2200\mu\text{F}$ capacitance, with a maximum rated voltage of 10V , maximum current rating of 1A and maximum operating temperature of 105°C was used for the study. These were the recommended capacitors by the manufacturer for DC-DC converters. The electrolytic capacitors under test were characterized in detail before the start of the experiment at room temperature.

The ESR and capacitance values were estimated from the capacitor impedance frequency response measured using an SP-150 Biologic SAS electro-impedance spectroscopy instrument. A lumped parameter model consisting of a capacitor with a resistor in series was assumed to estimate the ESR and capacitance. The average pristine condition ESR value was measured to be $0.056\text{ m}\Omega$ and average capacitance of $2123\mu\text{F}$ individually for the set of capacitors under test.

The measurements were recorded every 8-10 hours of the total 180 plus hours of accelerated aging time to capture the rapid degradation phenomenon in the ESR and capacitance values. The ambient temperature for the experiment was controlled and kept at 25°C . During each measurement the voltage source was shut down, capacitors were discharged completely and then the characterization procedure was carried out. This was done for all the six capacitors under test. For further details regarding the aging experiment results and analysis of the measured data refer to (Kulkarni, Biswas, Koutsoukos, Celaya, & Goebel, 2010; Celaya et al., 2011).

2.2 Physical interpretation of the degradation process

There are several factors that cause electrolytic capacitors to fail. Continued degradation, i.e., gradual loss of functionality over a period of time results in the failure of the component. Complete loss of function is termed a *catastrophic* failure. Typically, this results in a short or open circuit in the capacitor. For capacitors, degradation results in a gradual increase

in the equivalent series resistance (ESR) and decrease in capacitance over time.

In this work, we study the degradation of electrolytic capacitors operating under high electrical stress, i.e., $V_{applied} \geq V_{rated}$. During the charging/discharging process the capacitors degrade over the period of time. A study of the literature indicated that the degradation could be primarily attributed to three phenomena (IEC, 2007-03; MIL-C-62F, 2008):

1. Electrolyte evaporation,
2. Leakage current, and
3. Increase in internal pressure

An ideal capacitor would offer no resistance to the flow of current at its leads. However, the electrolyte (aluminum oxide) that fills the space between the plates and the electrodes produces a small equivalent internal series resistance (ESR). The ESR dissipates some of the stored energy in the capacitor. In spite of the dielectric insulation layer between a capacitor's plates, a small amount of 'leakage' current flows between the plates. For a good capacitor operating nominally this current is not significant, but it becomes larger as the oxide layer degrades during operation. High electrical stress is known to accentuate the degradation of the oxide layer due to localized dielectric breakdowns on the oxide layer (Ikonopisov, 1977; Wit & Crevecoeur, 1974).

The literature on capacitor degradation shows a direct relationship between electrolyte decrease and increase in the ESR of the capacitor (Kulkarni, Biswas, Koutsoukos, Goebel, & Celaya, 2010b). ESR increase implies greater dissipation, and, therefore, a slow decrease in the average output voltage at the capacitor leads. Another mechanism occurring simultaneously is the increase in internal pressure due to an increased rate of chemical reactions, which are attributed to the internal temperature increase in the capacitor.

During the experiments, as discussed earlier, the capacitors were characterized at regular intervals. ESR and capacitance are the two main failure precursors that tipify the current health state of the device. ESR and capacitance values were calculated after characterizing the capacitors. As the devices degrade due to different failure mechanisms we can observe a decrease in the capacitance and an increase in the ESR.

ESR and capacitance values are estimated by using a system identification using a lump parameter model consistent of the capacitance and the ESR in series as shown in Figure 3. The frequency response of the capacitor impedance (measured with electro-impedance spectroscopy) is used for the parameter estimation. It should be noted that the lumped-parameter model used to estimate ESR and capacitance, is not the model to be used in the prognostics algorithm; it only allows us to estimate parameters which provide indications of the degradation process through time. Parameters such as ESR and capacitance are challenging to estimate from the *in-*

situ measurements of voltage and current through the accelerated aging test.

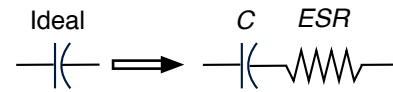


Figure 3. Lumped parameter model for a real capacitor.

Figure 4 shows percentage increase in the ESR value for all the six capacitors under test over the period of time. This value of ESR is calculated from the impedance measurements after characterizing the capacitors. Similarly, figure 5 shows the percentage decrease in the value of the capacitance as the capacitor degrades over the period under EOS test condition discussed. As per standards MIL-C-62F (2008), a capacitor is considered unhealthy if under electrical operation its ESR increases by 280 – 300% of its initial value or the capacitance decreases by 20% below its pristine condition value. From the plots in Figure 4 we observe that for the time for which the experiments were conducted the average ESR value increased by 54% – 55% while over the same period of time, the average capacitance decreased by more than 20% (the threshold mark for a healthy capacitor) (see Figure 5). As a result, the percentage capacitance loss is selected as a precursor of failure variable to be used in the degradation model development presented next.

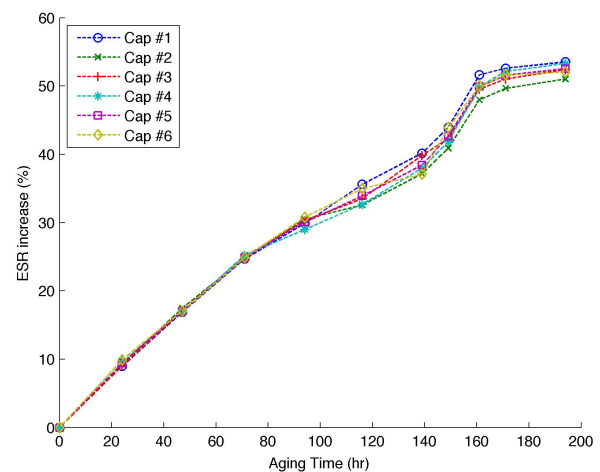


Figure 4. Degradation of capacitor performance, percentage ESR increase as a function of aging time.

3. PREDICTION OF REMAINING USEFUL LIFE

A model-based prognostics algorithm based on Kalman filter and a physics inspired empirical degradation model is presented. This algorithm is able to predict remaining useful life of the capacitor based on the accelerated degradation data

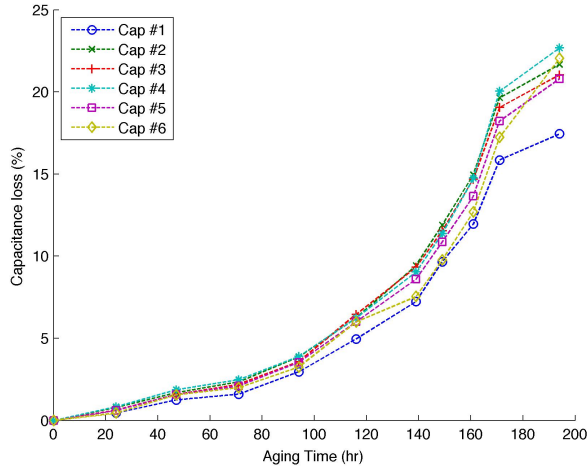


Figure 5. Degradation of capacitor performance, percentage capacitance loss as a function of aging time.

from the experiments described in previous sections. The percentage loss in capacitance is used as a precursor of failure variable and it is used to build a model of the degradation process. This model relates aging time to the percentage loss in capacitance and has the following form,

$$C_k = e^{\alpha t_k} + \beta, \quad (1)$$

where α and β are model constants that will be estimated from the experimental data of accelerated aging experiments. In order to estimate the model parameters, five capacitors are used for estimation (labeled capacitors #1 through #5), and the remaining capacitor (#6) is used to test the prognostics algorithm. A nonlinear least-squares regression algorithm is used to estimate the model parameters. Figure 6 shows the estimation results. The experimental data are presented together with results from the exponential fit function. It can be observed from the residuals that the estimation error increases with time. This is to be expected since the last data point measured for all the capacitors fall slightly off the concave exponential model. The estimated parameters are $\alpha = 0.0163$ and $\beta = -0.5653$.

The estimated degradation model is used as part of a Bayesian tracking framework to be implemented using the Kalman filter technique. This method requires a state-space dynamic model relating the degradation level at time t_k to the degradation level at time t_{k-1} . The formulation of the state model is described below.

$$\begin{aligned} \frac{dC}{dt} &= \alpha C - \alpha\beta \\ \frac{C_t - C_{t-\Delta t}}{\Delta t} &= \alpha C_{t-\Delta t} - \alpha\beta \\ C_t &= (1 + \alpha\Delta t)C_{t-\Delta t} - \alpha\beta\Delta t, \end{aligned} \quad (2)$$

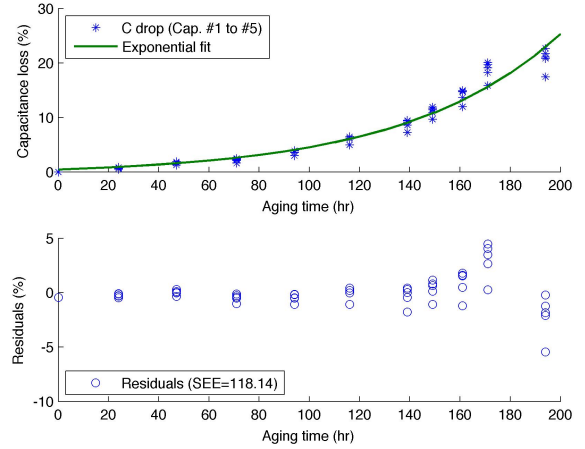


Figure 6. Estimation results for the empirical degradation model.

$$C_k = (1 + \alpha\Delta_k)C_{k-1} - \alpha\beta\Delta_k. \quad (3)$$

In this model C_k is the state variable and it represents the percentage loss in capacitance. Since the system measurements are percentage loss in capacitance as well, the output equation is given by $y_k = hC_k$, where the value of h is equal to one. The following system structure is used in the implementation of the filtering and the prediction using the Kalman filter.

$$C_k = A_k C_{k-1} + B_k u + v, \quad (4)$$

$$y_k = h C_{k-1} + w, \quad (5)$$

where,

$$\begin{aligned} A_k &= (1 + \Delta_k), \\ B_k &= -\alpha\beta\Delta_k, \\ h &= 1, \\ u &= 1. \end{aligned} \quad (6)$$

The time increment between measurements Δ_k is not constant since measurements were taken at non-uniform sampling rate. This implies that some of the parameters of the model in equations (4)-(6) will change through time. Furthermore, v and w are normal random variables with zero mean and Q and R variance respectively. The description of the Kalman filtering algorithm is omitted from this article. A thorough description of the algorithm can be found in Stengel (1994), a description of how the algorithm is used for forecasting can be found in Chatfield (2003) and an example of its usage for prognostics can be found in (Saha et al., 2009). Figure 7 shows the results of the application of the Kalman filter to the test case (Cap. #6). The model noise variance Q was estimated from the model regression residuals. The residuals have a mean very close to zero and a variance of 2.1829. This

variance was used for the model noise in the Kalman filter implementation. The measurement noise variance R is also required in the filter implementation. This variance was computed from the direct measurements of the capacitance with the electro-impedance spectroscopy equipment, the observed variance is $4.99E^{-7}$. Figure 7 shows the result of the filter tracking the complete degradation signal. The residuals show an increased error with aging time. This is to be expected given the results observed from the model estimation process.

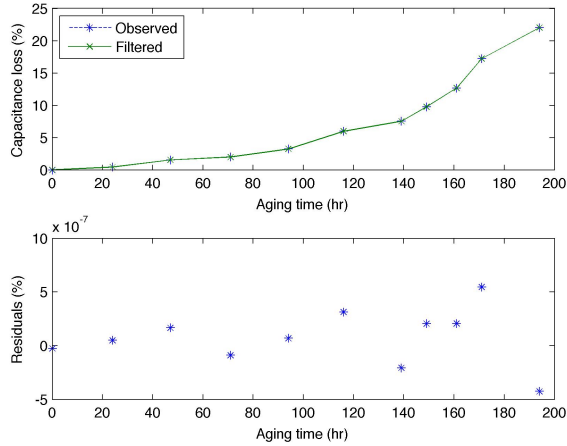


Figure 7. Tracking results for the Kalman filter implementation applied to test capacitor (capacitor #6).

The use of the Kalman filter as a RUL forecasting algorithm requires the evolution of the state without updating the error covariance matrix and the posterior of the state vector. The n step ahead forecasting equation for the Kalman filter is given below. The last update is done at the time of the last measurement t_l .

$$\hat{C}_{t+n} = A^n C_l + \sum_{i=0}^{n-1} A^i B \quad (7)$$

The subscripts from parameters A and B are omitted since a constant Δ_t is used in the forecasting mode (one prediction every hour). Figure 8 presents results from the remaining useful life prediction algorithm at time 149 (hr), which is the time at which an ESR and C measurements are taken. The failure threshold is considered to be a crisp value of 20% decrease in capacitance. End of life (EOL) is defined as the time at which the forecasted percentage capacity loss trajectory crosses the EOL threshold. Therefore, RUL is EOL minus 149 hours.

Figure 9 presents the capacitance loss estimation and EOL prediction at different points during the aging time. Predictions are made after each point in which measurements are available. It can be observed that the predictions become better as the prediction is made closer to the actual EOL. This is possible because the estimation process has more information to update the estimates as it nears EOL. Figure 10 presents a

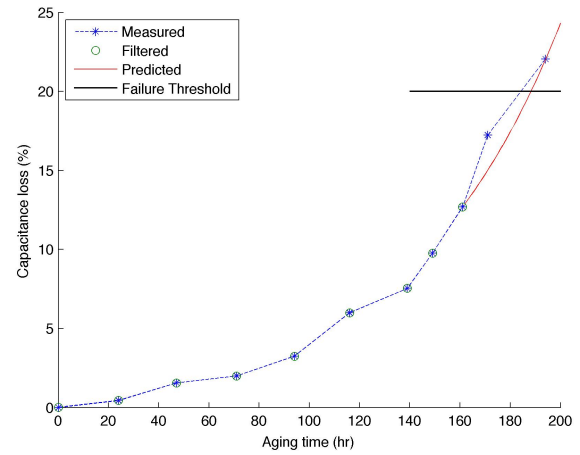


Figure 8. Remaining useful life prediction at time 149 (hr).

zoomed-in version of figure 9 focusing in the area close to the failure threshold.

Table 1 summarizes results for the remaining life prediction at all points in time where measurements are available. The last column indicates the RUL prediction error. The magnitude of the error decreases as the prediction time gets closer to EOL. The decrease is not monotonic which is to be expected when using a the tracking framework to estimate health state because the last point of the estimation is used to start the forecasting process. An α - λ prognostics performance metric is presented in Figure 11. The blue line represents ground truth and the shaded region is corresponding to a 30% ($\alpha = 0.3$) error bound in the RUL prediction. This metric specifies that the prediction is within the error bound halfway between first prediction and EOL ($\lambda = 0.5$). In addition, this metric allows us to visualize how the RUL prediction performance changes as data closer to EOL becomes available.

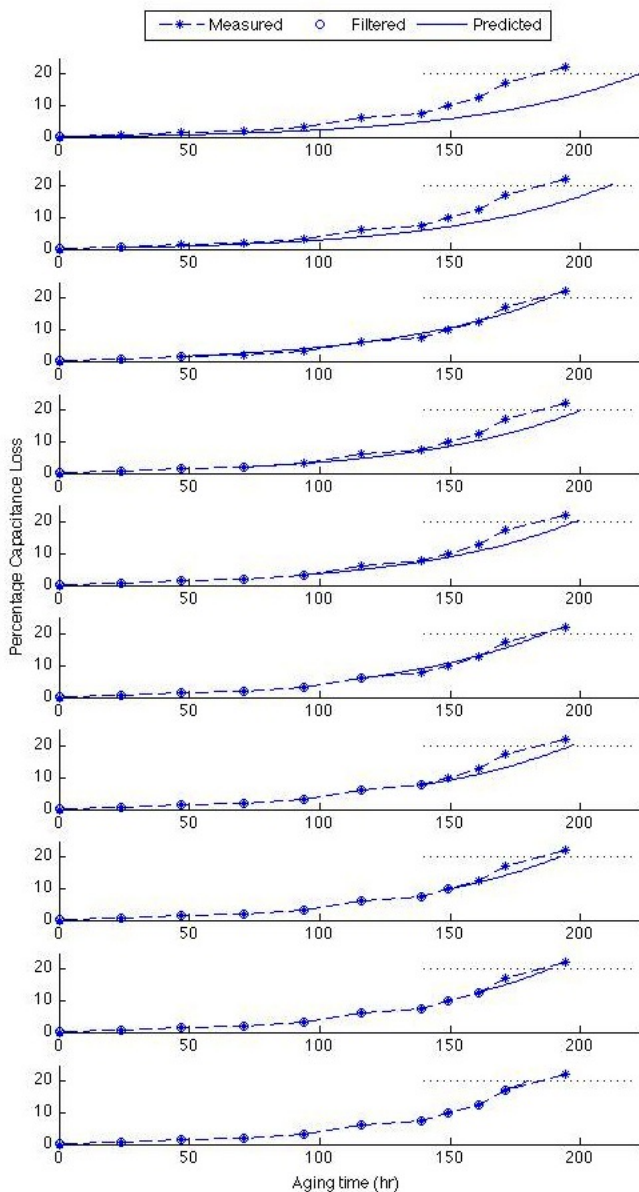


Figure 9. Health state estimation and forecasting of capacitance loss (%) at different times t_p during the aging time; $t_p = [0, 24, 47, 71, 94, 116, 139, 149, 161, 171]$.

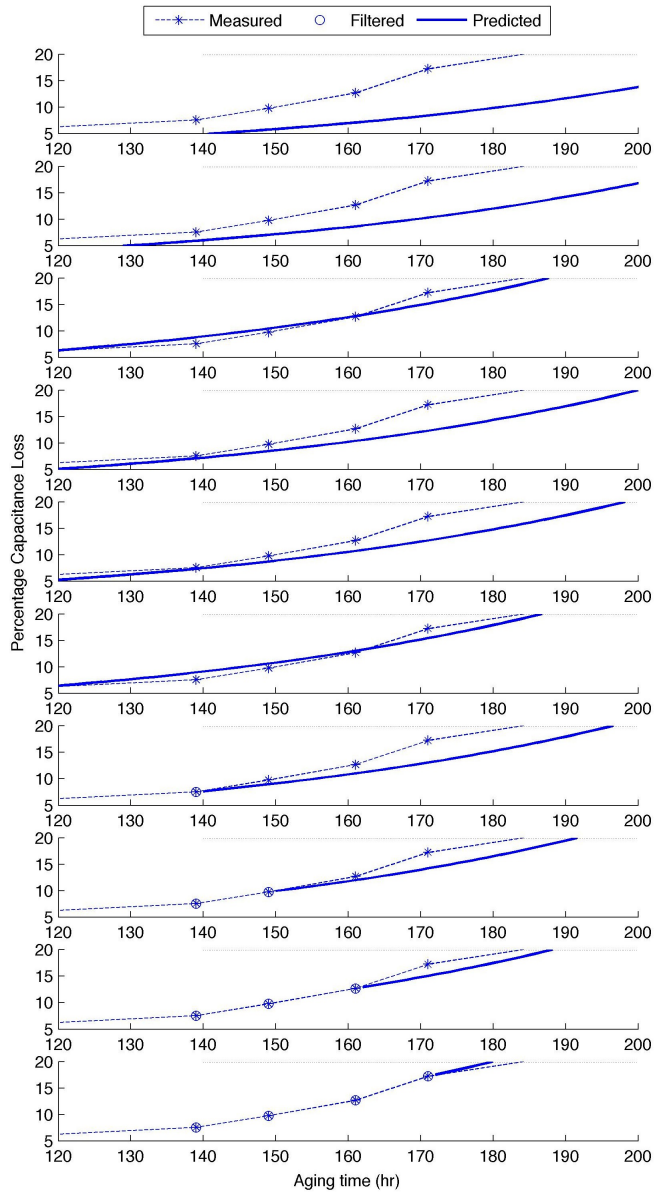


Figure 10. Detail of the health state estimation and forecasting of capacitance loss (%) at different times t_p during the aging time; $t_p = [0, 24, 47, 71, 94, 116, 139, 149, 161, 171]$.

4. CONCLUSION

This paper presents a RUL prediction algorithm based on accelerated life test data and an empirical degradation model. The main contributions of this work are: a) the identification of the lumped-parameter model (Figure 3) for a real capacitor as a viable reduced-order model for prognostics-algorithm development; b) the identification of the ESR and C model parameters as precursor of failure features; c) the development of an empirical degradation model based on accelerated life test data which accounts for shifts in capacitance as a func-

RUL forecasting time (hr)	RUL estimate (hr)	Ground truth (hr)	Error (hr)
0	222.2	184.24	37.96
24	186.55	160.24	26.31
47	140.66	137.24	3.42
71	128.98	113.24	15.74
94	104.18	90.24	13.94
116	70.71	68.24	2.47
139	57.58	45.24	12.34
149	42.61	35.24	11.37
161	27.20	23.24	3.96
171	8.94	13.24	-4.3

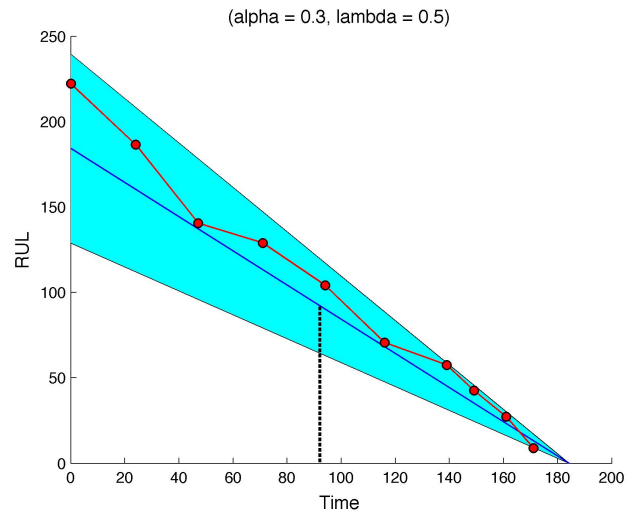
Table 1. Summary of RUL forecasting results.

tion of time; d) the implementation of a Bayesian based health state tracking and remaining useful life prediction algorithm based on the Kalman filtering framework. One major contribution of this work is the prediction of remaining useful life for capacitors as new measurements become available.

This capability increases the technology readiness level of prognostics applied to electrolytic capacitors. The results presented here are based on accelerated life test data and on the accelerated life timescale. Further research will focus on development of functional mappings that will translate the accelerated life timescale into real usage conditions time-scale, where the degradation process dynamics will be slower, and subject to several types of stresses. The performance of the proposed exponential-based degradation model is satisfactory for this study based on the quality of the model fit to the experimental data and the RUL prediction performance as compared to ground truth. As part of future work we will also focus on the exploration of additional models based on the physics of the degradation process and larger sample size for aged devices. Additional experiments are currently underway to increase the number of test samples. This will greatly enhance the quality of the model, and guide the exploration of additional degradation-models, where the loading conditions and the environmental conditions are also accounted for towards degradation dynamics.

ACKNOWLEDGMENT

This work was funded by the NASA Aviation Safety Program, SSAT project, IVHM project and NRA grant #NNX07ADIZA.

Figure 11. Performance based on α - λ performance metric.

NOMENCLATURE

C_p	Pristine state measured capacitance
ESR	Equivalent series resistance of the electrolytic capacitor
ESR_p	Pristine state measured equivalent series resistance
C_k	Measured capacitance at time t_k
RUL	Remaining useful life

REFERENCES

- Bhatti, U., & Ochieng, W. (2007). Failure modes and models for integrated gps/ins systems. *The Journal of Navigation*, 60, 327.
- Celaya, J., Kulkarni, C., Biswas, G., & Goebel, K. (2011, March). Towards prognostics of electrolytic capacitors. In *AIAA 2011 Infotech@Aerospace Conference*. St. Louis, MO.
- Chatfield, C. (2003). *The analysis of time series: An introduction* (6th ed.). Chapman and Hall/CRC.
- Goodman, D., Hofmeister, J., & Judkins, J. (2007). Electronic prognostics for switched mode power supplies. *Microelectronics Reliability*, 47(12), 1902-1906. (doi: DOI: 10.1016/j.microrel.2007.02.021)
- Gu, J., Azarian, M. H., & Pecht, M. G. (2008). Failure prognostics of multilayer ceramic capacitors in temperature-humidity-bias conditions. In *Prognostics and health management, 2008. phm 2008. international conference on* (p. 1-7).
- IEC. (2007-03). *60384-4-1 fixed capacitors for use in electronic equipment* (Tech. Rep.).
- Ikonopisov, S. (1977). Theory of electrical breakdown during formation of barrier anodic films. *Electrochimica Acta*, 22, 1077-1082.
- Judkins, J. B., Hofmeister, J., & Vohnout, S. (2007). A prognostic sensor for voltage regulated switch-mode power supplies. In *Aerospace conference, 2007 IEEE* (p. 1-8).

- Kulkarni, C., Biswas, G., Bharadwaj, R., & Kim, K. (2010). Effects of degradation in dc-dc converters on avionics systems: A model based approach. *Machinery Failure Prevention Technology Conference, MFPT 2010*.
- Kulkarni, C., Biswas, G., & Koutsoukos, X. (2009). A prognosis case study for electrolytic capacitor degradation in dc-dc converters. *Annual Conference of the Prognostics and Health Management Society, PHM 2009*.
- Kulkarni, C., Biswas, G., Koutsoukos, X., Celaya, J., & Goebel, K. (2010). Integrated diagnostic/prognostic experimental setup for capacitor degradation and health monitoring. In *2010 IEEE AUTOTESTCON* (p. 1-7).
- Kulkarni, C., Biswas, G., Koutsoukos, X., Goebel, K., & Celaya, J. (2010a). Experimental studies of ageing in electrolytic capacitors. *Annual Conference of the Prognostics and Health Management Society*.
- Kulkarni, C., Biswas, G., Koutsoukos, X., Goebel, K., & Celaya, J. (2010b). Physics of Failure Models for Capacitor Degradation in DC-DC Converters. *The Maintenance and Reliability Conference, MARCON 2010*.
- MIL-C-62F. (2008). *General specification for capacitors, fixed, electrolytic*.
- Nie, L., Azarian, M. H., Keimasi, M., & Pecht, M. (2007). Prognostics of ceramic capacitor temperature-humidity-bias reliability using mahalanobis distance analysis. *Circuit World*, 33(3), 21 - 28.
- Orsagh, R., Brown, D., Roemer, M., Dabnev, T., & Hess, A. (2005). Prognostic health management for avionics system power supplies. In *Aerospace conference, 2005 IEEE* (p. 3585-3591).
- Saha, B., Goebel, K., & Christophersen, J. (2009). Comparison of prognostic algorithms for estimating remaining useful life of batteries. *Transactions of the Institute of Measurement and Control*, 31(3-4), 293-308.
- Stengel, R. F. (1994). *Optimal control and estimation*. Dover Books on Advanced Mathematics.
- Wit, H. D., & Crevecoeur, C. (1974). The dielectric breakdown of anodic aluminum oxide. *Physics Letters A, December*, 365-366.

José R. Celaya is a research scientist with SGT Inc. at the Prognostics Center of Excellence, NASA Ames Research Center. He received a Ph.D. degree in Decision Sciences and Engineering Systems in 2008, a M. E. degree in Operations Research and Statistics in 2008, a M. S. degree in Electrical Engineering in 2003, all from Rensselaer Polytechnic Institute, Troy New York; and a B. S. in Cybernetics Engineering in 2001 from CETYS University, México.

Chetan S Kulkarni is a Research Assistant at ISIS, Vanderbilt University. He received the M.S. degree in EECS from Vanderbilt University, Nashville, TN, in 2009, where he is currently a Ph.D student.

Sankalita Saha is a research scientist with Mission Critical Technologies at the Prognostics Center of Excellence, NASA

Ames Research Center. She received the M.S. and PhD. degrees in Electrical Engineering from University of Maryland, College Park in 2007. Prior to that she obtained her B.Tech (Bachelor of Technology) degree in Electronics and Electrical Communications Engineering from the Indian Institute of Technology, Kharagpur in 2002.

Kai Goebel received the degree of Diplom-Ingenieur from the Technische Universitt Mnchen, Germany in 1990. He received the M.S. and Ph.D. from the University of California at Berkeley in 1993 and 1996, respectively. Dr. Goebel is a senior scientist at NASA Ames Research Center where he leads the Diagnostics and Prognostics groups in the Intelligent Systems division. In addition, he directs the Prognostics Center of Excellence and he is the technical lead for Prognostics and Decision Making of NASAs System-wide Safety and Assurance Technologies Program. He worked at General Electric Corporate Research Center in Niskayuna, NY from 1997 to 2006 as a senior research scientist. He has carried out applied research in the areas of artificial intelligence, soft computing, and information fusion. His research interest lies in advancing these techniques for real time monitoring, diagnostics, and prognostics. He holds 15 patents and has published more than 200 papers in the area of systems health management.

Gautam Biswas received the Ph.D. degree in computer science from Michigan State University, East Lansing. He is a Professor of Computer Science and Computer Engineering in the Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN.

A Structural Health Monitoring Software Tool for Optimization, Diagnostics and Prognostics

Seth S. Kessler¹, Eric B. Flynn², Christopher T. Dunn³ and Michael D. Todd⁴

^{1,2,3}*Metis Design Corporation, Cambridge, MA, 02141, USA*

skessler@metisdesign.com

eflynn@metisdesign.com

cdunn@metisdesign.com

⁴*University of California San Diego*

mtodd@ucsd.edu

ABSTRACT

Development of robust structural health monitoring (SHM) sensors and hardware alone is not sufficient to achieve desired benefits such as improved asset availability and reduced sustainment costs. For SHM systems to be practically deployed as part of an integrated system health management (ISHM), tools must be created for SHM life-cycle management (LCM). To that end, SHM-LCM software has been developed to expedite the adoption of SHM into ISHM. The SHM-LCM software is a flexible application intended to manage the cradle-to-grave life-cycle of an SHM system for generic applications. There are 4 core modules to facilitate critical roles: Optimization, Calibration, Visualization, and Action. The Optimization module seeks to devise optimal sensor placement and excitation parameters in order to achieve probability of detection (POD) coverage requirements. The Calibration module is designed to guide a user through a series of material level tests in order to customize algorithm variables to the system being designed. The Visualization module is dedicated to generating a diagnostic composite picture based on data downloaded from the diagnostic server, which is "stitched" to the original 3D mesh, providing users with a manipulatable GUI to toggle between probability of damage distributions for various calibrated damage modes. Finally, The Action module generates residual performance plots (ultimate load or deflection for example) as a function of probability of damage, so detection confidence can be weighed against impact to the vehicle's capabilities. SHM-LCM software will enable SHM systems to be incorporated into ISHM by engineers rather than experts, making the technology more accessible, and commercially practical.

Kessler et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

1. INTRODUCTION

Currently successful laboratory non-destructive testing and monitoring methods are impractical for service inspection of large-area structures due to the size and complexity of the support equipment required, as well as the time and cost associated with component tear-down. It is clear that new approaches for inspection are necessary. Structural Health Monitoring (SHM) denotes the ability to detect and interpret adverse "changes" in a structure to direct actions that reduce life-cycle costs and improve reliability. Essentially, minimally-invasive detection sensors are integrated into a structure to continuously collect data that are mined for information relating to damage such as cracks or corrosion. SHM is receiving increasing attention, particularly from the DoD community, to eliminate scheduled and/or manual inspections in lieu of condition-based maintenance for more efficient design practices and more accurate repair and replacement decisions. This methodology shift will result in significant savings in overall cost of ownership of a vehicle, as well as significant gains in operational safety.

For SHM to be successfully implemented, accurate diagnostic and prognostic models are essential. Not only do sensors need to be properly integrated to collect data, but diagnostic characterization of the health of the structure needs to be extracted and presented to the operator and/or maintainer in a timely and meaningful manner. Furthermore, the diagnostic information should be converted to prognostic predictions so that proper action regarding remaining useful life or necessary repair can be taken. There are presently limited methods for visualizing diagnostic data, mainly 2-D representations, and no proven software to explicitly link diagnostic and prognostic information. Some methods have been demonstrated for health & usage monitoring system (HUMS); however, these systems provide far less detailed information compared to what is expected from an SHM system.

The overall approach taken by the current investigators was a system optimization problem; attempting to maximize detection capabilities with minimal impact to the test structure and at minimal cost, both capitalized and risk-generated. Hundreds of sensors densely spaced over a test structure would certainly have the best opportunity to precisely resolve damage locations, but this would obviously be impractical for real-life applications due to the quantity of instrumentation required (cables, data acquisition hardware, etc) or other incurred penalties (e.g. weight on an aircraft). Therefore the chosen approach was to use Bayesian risk function to assign costs to missed-damage, false-positives, and localization error as well as associating a cost with each sensor (where cost here is not monetary necessarily, but a relative metric for comparing the value of each parameter).

The algorithms used were a hybrid collection of functions making use of both coherent and incoherent information in the data. Data for each sensor is processed separately, then ultimately summed in a weighted fashion across the test structure. Further logic is also deployed to eliminate anomalies and invalid features. Generally, this process is analogous to active sonar. Damage ("targets") are detected and/or localized by generating ultrasonic elastic waves and watching how they bounce off of potential targets. Because a test structure is arguably far more complex than the open ocean, producing potentially far more "false targets" (such as boundaries, stiffeners, rivets, size changes, material interfaces, etc.), this approach takes advantage of embedding probabilistic models into the wave propagation/scattering process so that likelihood-based judgments can be made about the damage targets. These judgments may be understood in appropriate performance terms—probability of detection, probability of localization, etc.—which directly supports the uncertainty quantification needed for decision-making.

Finally, the decomposed data must be displayed in a meaningful matter. Work was done to deploy a graphical-user-interface (GUI) that would allow 3D structures to be represented with damage predictions stitched-in. Controls are deliberately included to allow knowledgeable users to deviate from default algorithm and display parameter values to refine the image or search for smaller damage that is obfuscated by severe damage locations. The software is also built in such a way so that diagnostic results can be exported to commercial finite element tools to provide prognostic information such as residual strength or stiffness.

A major advantage of this overall approach is its power to serve also as a design tool. Through the overarching probabilistic framework, if a client-defined objective is established for a given application (e.g., "must detect fatigue cracks < 1 mm oriented at any random angle with a probability of 95% and use no more than 1 sensor per square meter"), this approach allows for an a priori optimization of

the sensor architecture before in-situ deployment to meet those objective(s). This provides tremendous potential cost savings, eliminating the "black box" and "trial and error" approaches to doing SHM system design.

2. SHM SYSTEM SENSORS AND HARDWARE

To achieve the overall goals of efficient damage detection, This research leverages hardware previously developed by the investigators, including distributed digitization hardware, piezoelectric-based damage and localization sensors. A patented method is used to determine relative phase information for the sensor responses, by surrounding a central actuating disk with multiple sensing disks, known as vector-based localization. The actuating and sensing component consists of seven piezoelectric wafers that are integrated into a custom flex-circuit assembly that connects to the digitization hardware. These elements are permanently mounted on the structure being monitored. The closely spaced set of piezoelectric elements in each node form a phased array, which enables the identification of both range and bearing to multiple damage sites using a single node. This is in contrast to isolated piezoelectric elements which can only identify range, necessitating the use of multiple spatially separated elements to localize damage sites through a triangulation process that has been shown to be susceptible to corruption by multiple damage sites. Also, if relative time of arrival at the sensor elements is used, a ray indicating angle to damage can be generated without any wavespeed information. Thus damage can be localized by simply finding the ray intersection of 2 of these vector-locator nodes. This method can be deployed actively using GW to determine the location of damage as described here, or passively in acoustic emission mode the same equations can be used to describe the position of an impact.

3. SENSOR PLACEMENT OPTIMIZATION

SHM systems are decision makers. At any given time, or according to any given measurement, the SHM system needs to be designed to let the operator know whether or not a potential problem in the structure requires action. As such, an SHM system will likely have to make hundreds or thousands of decisions while the structure is undamaged before a defect actually develops. During this time, it is important that the SHM system correctly decides that the structure is healthy as frequently as possible. If the SHM system constantly demands costly, unnecessary manual inspections then it provides no benefit to the monitored structure and its operation. It is important, then, that the design of SHM systems and the evaluation of their performance consider the total risk posed by all forms of decision errors.

3.1 Theory

The presented approach to SHM is the minimization of the expect cost, or Bayes Risk, associated with an SHM system through the optimal design of detection algorithms and hardware (i.e., sensor placement). Put simply, the Bayes Risk is sum of the costs of all possible detection outcomes (detection, missed detection, false alarm, etc) weighted by their probability of occurring. This can be represented as

$$R = \sum_{d, \theta} C(d, \theta) P(d | \theta, e) P(\theta) + C(e) \quad (1)$$

where d is the set of possible decisions the SHM system makes, θ is the damage state of the structure, and e is the design of the SHM algorithms and hardware. The first probability term describes the statistical performance of the SHM detection system and the second probability term reflects the prior probabilities of damage, if known. The optimal design is then defined as:

$$e^* = \arg \max_e R(e) \quad (2)$$

A key component of the approach to structural health monitoring is the optimization of the placement of sensor nodes according to the minimization of the expect cost, or Bayes Risk, associated with the decisions made by the SHM system. The calculation of the Bayes Risk for an arbitrary set of node placements then requires accurate models of the wave propagation process and detector statistics parameterized by the node coordinates. To simplify the modeling, the structure is divided into discrete regions. Then to determine the total Bayes Risk of the structure, the localized Bayes Risk is calculated for each region and sum. The statistical performance of detectors for each region is evaluated with any given set of node placements using an analytical model of the wave propagation and scattering process. This model includes beam spread, line of site, directional scattering, and transmission across the doublers. According to this stochastic model, the detector described above, and an optimal set of detector thresholds, maps can be constructed of the expected localized detection and false alarm rates. Examples of these maps for a two-node arrangement are shown in Figure 1. Note the effect of line of site and the doublers in the two maps. Nodes were optimally placed in a greedy algorithm fashion. Starting with one node, one at a time, each node is added so that it optimally compliments the existing fixed arrangement. As such, there always exists a subset of n nodes from the total N nodes that is near-optimal. Near-optimal in this case means a guaranteed performance of at least:

$$U_{Greedy}[n] \geq \left(1 - \left(\frac{n-1}{n}\right)^n\right) U^*[n] > 0.63 U^*[n] \quad (3)$$

where $U^*[n]$ is the performance (or utility) of the optimal arrangement of n nodes.

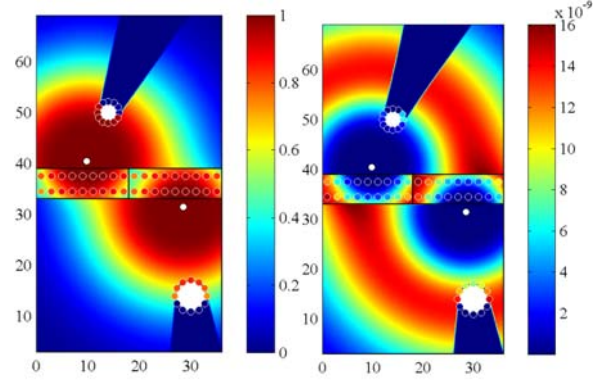


Figure 1: Local detection rates (left) and, false alarm rates (right) for a two-node arrangement. Nodes are indicated with small white-filled circles.

3.2 Optimization Example

A structure was divided into two sets of discrete regions. The first set forms a uniformly spaced grid covering the structure. The second set is assembled from the identified hot spots on the structure consisting of the localized area around each of the bolt holes. Each region from the uniform and hot-spot sets is then assigned a probability of damage. Wave scattering was modeled according to a 5 mm crack with uniform random orientation. Imaging noise was modeled as Raleigh distributed. Noise parameters were fitted from data acquired over two days from three-foot-square plate instrumented with identical nodes. The probabilities and error penalties were assigned as follows:

- Probability of damage being introduced: 80%
- Conditional probability of damage being introduced at hot spots: 60%
- Conditional probability of damage being introduced away from hot spots: 40%
- Penalty of missed detection / penalty of false alarm = 2/1

Figure 2 shows the optimized arrangement of six nodes on a map of the resulting normalized risk. The nodes are numbered in order of their placement by the greedy algorithm. The normalized risk for the greedy-chosen arrangement fell within 5% of the true optimal 6-node arrangement as found by an exhaustive genetic algorithm search. Figure 3 provides a graph of the normalized risk versus node count. When the cost of each additional node is added to the risk calculation, the risk versus node count will have a minimum that indicates the optimal number of sensors to use. Figure 3 demonstrates that adding additional nodes has diminishing returns when accounting for per-node costs.

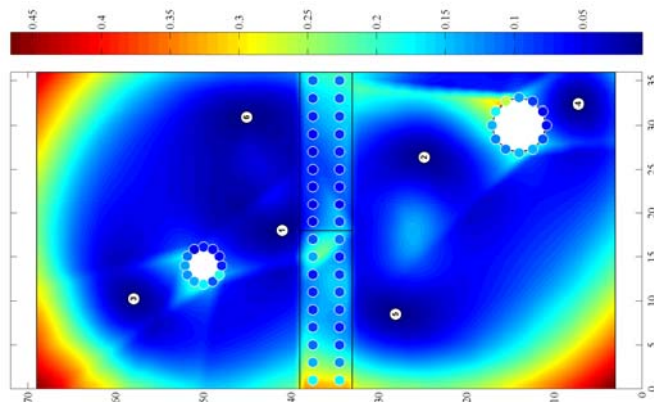


Figure 2: Optimal 6 node arrangement with map of normalized local risk

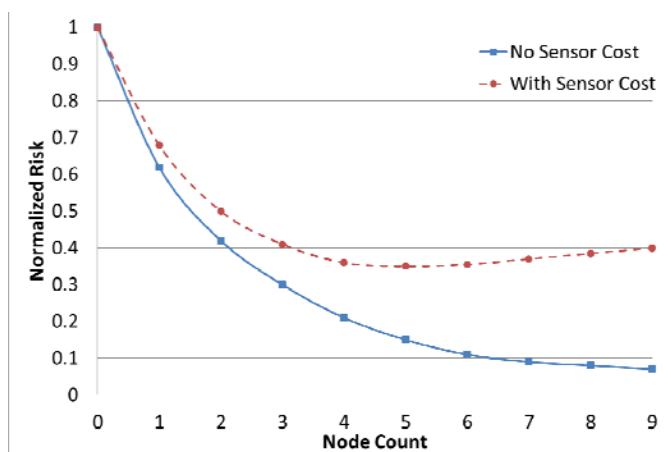


Figure 3: Normalized risk versus node count

4. DIAGNOSTIC ALGORITHMS

Active sensing involves mechanically exciting a structure and in turn measuring the response in order to gain information regarding the potential presence of damage. When one dimension of the structure being excited is relatively small compared to the other two, such as in a plate-like structure, and the wavelength(s) of excitation are of the same order as this dimension, the process is referred to as guided-wave active sensing.

The MD7 system works in an analogous fashion to active sonar. One at a time, each MD7 node actuates a series of narrow-band, ultrasonic mechanical pulses, or “pings”, using its central actuation transducer. These pulses propagate through the structure, reflect and scatter at geometric features, such as plate boundaries, as well as at potential damage, and are then sensed by the six sensing transducers on the node. The node digitizes the sensed responses and sends the data to the accumulation hub where it is stored for later retrieval and processing

The recorded responses are used to determine the range(s), bearing(s), and size(s) of potential damage in the structure relative to each node. In traditional active sonar applications, bearing is often determined in one of two ways. The first is to physically arrange the sonar array to maximize its sensitivity in one direction, and then mechanically orientate, or steer, the array to scan multiple directions. The second approach is to artificially introduce delays in the acquired, digitized responses in order to electronically steer the array through a processes known as beam forming. For the current application, the latter approach has two distinct advantages. First, the position of the array elements (i.e. sensing transducers) can be fixed so there are no moving parts. Second, a single actuated pulse and sensed response can be used to simultaneously scan for damage in every direction. This directional scanning through electronic steering forms the basis of the present approach to ultrasonic guided wave imaging.

4.1 Beamforming

Optimal detectors can be derived according to statistical likelihood tests on the measured responses for the presence and location of damage. Depending upon the specific objective(s), such detectors provide a means of combining measurement data to build a set of test statistics $T(x)$ (sometimes referred to as “damage features”) that can be compared to a threshold (determined by a risk analysis) in order to make decisions regarding the existence and/or location of damage on the structure. In most cases, where localization is of prime importance, the time of flight from the actuator to the potentially damaged region to the sensor for a given wave number can be reasonably estimated based on an average group velocity computed from the (likely heterogeneous) material and geometric properties along the propagation path. With this in mind, a common localization detection approach for each region in a structure is one that delays and sums the measurements from the different transducer pairs so that they will additively combine at the true location of damage, resulting in an “image” of highly constructive scatter relative to the background noise. However, the relative average phase velocities from each transducer pair to each region of the structure can be more difficult to predict. This leads to two basic forms of detectors based on the statistical model of the measurements: coherent and incoherent beam forming.

In the case where the relative phase velocity is different and unknown between transducer pairs, the envelopes of the waveforms must be summed together in order to eliminate the dependence on phase. Otherwise, the delayed and summed waveforms run the risk of destructively interfering at the true location of damage and/or constructively interfering away from damage. If we represent the baseline-subtracted acquired waveform from each transducer pair on node according to its complex analytic signal, then the

test statistic for the incoherent (“phase ignorant”) detector for damage at \mathbf{x} reduces to

$$T_I(\mathbf{x}) = \sum_{m=1}^M \left| w_m(t - \tau(m, \mathbf{x})) \right| \quad (4)$$

where $\tau(m, \mathbf{x})$ is time of flight from transducer pair m to \mathbf{x} .

In the case where the relative phase velocity between transducer pairs is the same, the delayed waveforms can be combined coherently, without enveloping, which is referred to as coherent beamforming. The test statistic for the coherent detector can then be expressed as:

$$T_C(\mathbf{x}) = \left| \sum_{m=1}^M w_m(t - \tau(m, \mathbf{x})) \right| \quad (5)$$

where the magnitude is taken after summation rather than before. Coherent beamforming is ideal since the summation of the delayed waves tend to destructively combine at all locations except the true location of damage. However, in order for the average phase velocities along the path to each region of the structure to be the same, the transducers must be very closely spaced (less than a characteristic interrogation wavelength apart), limiting their coverage of the structure. In practice, for narrowband signals, the time delays are substituted by computationally faster phase shifts. As such, arrays of sensors that make use coherent beamforming, such as those packaged in each MD7 node, are referred to as phased arrays.

Each sensor node implemented by MDC involves a single actuating transducer surrounded by six sensing transducers. Across the transducers in each node, the average phase velocity along the path to any given region is approximately equal, allowing for coherent beamforming. From node to node, however, the average phase velocity is generally not equal and as such the scattered signals must be combined incoherently. This hybrid approach enables both effective imaging through coherent beam forming within each node as well as effective coverage of large areas through the placement of multiple nodes.

$$T_H(\mathbf{x}) = \left| \sum_{n=1}^N \left| \sum_{m=1}^6 w_{nm}(t - \tau(n, m, \mathbf{x})) \right| \right| \quad (6)$$

Figure 4 shows a graphical representation of the summation process. The scans on the left are the result of coherent summation of the individual sensing-transducers’ measurements with appropriate time delays while the image on the right shows the result of the incoherent summation of multiple MD7 nodes.

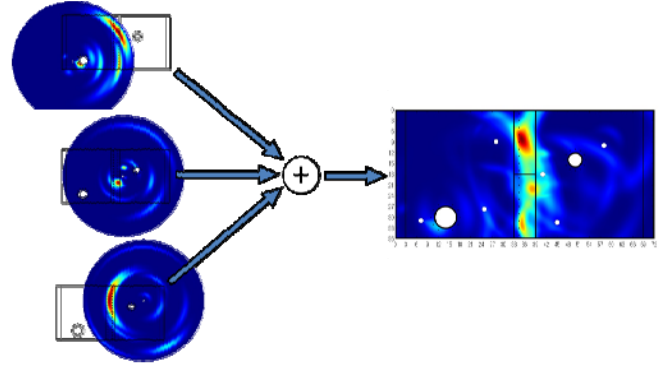


Figure 4: Summation of multiple single-node radial scans

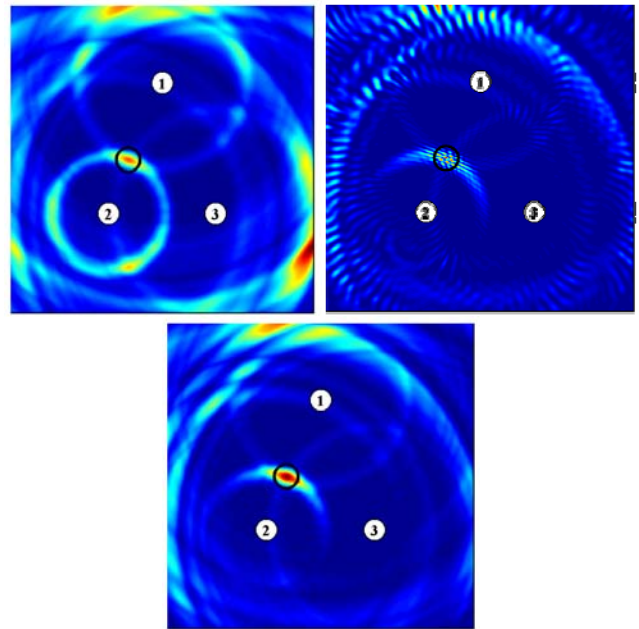


Figure 5: Incoherent (left) coherent (right) and hybrid (bottom) imaging using three nodes. The 0.25 inch disc magnet is located at the center of the open black circle.

Figure 5 shows a summary of results from these three imaging approaches for detecting a 0.25 inch magnet added to a three foot square plate. As shown, with coherent beamforming, a single node can identify both range and bearing of wave-scattering damage. Sensing systems that are not capable of coherent beamforming, such as sparse transducer arrays, can only identify range to a target, forcing them to rely on multiple, widely spaced, sensing elements in order to triangulate the damage location. This significantly reduces the necessary instrumentation footprint of the MD7 system when compared to traditional ultrasonic guided wave systems.

4.2 Matched Pursuits

One of primary and most unique aspects of the present data processing approach is the using of matching pursuit algorithms for identifying scatter targets. This is done by decomposing the 2D radial scans for each node into a sum of wave reflection packets, so that the scans can be approximated as

$$I^*(r, \theta) = \sum_n a_n K(r - r'_n, \theta - \theta'_n) \quad (7)$$

where a'_i , r'_i and θ'_i are the maximum likelihood estimates of the amplitude, range and bearing of the largest wave reflection and $K(r, \theta)$ is the wave reflection shape function. The wave reflection shape function depends on the shape and frequency of the excitation pulse as well as the layout of the sensing array within each node.

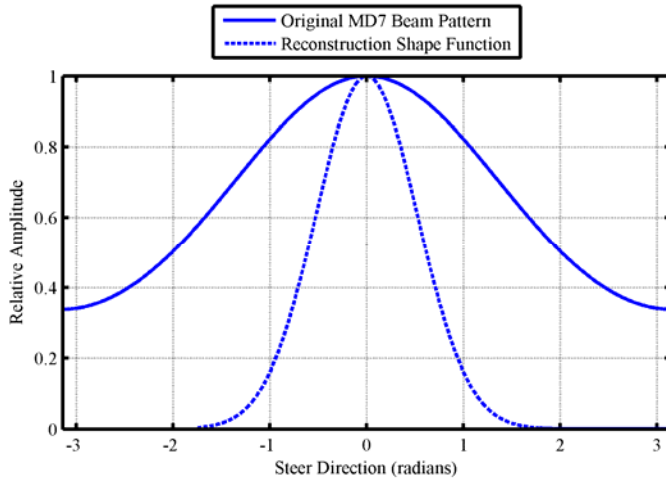


Figure 6: Beam pattern for 46 mm wavelength wave incident on the MD7 node

In the case of the MD7 node and the pulse width and frequency used in this test, the shape function can be expressed as

$$K(r, \theta) = \exp\left(\frac{-r^2}{2\sigma_r^2}\right) B(\theta) \quad (8)$$

where σ_r^2 is the width of the excitation pulse and is the beam pattern for a wave incident at broadside (zero degrees). The beam pattern is graphed in Figure 6 (solid line) for the primary wavelength used in testing and the circular sensor configuration on the MD7 nodes. The function represents the leakage of a wave incident at zero degrees into other look directions in the radial scan.

The amplitudes, ranges, and bearings of the wave packets are estimated according to the following matching pursuit algorithm:

1. Identify range, bearing, and amplitude corresponding the global maximum of the radial scan image

$$\{r'_n, \theta'_n\} = \arg \max_{r, \theta} I(r, \theta), \quad a'_n = I(r'_n, \theta'_n) \quad (9)$$

2. Subtract the reconstructed wave packet from the radial scan image

$$I(r, \theta) = I(r, \theta) - a_n K(r - r'_n, \theta - \theta'_n) \quad (10)$$

3. Repeat until the error the between the original image and the reconstructed image reaches a minimum

$$N_{Optimal} = \arg \min_N \sum_{r, \theta} \left(I(r, \theta) - \sum_{n=1}^N a_n K(r - r'_n, \theta - \theta'_n) \right)^2 \quad (11)$$

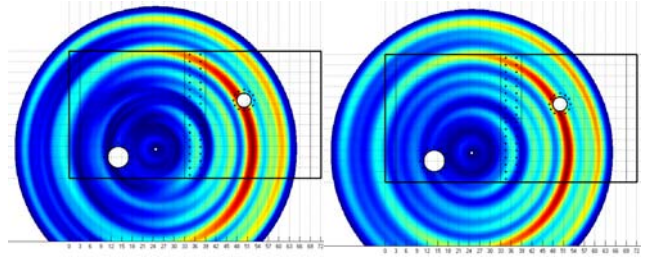


Figure 7: Original radial scan for single MD7 node (left) and reconstructed scan (right) using reflection packets estimated using matching pursuit algorithm

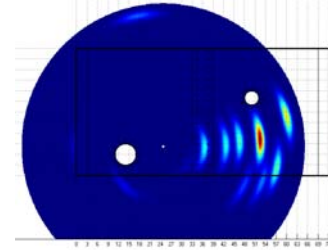


Figure 8: Reconstructed scan using narrowed-angle reflection shape function

Figure 7 shows the original radial scan for a single MD7 node (top) and a reconstructed image using discrete reflection packets. As can be seen in the figure, the natural wave reflection shape functions leave a large degree of ambiguity in the target bearing. When the responses from multiple nodes are combined, this can lead to significant error in the target localization. To remedy this, the imaging software alternatively reconstructs the images using the same estimated target amplitudes, ranges, and bearings, but with a narrower shape function, as depicted in Figure 6 (broken line). Figure 8 shows the same reconstructed radial scan image using the narrower shape function. Here, the precise locations of the potential reflection targets can be more readily identified.

5. PATH TO PROGNOSTICS

The development of sensors, hardware and diagnostic algorithms alone is not sufficient to achieve desired benefits for SHM. At best, current SHM systems can provide diagnostic information—typically in a proprietary and/or stand-alone format—and furthermore require a team of subject-matter experts to properly devise an installation strategy, calibrate algorithms and interpret the data. It is evident that for SHM system to be practically deployed as part of an integrated system health management (ISHM), tools must be created for SHM life-cycle management (LCM). To that end, SHM-LCM software has been developed to manage the cradle-to-grave life-cycle of an SHM system for generic applications. The initial version focuses on the MD7 pulse-echo style guided-wave SHM sensors previously described; however, the intent is to develop a framework that could eventually be sensor agnostic. There are 4 core modules to facilitate critical roles: Optimization, Calibration, Visualization, and Action.

The Optimization module seeks to devise optimal sensor placement (using the Bayesian principals previously described) and excitation parameters in order to achieve probability of detection (POD) coverage requirements. This module is fueled by a 3D mesh of the structure to be monitored, and allows a user to impose POD distribution through a graphical user interface (GUI), resulting in a list of grid point to locate SHM sensors to meet these requirements.

The Calibration module is designed to guide a user through a series of material level tests in order to customize diagnostic algorithm variables (using the hybrid beamforming approach as previously described) to the system being designed. The output would be a file to be uploaded onto the SHM system diagnostic server (could be a local data accumulator or remote slot-card in a HUMS or AHM system box) that would take individual sensor raw data, translate it to diagnostic results, and fuse data from both active and passive sensor sources to compile a complete diagnostic picture including both structural and sensor health with quantified uncertainty.

5.1 Visualization Software

The Visualization module is dedicated to generating a diagnostic composite picture based on data downloaded from the diagnostic server. A prototype of the visualization tool was developed to help present ultrasonic imaging data to the user, seen in Figure 9. The idea is that the input to the software would be a) finite element mesh from a designer, and b) probability distribution as a function of damage size from diagnostics algorithms. The software would then stitch these results to the mesh and allow 3D visualization and manipulation (zoom, rotate, etc) of the diagnostic results on the actual geometry. Controls in the form of “sliders” are provided to the user to be able to control key

algorithms variables, as well as adjust the upper and lower visualization thresholds. The intention is that eventually users will be able to toggle between probability of damage distributions for various calibrated damage modes within the GUI as well, as separated using time-windowed pattern recognition techniques such as K nearest neighbor (KNN).

This all contributes to providing a system that “feels” more like conventional NDE, where, while there are default settings, a knowledgeable/advanced user could refine the results for a more precise location, or alternatively find smaller damage that is hidden by the effects of large damage response. A screen-shot of the full three dimensional visualization of the software is shown in Figure 10.

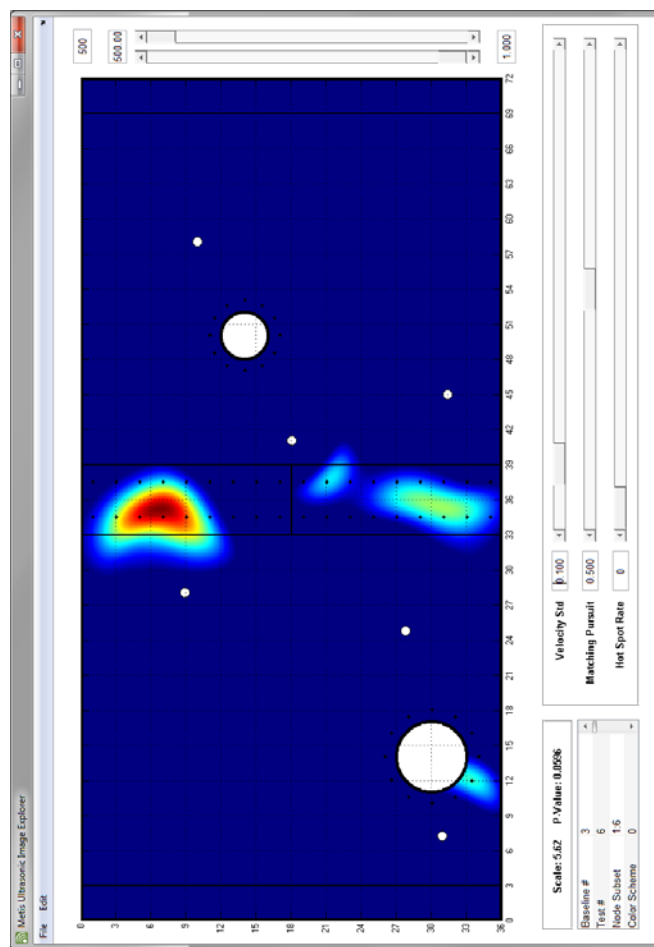


Figure 9: Prototype diagnostic visualization software (2D)

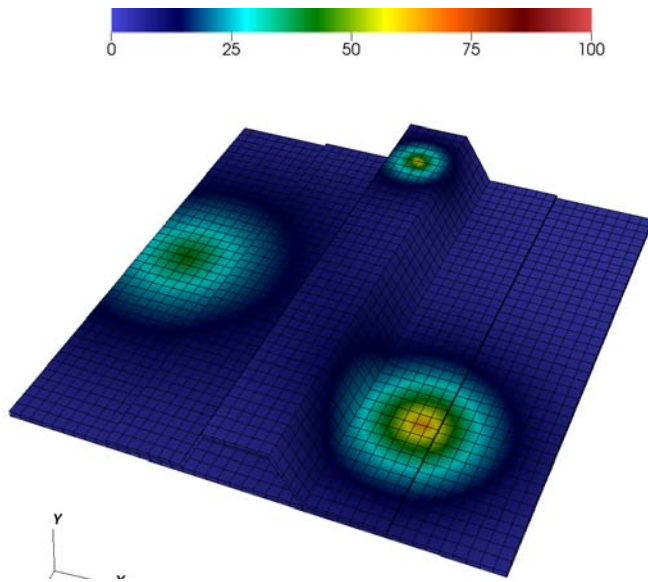


Figure 10: Prototype diagnostic visualization software (3D)

5.2 Residual Performance

The final Action module completes the life-cycle management by providing users with guides for responses to the diagnostic results. This includes the generation of residual performance plots (ultimate load or deflection for example) as a function of probability of damage, using an embedded finite element engine that compares baseline models to those with reduced material properties or un-tied coincident nodes. Using this module, users can weigh detection confidence against the impact to the vehicle's capabilities; and eventually this type of methodology could be embedded for real-time usage to enable fly-by-feel methodologies. Finally, repair optimization tools are planned to be incorporated in order to suggest means of restoring original performance for an assumed damage confidence-level design point.

6. CONCLUSION

This paper presents the framework of a software tool being developed to manage the life-cycle for SHM systems. The core elements include optimization, calibration, visualization and action modules. Much of the present research has focused on the optimization piece, using a Bayesian risk minimization approach to determine optimal sensor placement to minimize false positives while providing the desired coverage, attempting to use the minimum number of sensors to convey efficiency. Furthermore work was performed with regards to diagnostic algorithm calibration using a hybrid beamforming method. Finally, a visualization approach was demonstrated with an intuitive and fast GUI for near-real time display of

diagnostic results with NDE-like controls. Overall, while the proposed framework was demonstrated using pulse-echo style guided wave sensors, it was developed such that it will be able to become sensor agnostic, and also be able to easily link up with prognostic methods for evaluating residual performance. The SHM-LCM software will enable SHM systems to be incorporated into ISHM by engineers rather than experts, making the technology more accessible, and commercially practical.

ACKNOWLEDGEMENT

This work was sponsored by the Office of Naval Research, under contract N00014-10-M-0301, monitored by Dr. Ignacio Perez. The authors would like to additionally thank Dr. Liming Salvino, Dr. Roger Crane, Dr. Mark Seaver and Dr. Benjamin Grisso for their guidance during this program. Metis Design Corporation was the prime contractor under this STTR topic N10A-T042, and University of California San Diego was the subcontracted research institute.

REFERENCES

- Fasel T. R., Kennel M. B., M. D. Todd, E. H. Clayton, M. Stabb, and G. Park, (2009). "Damage State Evaluation of Experimental and Simulated Bolted Joints Using Chaotic Ultrasonic Waves," *Smart Structures and Systems*, vol 5(4), pp. 329-344.
- Flynn E. and M. D. Todd (2010). "Optimal Placement of Piezoelectric Actuators and Sensors for Detecting Damage in Plate Structures," *Journal of Intelligent Material Structures and Systems*, vol. 21(2), pp. 265-274.
- Holmes C, Drinkwater BW, Wilcox PD (2005). Post-processing of the full matrix of ultrasonic transmit-receive array data for non-destructive evaluation. *NDT and E International*. vol. 38, pp.701-711.
- Kay SM (1998). *Fundamentals of Statistical signal processing, Volume 2: Detection theory*. Prentice Hall PTR.
- Kessler S.S. and P. Agrawal.(2007) "Application of Pattern Recognition for Damage Classification in Composite Laminates." *Proceedings of the 6th International Workshop on Structural Health Monitoring*, Stanford University
- Kessler S.S. and A. Raghavan (2008). "Vector-Based Localization for Damage Position Identification from a Single SHM Node." *Proceedings of the 1st International Workshop on Prognostics & Health Management*, Denver, CO
- Kessler S.S. and A. Raghavan (2009). "Vector-based Damage Localization for Anisotropic Composite Laminates." *Proceedings of the 7th International Workshop on Structural Health Monitoring*, Stanford University

A Study on the parameter estimation for crack growth prediction under variable amplitude loading

Sang Hyuck Leem¹, Dawn An², Sangho Ko³, and Joo-Ho Choi⁴

^{1,2} Department of Aerospace & Mechanical Engineering, Korea Aerospace University, GoyangCity, Gyeonggido, 412-791, Korea

sanghyuck@naver.com
skal34@nate.com

^{3,4} School of Aerospace & Mechanical Engineering, Korea Aerospace University, GoyangCity, Gyeonggido, 412-791, Korea

sanghoko@kau.ac.kr,
jhchoi@kau.ac.kr

ABSTRACT

Bayesian formulation is presented to address the parameters estimation under uncertainty in the crack growth prediction subjected to variable amplitude loading. Huang's model is employed to describe the retardation and acceleration of the crack growth during the loadings. Model parameters are estimated in probabilistic way and updated conditional on the measured data by Bayesian inference. Markov Chain Monte Carlo (MCMC) method is employed for efficient sampling of the parameter distributions. As the model under variable amplitude loading is more complex, the conventional MCMC often fails to converge to the equilibrium distribution due to the increased number of parameters and correlations. An improved MCMC is introduced to overcome this failure, in which marginal PDF is employed as a proposal density function. A center-cracked panel under a mode I loading is considered for the feasibility study. Parameters are estimated based on the data from specimen tests. Prediction is carried out afterwards under variable amplitude loading for the same specimen, and validated by the ground truth data.

Key Words : Prognostics and Health Management (PHM), Markov Chain Monte Carlo (MCMC), Crack growth, Variable amplitude loading.

1. INTRODUCTION

Although the reliability-based design technology for lifecycle is in its mature stage, it has its limited value due to the inability to account for the unexpected incidences during the in-service condition. Besides, critical systems such as aircraft tend to be operated without retirement even after the

design lives. In such cases, efficient maintenance techniques should be incorporated during the operation. Frequent preventive maintenance can, however, increase operating cost significantly, especially for aging aircraft. Recently, prognostics and health management (PHM) techniques are drawing considerable attention, which detect, monitor and predict the damage growth, from which only the faults indicating impending failure are repaired. As a result, condition-based maintenance (CBM) can be achieved, which significantly reduce the number of maintenance visits and repairs.

Prognosis of crack growth is one of the active research topics in the PHM study because the physical model underlying the feature is relatively well known. Numerous literatures have been devoted to this topic, mainly focused on the probabilistic methods to address the associated uncertainties. Orchard and Vachtsevanos (2007) introduced an on-line particle-filtering-based framework for failure prognosis, and applied to a crack growth problem of UH-60 planetary carrier plate. They assumed that the crack growth is described by a simple Paris model and the model parameters are known a priori, which is questionable in practical applications. Cross et al. (2007) developed a Bayesian technique for simultaneous estimation of the equivalent initial flaw size (EIFS) and crack growth rate distributions. AFGROW is used for the crack growth calculation for the fastener hole crack under constant amplitude load. Coppe et al. (2009, 2010) employed Bayesian formulation using the Paris model in which the model parameters are estimated and updated conditional on the measured crack data. A center-cracked panel under a mode I loading is considered for the study. An et al. (2011) conducted similar study by introducing Markov Chain Monte Carlo (MCMC) method for more efficient sampling of the parameters' distribution. They paid particular attention to the parameters correlation as well as the imprecise data due to the noise and bias, which may make

* This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

the Bayesian estimation more difficult. It should be noted that all the previous studies have employed Paris model for the sake of simplicity which holds good under a constant amplitude loading.

In this paper, the study by An et al. (2011) is extended to the case of variable amplitude loading, which involves more parameters in the crack growth model. The feasibility of Bayesian approach is studied to cope with the increased parameters in which the correlations are encountered. We have experienced that the MCMC does not work well, i.e., fails to converge at the equilibrium distribution. An improved MCMC method is introduced to relieve this problem by employing marginal PDF as a proposal density function. Feasibility of the method is illustrated by a center-cracked panel under a mode I loading with constant and variable amplitudes, respectively. In the case of variable amplitude loading, the unknown model parameters are estimated based on the crack data by lab specimens under multiple set of constant amplitude loadings. The prognosis under variable loading is then conducted for the same specimen using the obtained parameter samples, and the remaining useful life (RUL) is predicted accordingly.

2. CRACK GROWTH MODEL

When the load is applied in a constant amplitude, Paris model best describes the crack growth:

$$\frac{da}{dN} = C(\Delta K)^m, \quad \Delta K = \Delta\sigma \cdot \alpha\sqrt{\pi a} \quad (1)$$

where a is the half crack size, N is the number of cycles (flights), ΔK the range of stress intensity factor (SIF) and α the geometric correction factor. In the case of the variable amplitude loading, however, the crack growth behavior is significantly different from that under constant loading, presenting the crack growth retardation and acceleration caused by the overload. Numerous models have been developed to adequately describe this behavior. A model based on the crack closure approach, which considers plastic deformation and crack face interaction in the wake of the crack, was proposed by Eiber (1971). Willenborg (1971) and Wheeler (1972) proposed other models based on the calculations of the yield zone size ahead of the crack tip. In this paper, crack growth model by Huang et al. (2007) is used, which is based on a modified Wheeler model to account for the overload and underload effect. Huang's model consists of two parts, one being the scaling factor M_R which accounts for the crack growth under constant amplitude loading and the other the correction factor M_p which accounts for the loading sequence interaction such as retardation and acceleration under variable amplitude. The expression is given as follows.

$$\frac{da}{dN} = C[(\Delta K_{eq0})^m - (\Delta K_{th0})^m] \quad (2)$$

$$\Delta K_{eq0} = M_R M_p \Delta K \quad (3)$$

$$M_R = \begin{cases} (1-R)^{-\beta_1} & (-5 \leq R < 0) \\ (1-R)^{-\beta} & (0 \leq R < 0.5) \\ (1.05 - 1.4R + 0.6R^2)^{-\beta} & (0.5 \leq R < 1) \end{cases} \quad (4)$$

Here, R is the stress ratio, ΔK_{eq0} and ΔK_{th0} are equivalent and threshold SIF range respectively, C, m are the Paris model parameters, and β, β_1 are the shaping parameters for M_R . The parameters $C, m, \Delta K_{th0}, \beta$ and β_1 are the fitting parameters under a constant amplitude loading, which determines the relationship between the crack growth rates da/dN and SIF range ΔK . The correction factor M_p is given by

$$M_p = \begin{cases} \left(\frac{r_y}{a_{OL} + r_{OL} - a - r_\Delta} \right)^n & (a + r_y < a_{OL} + r_{OL} - r_\Delta) \\ 1 & (a + r_y \geq a_{OL} + r_{OL} - r_\Delta) \end{cases} \quad (5)$$

where

$$r_y = \alpha \left(\frac{K_{\max}}{\sigma_y} \right)^2 \quad (6)$$

$$r_{OL} = \alpha \left(\frac{K_{\max}^{OL}}{\sigma_y} \right)^2 \quad (7)$$

where r_y is the plastic zone size ahead of the crack tip, r_Δ is the increment in the plastic zone size due to the underload following an overload, n is a shaping parameter determined by fitting to the test data under variable amplitude loading, and parameters with the subscript OL denote those under the overload. In Eq.(6) and Eq.(7), α is the plastic zone size factor which is dependent upon the constraints around the crack tip and the maximum applied stress, yield strength of the material, and specimen thickness (Voorwald et al. 1991). The size of the each plastic zone is calculated in terms of the applied maximum SIF and yield strength σ_y . The crack growth under variable amplitude loading is accounted for by incorporating the correction factor M_p after decomposing the variable loading into the successive series of different constant amplitude loadings. Consequently, only the parameters $C, m, \Delta K_{th0}, \beta$ and β_1 are the unknown parameters to be estimated in this study.

3. MARKOV CHAIN MONTE CARLO FOR PARAMETER ESTIMATION

In this study, Bayes rule is used to account for the uncertainties in the parameters estimation (Bayes, 1763):

$$p(\boldsymbol{\theta} | \mathbf{y}) \propto L(\mathbf{y} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) \quad (8)$$

where $L(\mathbf{y} | \boldsymbol{\theta})$ is the likelihood of observed data \mathbf{y} conditional on the given parameters $\boldsymbol{\theta}$, $p(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$, and $p(\boldsymbol{\theta} | \mathbf{y})$ is the posterior distribution of $\boldsymbol{\theta}$ conditional on \mathbf{y} . The equation states that our degree of belief on the parameter $\boldsymbol{\theta}$ is expressed as posterior PDF in light of the given data \mathbf{y} . In general, the posterior distribution is given by complex expression in terms of the parameters, of which the sample drawing is cumbersome, and prohibiting the use of standard techniques of probability functions. MCMC has been recognized as an effective sampling method, which is based on a Markov chain model of random walk with the stationary distribution being the target distribution. Metropolis-Hastings is the most typical variants of the MCMC algorithm:

1. Initialise $x^{(0)}$
2. For $i = 0$ to $nm - 1$
 - Sample $u \sim U_{[0,1]}$
 - Sample $x^* \sim q(x^* | x^{(i)})$
 - if $u < A(x^{(i)}, x^*) = \min \left\{ 1, \frac{p(x^*)q(x^{(i)} | x^*)}{p(x^{(i)})q(x^* | x^{(i)})} \right\}$ (9)
 - $x^{(i+1)} = x^*$
 - else
 - $x^{(i+1)} = x^{(i)}$

In this equation, $x^{(0)}$ is the initial value of an unknown parameter to estimate, nm is the number of iterations or samples, U is a uniform distribution, $p(x)$ is the posterior distribution (target PDF), and $q(x^* | x^{(i)})$ is an arbitrary chosen proposal distribution which is used when a new sample x^* is to be drawn conditional on the current point $x^{(i)}$. Uniform or Gaussian distribution at the current point are the most common choices for the proposal distribution. Success and failure of the algorithm relies heavily on a proper design of the proposal distribution. In order to illustrate this, a target distribution of x is considered (Andrieu et al, 2003):

$$p(x) \propto 0.3 \exp(-0.2x^2) + 0.7 \exp(-0.2(x-10)^2) \quad (10)$$

As the candidates of proposal distribution, normal distributions with three different standard deviation, $\sigma = 1$, $\sigma = 10$ and $\sigma = 100$, are attempted. The shapes of each distribution are compared in Figure 1(a). The MCMC sampling results using each three proposal distributions with the number of samples $nm = 5000$ are shown in Figure

1(b)~(d), respectively. Only the proposal distribution with $\sigma = 10$ gives acceptable result. In the general case with increased parameters and correlations, however, this would be much more difficult.

An improved MCMC method is introduced in this study, which is to employ a marginal PDF as a proposal distribution:

1. Initialise $x^{(0)} = \text{mean}(q(x))$
2. For $i = 0$ to $nm - 1$
 - Sample $u \sim U_{[0,c]}$
 - Sample $x^* \sim q(x)$
 - if $u < A(x^{(i)}, x^*) = \min \left\{ 1, \frac{p(x^*)}{p(x^{(i)})} \right\}$ (11)
 - $x^{(i+1)} = x^*$, $i = i + 1$
 - else
 - $i = i$

where $q(x)$ is the marginal PDF of x defined by

$$q(x_i) = \int \dots \int p(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_{n_p}) dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_{n_p} \quad (12)$$

Conventional way to construct the marginal PDF requires intensive computation which requires large number of joint PDF evaluation. In this paper, a simpler approach, which employs Latin Hypercube Sampling (LHS), is used to facilitate efficiency because the marginal PDF needs not be precise in view of the proposal density function.

In the algorithm (11), unlike the conventional MCMC, if the new sample x^* is not accepted, the $i+1$ 'th sample is not assigned and the sampling is repeated until $i+1$ 'th sample satisfies the MH criteria, which results in a little longer

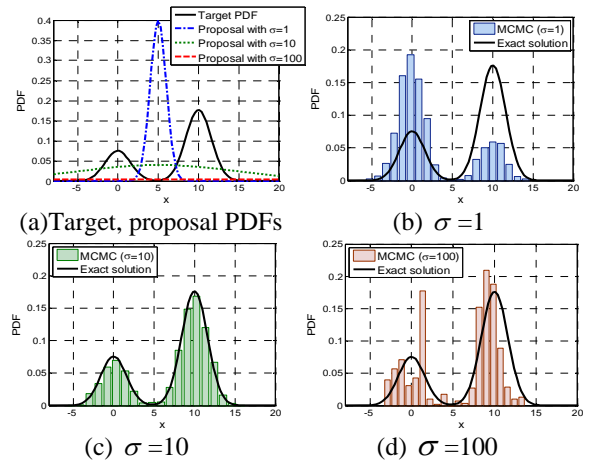


Figure 1. MCMC sampling results of the target PDF given by Eq. (10)

computing time. The uniform distribution, $U_{[0,1]}$ in the conventional MCMC is replaced here by $U_{[0,c]}$ where c is a constant less than 1. By the authors' experience, it was found that as c gets smaller, the overall time was decreased dramatically, while the obtained samples distribution did not change much.

4. CRACK GROWTH UNDER CONSTANT AMPLITUDE LOADING

In order to verify the new MCMC method, the data generated with fixed parameter values are used. Crack growth of a center-cracked panel of Al 7075-T6 under a mode I loading as shown in Figure 2 is considered. Assuming the effect of finite plate size is ignored, Paris model predicts the crack growth in terms of the fatigue cycles in the closed form expression as:

$$a(N) = \left[NC \left(1 - \frac{m}{2} \right) (\Delta\sigma\sqrt{\pi})^m + a_i^{1-\frac{m}{2}} \right]^{\frac{2}{2-m}} \quad (13)$$

where a is the half crack size at cycle N , C and m are the two damage growth parameters to be estimated, a_i is the initial crack size which is assumed to be known, and $\Delta\sigma$ is the stress range due to the fatigue loading. Synthetic curve is generated for the case $a_i = 10\text{mm}$ and $\Delta\sigma = 78.6\text{MPa}$. Assuming that the true parameters, m_{true} and C_{true} are given by 3.8 and $1.5\text{E-}10$ respectively, crack sizes are calculated according to Eq. (13) for a given N . Then, measurement errors with a deterministic bias $b = 2\text{mm}$ and random noise $N(0, \sigma = 1.33)$ are added intentionally to the synthetic curve for the generated data. 10 sets of generated data are made at the interval of 100 cycles. In this case, the unknown parameters consist of the two model parameters m, C and the two measurement errors b, σ . The joint posterior distribution of these parameters is given by

$$p(m, C, b, \sigma) \propto \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^{10} f \cdot p(m) \cdot p(C) \quad (14)$$

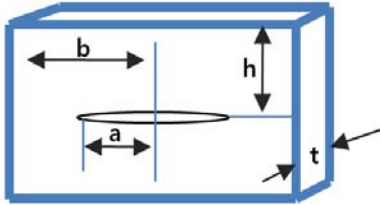
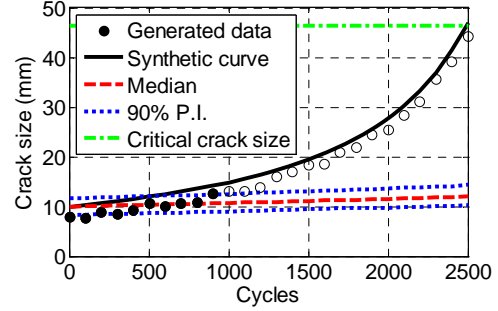
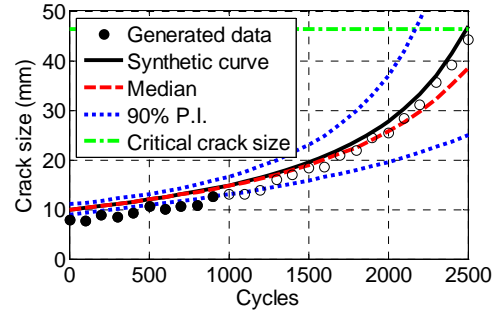


Figure 2. Specimen geometry ($t=4.1$, $b=152.5$, $a=6$ (mm))



(a) Conventional MCMC.



(b) Improved MCMC.

Figure 3. Prediction of the crack growth

where f and $p(m), p(C)$ are the likelihood and prior PDFs of the two parameters respectively, given by

$$f = \exp \left[-\frac{1}{2\sigma^2} \sum_{k=1}^{10} (a_{meas_k} - a(N_k) - b)^2 \right] \quad (15)$$

$$p(m) = U_{[3.3, 4.3]}, \quad p(C) = U_{[\log(5 \times 10^{-11}), \log(5 \times 10^{-10})]}$$

The synthetic curve and the generated data are plotted as black curve and solid dots with 10 numbers in Figure 3 respectively. The unknown parameters are to be estimated conditional on this data based on the MCMC process with the number of samples being 5000. Using the conventional MCMC, proper sampling could not be achieved in spite of lot of trials. One instance of such result is given in Figure 3(a). In Figure 3(a), the incorrect prediction using the failed samples is also given, in which the three dashed curves denote the median and 90% confidence bounds obtained from the distribution respectively. The green horizontal line denotes the critical crack size. On the other hand, the result of the improved MCMC is shown in Figure 3(b), which is instantly obtained at one attempt. The obtained PDF shapes look quite plausible and the correlation between m and C is also identified clearly. The posterior predictive distribution of the crack growth obtained by the sampling results of the unknown parameters is shown in Figure 4. The improved MCMC predicts the crack growth quite well, following the synthetic curve by correcting the bias while the conventional MCMC could not. Therefore, the improved

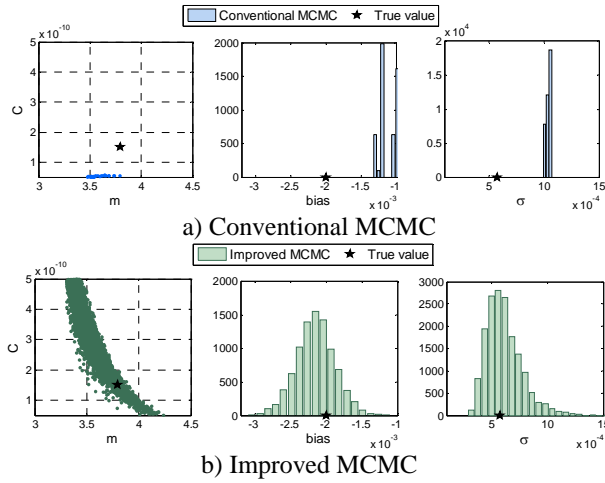


Figure 4. Posterior PDFs of four parameters in the crack growth problem

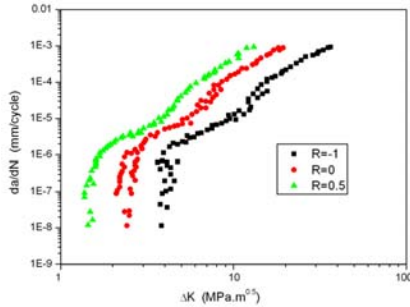


Figure 5. Fatigue crack growth data under constant amplitude loading for Al 7075-T6 (Huang et al, 2007)

MCMC is verified by predicting the synthetic curve with correct parameter estimation.

5. CRACK GROWTH UNDER VARIABLE AMPLITUDE LOADING

In the prognosis of crack growth under variable amplitude loading, the unknown model parameters C , m , ΔK_{th0} , β and β_1 are to be estimated conditional on the measured crack data under study. In this study, the unknown model parameters are regarded as the intrinsic property of the material such as the Elastic modulus. Therefore, the unknown model parameters under constant amplitude loading are assumed as identical to those under variable amplitude loading. In view of this, data by Huang et al. (2007) are used for the prognosis, in which the cracks are grown for the lab specimens of Figure 2 under multiple sets of constant amplitude mode I loadings.

Assuming the error between the data and true crack growth model follows Gaussian distribution with $N(0, \sigma)$, the joint posterior distribution of the parameters is given by Eq. (8)

which θ denote $C, m, \beta, \beta_1, \Delta K_{th0}$ and σ , and \mathbf{y} are the measured crack data. L is the likelihood given by

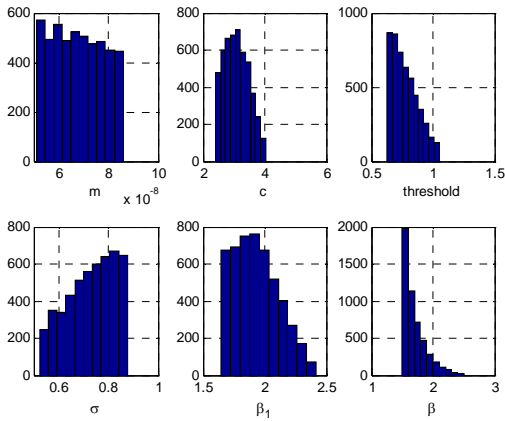
$$L(\mathbf{y} | C, m, \beta, \beta_1, \Delta K_{th0}, \sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^k \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^k (y_{estimation}^i - y_{test}^i)^2 \right] \quad (16)$$

MCMC simulation is implemented to obtain the samples that satisfy the distribution. In this case, the conventional MCMC does not work at all due to the large number of parameters, and fails to obtain the target distributions. Even the improved MCMC gives inadequate distributions as given in Figure 6(a). The reason may be attributed to the Eq. (2)~Eq.(4), in which the parameters β, β_1 exist only when $R \neq 0$ whereas the data set include the case $R=0$. Ignoring this characteristics and taking all three data set equally into account in Eq.(16) leads to the improper marginal PDF. In order to resolve this issue, following four steps are taken during the MCMC simulation.

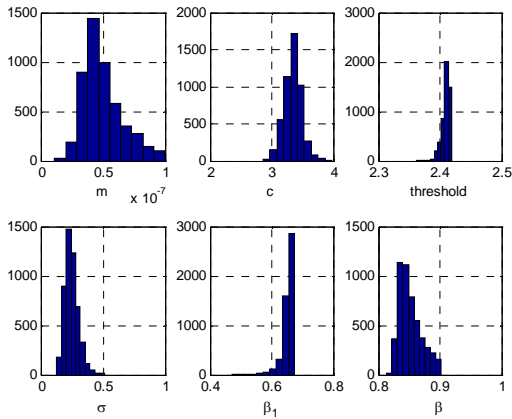
1. The marginal PDFs of $C, m, \Delta K_{th0}$ are constructed from $R=0$ data set. In this process, β, β_1 is not necessary since $R=0$ makes M_R independent on β, β_1 .
2. The ranges of $C, m, \Delta K_{th0}$ are given from the percentiles of the marginal PDF of $C, m, \Delta K_{th0}$.
3. The marginal PDFs of β and β_1 are constructed from the remaining two sets $R=-1$ and $R=0.5$ under the ranges of $C, m, \Delta K_{th0}$ of the process 2.
4. All the marginal PDFs thus obtained are then used in the main process of improved MCMC as given by (11).

As a result, Figure 6(b) is obtained, in which the distributions of the parameters θ exhibit plausible shape, and represent our degree of confidence due to the uncertainties caused by the insufficient data and measurement errors.

Once the distributions are obtained by the MCMC, the prognosis under variable amplitude loading is conducted using the obtained parameter samples. This is just to implement the crack growth simulation by integration of Eq.(2) to obtain the future crack size distribution using each of the parameter samples. The remaining useful life (RUL) can be predicted from this result. The same specimen is used in this study since the actual data of crack growth are available by Huang et al. (2007) under the variable loading condition as a ground truth data. The loading condition for prognosis process is given in Figure 7, in which a single cycle consists of the p numbers of repeated load between 3.48~68.13 MPa and the q numbers of overload with 3.48~103.02 MPa. This loading condition is repeatedly applied to the specimen generating total load cycles. Two



(a) Sample data from direct application of improved MCMC



(b) Sample data after taking the four step process in the improved MCMC

Figure 6. Histogram of samples for the parameters generated by the improved MCMC method

cases of $p = 50$, $q = 1$ and $p = 50$, $q = 6$ are considered. The results of the predictive simulation are shown in Figure 8, in which each blue curve represents a single result using realized parameters while the red curve represents the ground truth data made by the test of identical loading condition. Figure 9 also represents the confidence bounds obtained from the predictive distribution. The width of the curve in this figure may be attributed to the uncertainty of insufficient data and measurement errors. The RUL distribution shown in Figure 10 is obtained by calculating the cycles at which the crack of each sample grows to a critical crack size. 10% percentile as well as the true RUL

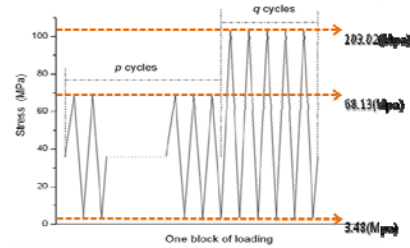


Figure 7. Variable amplitude loading

values are indicated by the marks respectively. Recall that in this study, the parameters were first estimated using the three specimens under constant amplitude loadings, followed by prognosis for the fourth specimen under variable loadings using the estimated parameters. The test data of the last specimen was used just for validation of the prognosis.

6. CONCLUSION

In this paper, Bayesian formulation is presented to identify the uncertain parameters in the crack growth problem under variable amplitude loading. Huang's model is employed to describe the retardation and acceleration of the crack growth during the loadings. As the conventional MCMC does not work well in the case of increased parameters and correlations as in this problem, improved MCMC method is introduced by employing marginal PDF as a proposal density function. Feasibility of the method is illustrated by a center-cracked panel under a mode I loading with constant and variable amplitudes, respectively. In the case of variable amplitude loading, parameters are first estimated based on the data from specimen tests under a multiple constant amplitude loadings, and prognosis is followed based on the parameters with another specimen under variable loading. The result is validated by the actual test data. The drawback of this approach is that the model parameters are identified by the lab experiments, and are used for the prognosis of a real part (although, in this case, the same specimen is chosen), of which the material and operating conditions may be somewhat different. Therefore, the estimated RUL has wide range to represent the general life of the entire specimen.

More desirably, the measured data from the real part undergoing variable amplitude loading may be utilized for the parameters estimation as well as the prognosis. Additional work toward this direction will be made in the final draft.

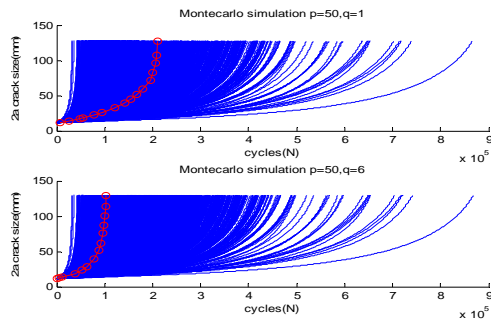


Figure 8. Crack growth simulation under variable amplitude loading using each sample of parameters (red curve :ground truth data)

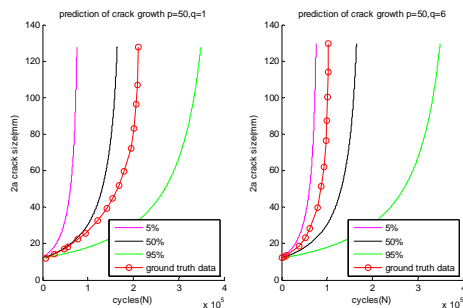


Figure 9. Confidence bounds of crack growth simulation under variable amplitude loading

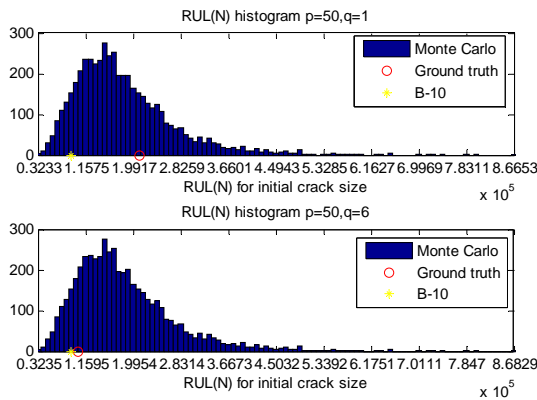


Figure 10. RUL distribution and its 10% percentile value

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2009-0081438).

REFERENCES

Andrieu, C., Freitas, N. D., Doucet, A. & Jordan, M. (2003). An introduction to MCMC for Machine Learning.

Machine Learning, vol. 50(1), pp. 5-43.

Bayes, T. (1763). An Essay towards solving a Problem in the Doctrine of Chances, *Philosophical Transactions of the Royal Society of London*, vol. 53, pp. 370-418.

Coppe, A., Haftka, R. T. & Kim, N. H. (2009). Reducing Uncertainty in Damage Growth Properties by Structural Health Monitoring. *Annual Conference of the Prognostics and Health Management Society 2009* September 27 – October 1, San Diego CA

Coppe, A., Haftka, R. T., & Kim, N. H. (2010). Identification of Equivalent Damage Growth Parameters for General Crack Geometry. *Annual Conference of the Prognostics and Health Management Society 2010*, October 10-16, Portland, Oregon

Cross, R. J., Makeev, A. & Armainios, E. (2007). A comparison of prediction from probabilistic crack growth models inferred from Virkler’s data. *Journal of ASTM International*, Vol. 3(10)

An, D., Choi, C. H. & Kim, N. H. (2011). Statistical Characterization of Damage Growth Parameters and Remaining Useful Life Prediction Using Bayesian Inference, *13th AIAA Non-Deterministic Approaches Conference*, April 4-7, Denver, CO.

Eiber, W. (1971). The significance of fatigue crack closure in fatigue. *ASTM STP*, Vol.486, pp. 230-242

Huang, X., Torgeir, M. and Cui, W. (2007). An engineering model of fatigue crack growth under variable amplitude loading. *A International Journal of Fatigue*. Vol. 30. pp. 1-10

Orchard, M., & Vachtsevanos, G. (2007). A Particle Filtering Approach for On-Line Failure Prognosis in a Planetary Carrier Plate. *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 7(4), pp. 221-227.

Patrick, R. & Orchard, M. (2007). An integrated approach to helicopter planetary gear fault diagnosis and failure prognosis. *Autotestcon 2007 IEEE* , pp. 547-552

Voorwald HJC, Torres MAS. (1991). Modeling of fatigue crack growth following overloads. *International Journal of Fatigue 1991*, Vol.13(5), pp.423-427

Wheeler, OE. (1972). Spectrum loading and crack growth. *Journal of Basic Engineering*, Vol. 94. pp. 181-186

Willenborg, J., Engle, R.M. & Wood, H. A. (1971). A crack growth retardation model using effective stress concept. *AFDL-TM-71-1-FBR Air Force Flight Dynamics Laboratory*.

Sang Hyuck Leem received the B.S. degree of mechanical engineering from Korea Aerospace University in 2011. He is now master student at Department of Aerospace & Mechanical Engineering. His research interest is the condition based maintenance by implementing PHM and its application to the Nuclear power plant.

Dawn An received the B.S. degree and M.S. degree of mechanical engineering from Korea Aerospace University

in 2008 and 2010, respectively. She is now a joint Ph.D. student at Korea Aerospace University and the University of Florida. Her current research is focused on the Bayesian inference, correlated parameter identification and the methodology for prognostics and health management and structural health monitoring.

Joo-Ho Choi received the B.S. degree of mechanical engineering from Hanyang University in 1981, the M.S. degree and Ph.D. degree of mechanical engineering from Korea Advanced Institute of Science and Technology (KAIST) in 1983 and 1987, respectively. During the year 1988, he worked as a Postdoctoral Fellow at the University of Iowa. He joined the School of Aerospace and Mechanical Engineering at Korea Aerospace University, Korea, in 1997 and is now Professor. His current research is focused on the reliability analysis, design for life-time reliability, and prognostics and health management.

A Testbed for Real-Time Autonomous Vehicle PHM and Contingency Management Applications

Liang Tang, Eric Hettler, Bin Zhang and Jonathan DeCastro

Impact Technologies, LLC. Rochester, NY, 14623, U.S.A
Liang.Tang@impact-tek.com

ABSTRACT

Autonomous unmanned vehicles are playing an increasingly important role in support of a wide variety of present and future critical missions. Due to the absence of timely pilot interaction and potential catastrophic consequence of unattended faults and failures, a real-time, onboard health and contingency management system is desired. This system would be capable of detecting and isolating faults, predicting fault progression and automatically reconfiguring the system to accommodate faults. This paper presents a robotic testbed that was developed for the purpose of developing and evaluating real-time PHM and Automated Contingency Management (ACM) techniques on autonomous vehicles. The testbed hardware is based on a Pioneer 3-AT robotic platform from Mobile Robots, Inc. and has been modified and enhanced to facilitate the simulations of select fault modes and mission-level applications. A hierarchical PHM-enabled ACM system is being developed and evaluated on the testbed to demonstrate the feasibility and benefit of using PHM information in vehicle control and mission reconfiguration. Several key software modules including a HyDE-based diagnosis reasoner, particle filtering-based prognosis server and a prognostics-enhanced mission planner are presented in this paper with illustrative experimental results. This testbed has been developed in hope of accelerating related technology development and raising the Technology Readiness Level (TRL) of emerging ACM techniques for autonomous vehicles.*

1. INTRODUCTION

Autonomous unmanned vehicles (AUVs) are finding increasing use in real-world applications ranging from the ground (e.g. unmanned ground vehicles, or UGVs),

* This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

to sea surface (e.g. unmanned surface vehicles, or USVs), underwater (e.g. unmanned undersea vehicles, or UUVs), airborne (e.g. unmanned aerial vehicles, or UAVs), and space exploration missions (e.g. unmanned rovers and unmanned space vehicles). Unmanned missions using these vehicles include surveillance and patrolling, search and rescue, operations in contaminated and denied areas, space exploration and more (Army UAS CoE, 2010; Navy, 2007). Due to communication delay and bandwidth limitations, there has been increasing dependence on AUVs for critical tasks. This makes it vital to assure the performance of the vehicles under off-nominal conditions in an autonomous fashion without relying on remote operators.

In recent years, growing demand for improving the reliability and survivability of autonomous vehicles has led to the development of prognostics and health management (PHM) and automated contingency management (ACM) systems (Vachtsevanos et al, 2006). In this context, the term Automated Contingency Management has been introduced to describe intelligent systems capable of mission re-planning and control reconfiguration based on health diagnostic and prognostic information (Tang et al, 2008). As a new emerging technology, the development of real-time autonomous vehicles PHM and ACM techniques can greatly benefit from a testbed that is built on a real vehicle platform using commercial-off-the-shelf (COTS) computing devices. The use of such a testbed can accelerate the development and raise the Technology Readiness Level (TRL) of the enabling techniques, as well as provide a technology demonstrator for commercialization efforts. This paper presents the development of a ground robotic testbed for real-time autonomous vehicles PHM and ACM techniques. The testbed has been built to fulfill the following objectives:

- (1). to demonstrate the benefits of real-time PHM and ACM technologies for autonomous vehicles in terms of improved reliability, survivability and overall mission success;

(2). to demonstrate the advantages of using PHM information (particularly the prognostic information) in control reconfiguration and mission planning by applying a novel hierarchical ACM architecture;

(3). to demonstrate that real-time implementation of selected diagnostic and prognostic routines are feasible on affordable COTS computing devices;

(4). to raise the TRL of several selected techniques by deploying them on hardware and testing them on field systems in realistic environments;

The rest of this paper is organized as follows. In section 2, the system architecture and common features of a generic, hierarchical PHM-enabled ACM system for autonomous vehicles are briefly introduced. Section 3 presents the development of a robotic testbed on which the techniques described in Section 2 are applied. Hardware configuration and modifications, as well as a Failure mode, effects, and criticality analysis (FMECA) study of selected components and fault simulations are presented. Section 4 presents the real-time PHM and ACM software modules implemented on the testbed with illustrative experimental results. These modules include a diagnostic reasoner based on NASA's Hybrid Diagnostic Engine (HyDE), several particle filter-based real-time prognostic routines, and prognostics-enhanced control configuration and mission re-planning modules. The paper concludes with remarks on the main contributions of the presented work and planned future developments.

2. PHM-ENABLED ACM SYSTEM FOR AUTONOMOUS VEHICLES

Conceptually, an ACM system is a system that is designed to provide the ability to proactively and autonomously adapt to current and future fault and/or contingency conditions while either achieving all or an acceptable subset of the mission objectives. An ACM system is different from a fault tolerant control system mainly in two aspects: 1) it consists of not only low level control reconfiguration, but also high level (mission) planning and optimization; 2) it uses not only diagnostic information, but also prognostic information.

A typical ACM+P (PHM-enabled ACM system) implementation usually utilizes a hierarchical architecture as shown in Figure 1. The PHM and situation awareness modules provide fault diagnostics, prognostics and contingency information to the ACM+P system, which in turn, identifies and executes the optimal fault accommodation and/or mitigation strategies. Note that the PHM system is a precondition for implementing ACM strategies, thus the whole system architecture is referred to as a PHM-enabled ACM system.

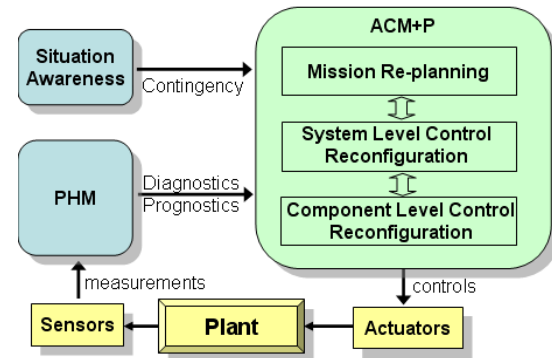


Figure 1: Conceptual PHM-enabled ACM system hierarchy

Some important features of the PHM-enabled ACM system include:

(1) Hierarchical architecture;

A component fault can often be accommodated at different levels in the ACM hierarchy and the decision should be made based on performance requirement and safety consideration. For example, if the left engine on a twin-engine, fixed-wing UAV is experiencing severe degradation, the thrust difference may generate an unwanted yaw movement. This fault can be accommodated at the lowest (component) level by adjusting the fan speed set point value in the engine controller or by adjusting the rudder position in the trajectory following auto-pilot at system level.

(2) Use of redundancy and trade-off;

Typically, it is possible to accommodate faults only in a system with redundancy, either physical redundancy or analytical redundancy. More advanced systems may include online healing concepts, including self-healing. When system performance cannot be totally recovered by the fault accommodation strategies, trade-off of mission objectives has to be made to secure the most important tasks.

(3) Online optimization;

If an ACM system is to be applied to an unmanned vehicle conducting complicated autonomous missions, it is often unavoidable to phrase the solution search as a dynamic optimization problem especially at mission planning level. This optimization problem may need to be solved online to arrive at the optimal strategies constrained by the available performance and resources to meet multiple (sometimes conflicting) mission objectives. It is important to realize that the optimization problems at different levels in an ACM+P system have different time horizons and real-time execution considerations.

(4) Uncertainty management and false alarm mitigation;

The use of prognostic information in the ACM system brings new challenges to both uncertainty management and false alarm mitigation. Since

prognosis projects the current system condition in the future using a prognostic model in the absence of future measurements, it necessarily entails large-grain uncertainty. This uncertainty has to be handled both in high level mission re-planning and middle/low level control reconfiguration modules. The implementation of this ACM architecture on the testbed and the enabling algorithms for the main functional modules are presented in section 4. More details regarding ACM technologies can be found in our previous publications (Tang et al, 2010; DeCatro et al, 2011; Zhang et al, 2011).

3. THE ACM TESTBED

The ACM Testbed is based on a Pioneer 3-AT robotic platform from Mobile Robots, Inc. The dimensions of the Pioneer 3-AT robot (without additional computer and sensors) are about 20" long by 19" wide by 11" high and it weighs about 26 pounds (see Figure 2). The robot is a four wheeled "skid steered" design with 8.5" diameter tubeless pneumatic rubber wheels. This means that the wheels are fixed in place and it is driven in a "tank drive" fashion. The wheels on either side of the robot are driven independently at different speeds to provide the freedom to turn. The wheels on one side of the robot are linked through timing belts and therefore always turn at the same speed. Each side is driven by two mechanically linked DC motors. The platform offers a build-in computer that hosts the baseline vehicle controller, serial communications, sonar sensors, encoders and other autonomous functions. This built-in controller uses PID control using the motor encoder signals to drive the robot at a commanded speed and calculate the robot's position using dead reckoning. It carries up to 3 hot swappable batteries. The eight forward and eight rear sonar array senses obstacles from 15 cm to 7 m. The robot can reach speeds of 0.8 meters per second and carry a payload of up to 32 kg. The robot uses 100-tick encoders which have been enhanced with inertial correction from a rate gyro measuring yaw movement for dead reckoning to compensate for skid steering.

3.1 Hardware Modifications

Several modifications have been made to the robot to enhance its sensing, computing and fault simulation capabilities required for hosting the PHM and ACM functions. The major additions to the platform described in this section include the following: on-board computer and data acquisition system, batteries, load simulator, tire leakage simulator, vision system, and diagnostic/prognostic server.



Figure 2: The ACM testbed

An onboard computer is mounted on the robot and is powered by the robot's auxiliary power ports. The onboard computer is dual boot, running both Windows XP and LabVIEW RTOS. The onboard computer features a 1.40 GHz Intel Pentium CPU, 512 MB RAM, two 40GB Hitachi HDDs and communicates with the build-in controller through a serial port. It can communicate with other computers (such as a remote client laptop) on a network through a WiFi access point plugged into the computer's Ethernet port.

A NI PCI-6229 data acquisition (DAQ) card has been added to the onboard computer to monitor the health of the robot. The DAQ card has 32 analog inputs, 4 analog outputs and 48 digital I/O. With a few added circuit boards, wires and electrical components, the DAQ card monitors the current and voltage of the battery and motors while sensors are being added to monitor the air pressure of the tires.

To perform prognosis demonstrations using batteries, the 12 V sealed lead acid (SLA) batteries have been replaced with LiFePO₄ Li-Ion batteries. A 200-ohm, 250-watt variable resistor is used to simulate an aging battery, and can be re-wired to simulate a winding short fault in the motor.

To simulate varying loads on the drive system of the robot in the lab, which is often needed for the development and testing of PHM algorithms for batteries and motors, a dynamometer rig has been created as shown in Figure 3. The front two wheels of the robot rest on two rollers. One of the rollers is attached to a hysteresis brake, which can supply a constant braking force when a voltage is applied.

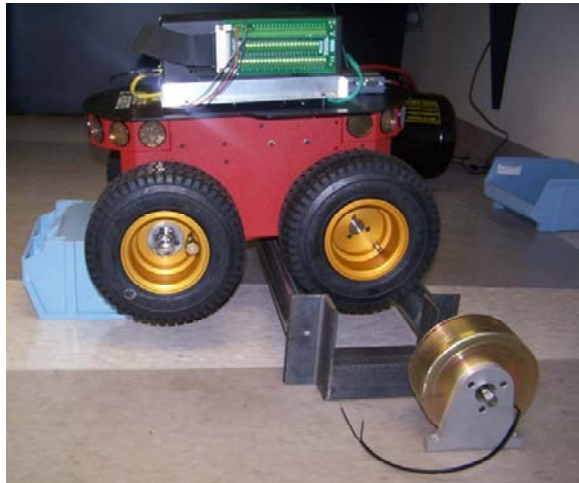


Figure 3: Simulated load test rig setup

A system for simulating tire leakage has been implemented as shown in Figure 4. On the right side of the robot, a rotary union has been connected to the hub of each tire. These rotary unions connect directly to the tire valve outlets, which have had the valves removed. Both of these unions are connected to a central manifold mounted on the top of the robot. This manifold provides ports for a pressure gauge for monitoring tire pressure, a needle valve for simulating slow tire leakage, and a Schrader valve for refilling the tires. Each tire also has its own shut off valve to allow for independent deflation. With this set up, a slow leak in the front, rear or both tires can be simulated.

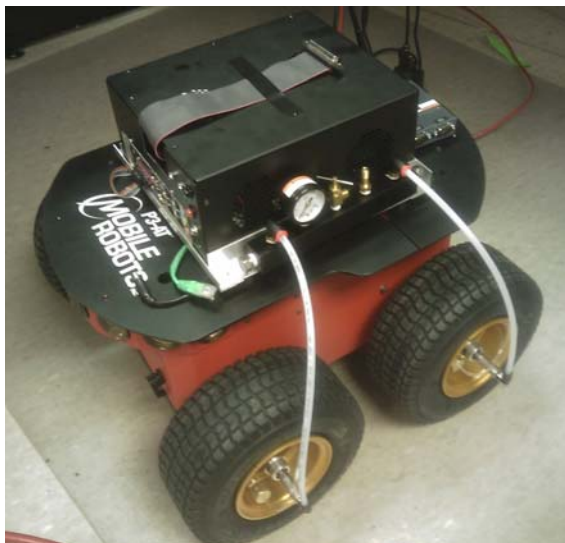


Figure 4: Tire leakage simulation system

To perform image processing, a Surveyor Stereo Vision System mounted on a Lynxmotion BPT-KT pan/tilt head is mounted on top of the onboard computer (see Figure 2). The stereo camera system is

intended for high level applications such as terrain classification, target tracking and classification.

The onboard computer acts as an autonomous server, receiving command signals from a client computer, in this case a laptop computer connected to the same network via WiFi. The laptop is a Dell Latitude D505 with a 1.60GHz Intel Pentium processor, 1.5 GB RAM, and an 80 GB HDD running Windows XP. This computer also acts as the server and image processor for the Stereo Vision System.

The server software on the robot can be configured in several different ways. To perform low-level control functions remotely, the remote client laptop can be configured to send signals that control either the speed of each individual motor, or the overall speed and angular speed of the robot. In this case, low-level sensor signals are sent to the client for processing. As an alternative, the client laptop can be configured to perform only high-level functions by simply sending waypoints to the robot. In this scenario, the onboard computer implements all lower level processing including path planning to the given waypoint, sending velocity signals to the motors, sensor signals processing, and obstacle avoidance and indoor localization using sonar array. This localization takes inputs of a pre-defined map of surroundings and information from the 16 sonar range finders to determine the accurate position of the robot which may be skewed due to the skid steering. The layout of the client software GUI can be seen in Figure 5.

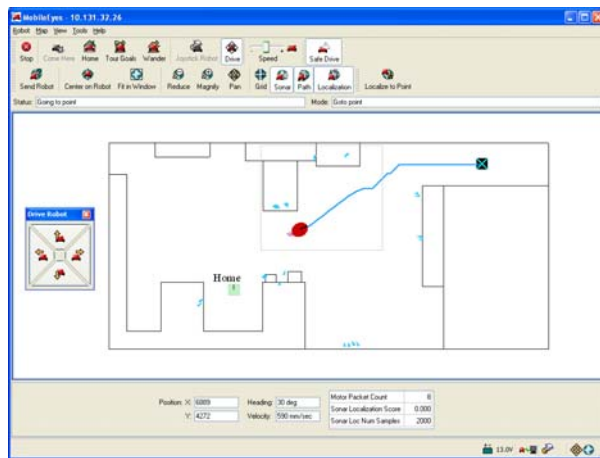


Figure 5: Client software GUI

3.2 FMECA Study

A Failure mode, effects, and criticality analysis (FMECA) study is conducted to identify possible faults and failure modes on the testbed platform. A subset of the failure modes, as well as their criticalities and possible diagnostic approaches and related sensor measurements, is listed below.

Table 1: Representative Failure Modes

Failure Mode	Cr.	Diag.
Motor Assembly		
Motor winding shorted/open	3	A, B, C
Motor shaft eccentricity	2	B, C
Motor bearing spall/wear	2	A, B, C, D
Seized motor bearing/shaft	4	A, B, C
Drivetrain		
Timing belt failure	4	A, B, E
Pulley degradation	2	A, B, E
Leak in tire	2	B, E, F
Tire blowout	4	B, E, F
Battery		
Battery short	3	G, H
Battery degradation	3	G, H
Sensors		
Encoder produces incorrect readings	3	B, E
Encoder produces no signal	3	B, E
Gyroscope produces incorrect readings	3	B, E
Gyroscope produces no signal	3	B, E
Cr.: criticality; Diag.: diagnostic approach and related sensors. A: Motor current; B: Encoder feedback; C: Motor accelerometer; D: Motor spindle; E: Gyroscope feedback; F: tire pressure sensor; G: Battery voltage; H: Battery Current;		

Some of the identified failure modes can be inserted or simulated on the testbed without causing permanent damage to the vehicle. These failure modes include tire leakage, tire blowout, battery short, battery degradation, incorrect encoder reading, no encoder signal, incorrect gyroscope reading, no gyroscope signal, motor wiring short, etc.

3.3 Load Simulations

Since battery end of charge and battery end of life are currently being tracked as part of the prognosis of the robot, it is necessary to simulate different loading scenarios due to terrain changes and other factors while the robot autonomously performs its mission. The robot is currently configured only for indoor use on 2D terrain. In an actual mission, however, the robot would experience different battery loading scenarios based on terrain. To simulate this in an indoor, 2D environment, a variable load has been attached to the battery. This variable load is made up of three resistors, each wired in parallel to the battery. Each resistor can be activated via a relay controlled by the onboard computer. It

provides 8 different loading scenarios progressing linearly in magnitude. The onboard computer has a map of simulated terrain and when the robot crosses into an area of higher simulated difficulty to traverse, the onboard computer activates a larger loading scenario using the variable load. This allows for many simulated terrains while keeping the robot in a safe, indoor environment.

4. IMPLEMENTATION OF PHM-ENABLED ACM SYSTEM ON THE TESTBED

The software architecture of the prototype PHM-enabled ACM system is shown in Figure 6. Starting from the bottom of this hierarchy, the DAQ Server/Monitor collects signals from various components (such as battery, motor, sonar, encoder, gyroscope, etc) and sends the observations to the Diagnosis Reasoner and Prognosis Server. Typically, the prognosis service is only activated after a fault has been detected. The diagnostic and prognostic information are sent to the ACM system where control reconfiguration and mission re-planning take place to accommodate and mitigate both present and potential future faults and failures. The ACM modules send waypoints to the Auto-pilot to adjust the mission to optimize the usage of the vehicle. Set-point commands may also be sent directly to the Vehicle Controller when lower level control reconfiguration is required. The situational awareness sensors, such as the onboard stereo vision cameras and sonars, provide obstacle, target and terrain information to the Situation Awareness module and the Auto-pilot to avoid external threats.

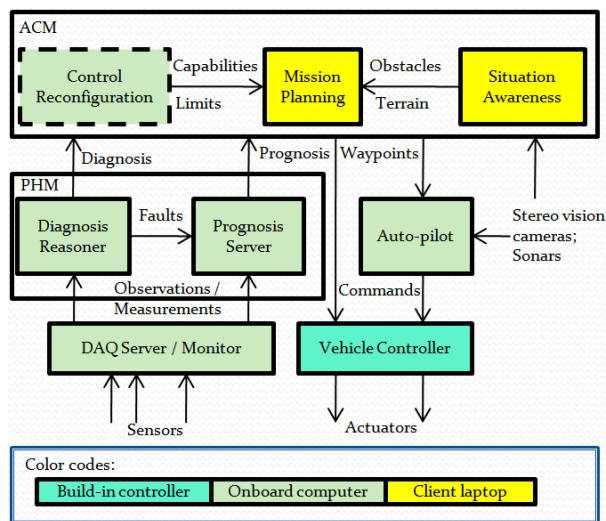


Figure 6: ACM System Software Architecture on the ACM Testbed

4.1 HyDE-Based Diagnosis Reasoner

HyDE (Hybrid Diagnostic Engine) is a model-based diagnosis engine that uses candidate generation and consistency checking to diagnose discrete faults in stochastic hybrid systems. HyDE uses hybrid (combined discrete and continuous) models and sensor data from the system being diagnosed to deduce the evolution of the state of the system over time, including changes in state indicative of faults (Narasimhan and Brownston, 2007).

To demonstrate the generic applicability of HyDE diagnostic reasoning techniques to autonomous vehicles, a Diagnosis Reasoner was developed and deployed on the onboard computer on the robot testbed. The deployed reasoner is essentially a HyDE reasoner that receives sensor observation from the DAQ Server/Monitor module on the onboard computer and outputs diagnostic reasoning result. The Diagnosis Reasoner on the testbed has been developed to diagnose the following fault modes for a proof of concept demonstration.

- 1) Encoder: missing counts; lost – no output
- 2) Timing belt: slipping; failed
- 3) Rate gyro: drifting, lost – no output
- 4) Tires: leaking; deflated
- 5) Tire pressure sensors: biased reading
- 6) Sonar Sensors: erroneous reading

A part of the HyDE diagnostic model involving the encoders, timing belt, rate gyro, tire pressure sensors and the tires is shown in Figure 7.

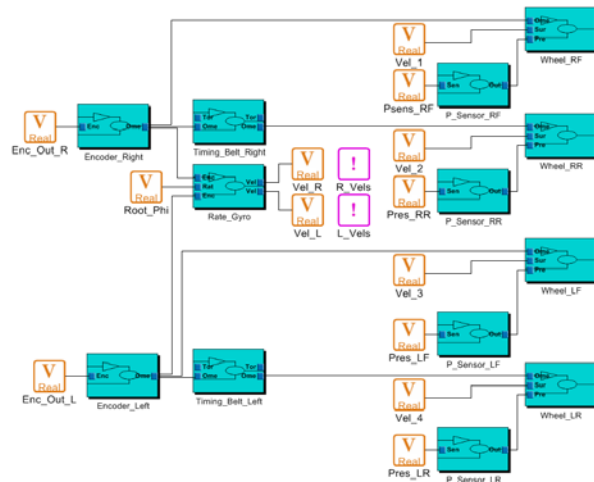


Figure 7: HyDE model for a few selected components on the testbed

To illustrate the reasoning capability of the HyDE-based solution as compared to simple logical calculation (which is often utilized in Expert System-based diagnostic systems), hybrid state reasoning has been used to determine the state of the tires. Using encoder measurements in addition to abnormal rotation

measured by the rate gyro during commanded forward movement, the diagnostic model can determine which side of the robot has a low pressure tire. With this information and the calculated logical constraints, the model can then determine whether the front or the rear tire is causing the reduced velocity. When combining this reasoning capability with tire pressure sensor reading, a pressure sensor fault can be diagnosed.

To test the model, real-time data was collected from the robot while varying both rate of movement in forward and reverse as well as individual tire pressures. Using this data, it was shown that a calculated channel could indicate an individual tire fault using accumulated significant error from the rate gyro. With the calculated channel input to the diagnostic model, it is possible to isolate an individual tire fault. The diagnosis result in a GUI is shown in Figure 8.

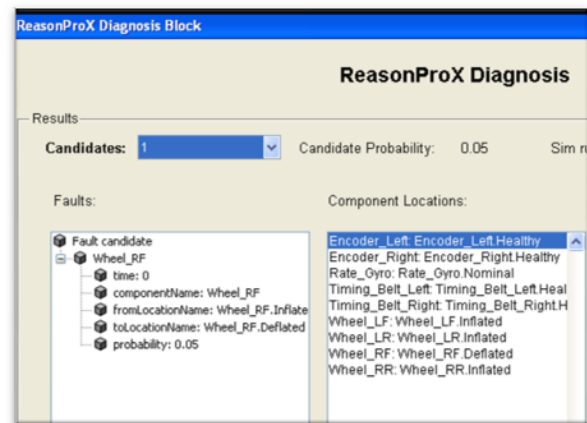


Figure 8: HyDE diagnosis given deflated right-front wheel data

4.2 Particle Filtering-Based Prognosis Server

The purpose of prognosis is to predict the remaining useful life (RUL) of a system/subsystem or a component when a fault is detected. Various prognostic algorithms have been developed and applied to various mechanical and electrical systems in the past decade (Schwabacher and Goebel, 2007; Uckun et al, 2008; Saxena et al, 2010). Among these approaches, particle filtering-based approaches have been shown to be theoretically sound, generically applicable and demonstrating promising results especially on applications where online prognosis is required (Goebel et al, 2008; Saha et al, 2009). To illustrate the effectiveness and computational efficiency of real-time PF-based prognosis approach, three parallel PF-based prognostic routines have been implemented on the Prognosis Server to predict three failure modes: the RUL of the battery, state of charge of the battery and tire leakage situation. Details regarding the particle filtering-based prognosis and uncertainty management

algorithms can be found in (Orchard et al, 2010; Edwards et al, 2010).

4.2.1 Battery Life Prognosis

To predict the battery end-of-life (EOL), features that reflect the aging condition of the battery such as capacity or stored energy must be tracked over time. Other features derived from electrochemical impedance spectrometry (EIS) measurement data (Goebel et al, 2008) may also be used but these features can only be obtained when onboard EIS devices are available. In this case study, the capacity of battery is used as the feature.

Because of its generic applicability and rich uncertainty management capabilities, a particle filtering-based approach is chosen. To date, three particle filtering-based prognostic approaches have been developed: i) the classic particle filtering algorithm (Orchard, 2009), ii) a routine that incorporated an Outer Feedback Correction Loop (Orchard et al., 2008), and iii) Risk Sensitive Particle Filter (RSPF) based routines (Orchard et al., 2010). The algorithm deployed on our testbed is the RSPF-based approach. Prognosis results obtained when applying the algorithm to a set of real Lithium-Ion battery data are shown in Figure 9. The capacity data measured per cycle is plotted in green while the estimated capacity which is the feature being tracked is plotted in magenta in the upper subplot. The critical capacity limit is centered around 1200 mAh as shown by the orange zone. The lower subplot shows the scaled probability density function (PDF) of the EOL predicted at cycle 100 (the peak of the PDF has been scaled to 1 for plotting purposes). In this case, the ground truth life of the battery is 168 cycles which is very close to the mean of the predicted PDF.

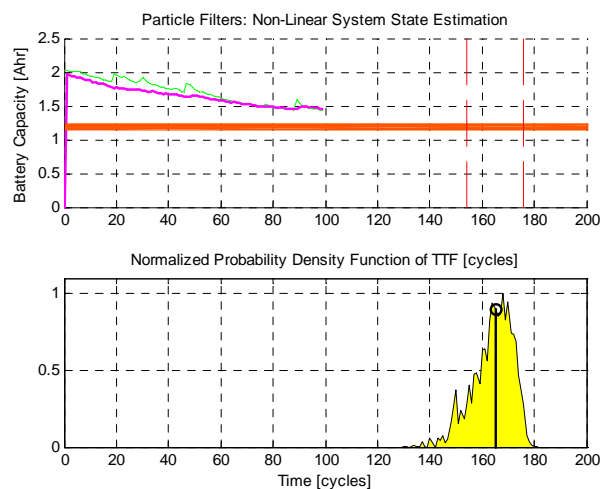


Figure 9: Battery life prognosis using RSPF-based approach

The algorithm and the implemented software module have also been tested on the battery data set provided by the Prognostics Center of Excellence at NASA Ames research center (Saha and Goebel, 2007) with comparable results. It should be noted that since the capacity data is collected per charge-discharge cycle, the software does not need to run in real-time.

4.2.2 Battery End-of-Charge Prediction

In contrast to battery end-of-life prognosis which is important for the planning of long term or future missions, battery end-of-charge prognosis focuses on the prediction of battery charge state for the current mission given the health of the battery which has degraded over the course of use. When a vehicle is powered by batteries, its mission plan can be optimized in real-time if an accurate battery end-of-charge prediction capability is available on board the vehicle.

A particle-filtering based algorithm that uses a combined voltage and stored energy feature has been implemented in the Prognosis Server on the onboard computer. One of the challenges in predicting battery end-of-charge is to handle the uncertainties associated with the prognosis due to uncertain initial state of charge, ambient temperature, future load (discharge) profile and battery health, among other factors. Therefore, a particle filtering-based algorithm is chosen in this case study due to its unique uncertainty management capability and computational efficiency which enables real-time execution of highly accurate predictions.

Figure 10 below shows a set of battery voltage and current data collected on the testbed running a random varying load and the prognosis results. It is clear that the voltage signal (first subplot) is mainly affected by the load before the battery charge reaches a critical level at about 3500 seconds, then the voltage drops drastically to 10 V within 50 seconds. The current (as shown in second subplot) is totally determined by the load and only a trivial increase tending is observed towards the end when voltage drops. In the 3rd subplot, three end-of-charge predictions (PDFs) made at 1226, 2451 and 3677 seconds respectively are shown. Since we know the true end-of-charge time in this case (3990 seconds), it can be seen that the first and second predictions have been made conservatively due to the uncertainties associated with initial charge state and future load. The third prediction which was made when the voltage signal started to drop was rather accurate.

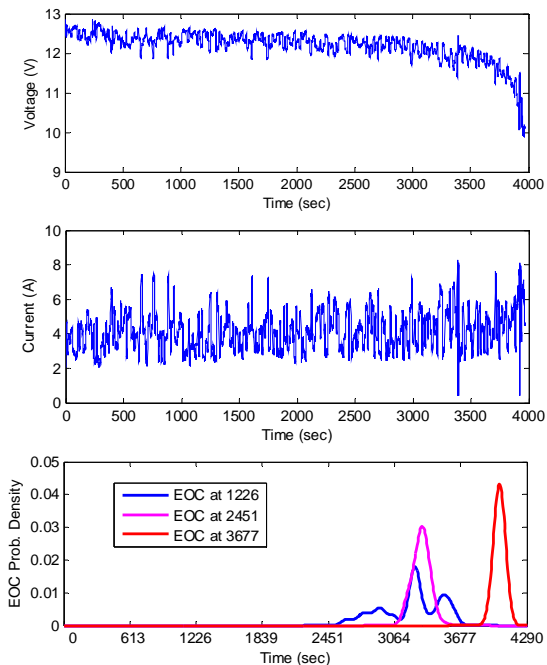


Figure 10: Battery data and end-of-charge prediction

4.3 Prognostics-Enhanced Mission Planning

A Mission Planning module which uses prognostic information to enhance the mission/path planning in a non-uniform environment has been implemented and deployed on the Client Laptop as shown in Figure 6. Prognostic information is introduced in order to ensure that the fault progression, or mission failure risk, can be minimized after the occurrence of a fault. This will enhance the performance of autonomous vehicles that often work in harsh environments that cause aging, fatigue, and fracture.

When a fault occurs, the mission planning module sends periodic RUL service requests to the Prognosis Server (Figure 6). The received RUL estimate (typically represented as a probability density function) is then used either as a constraint or an additional element in the cost function in the path planning algorithm, in this case, a field D^* search algorithm, in a receding horizon planning framework (Zhang et al, 2011).

In the field D^* algorithm, the map is described by grids while nodes are defined on corners of grids. The planning algorithm divides the map into three areas: implementation area, observation area, and unknown area as shown in Figure 11. The autonomous vehicle is equipped with onboard sensors that are able to detect and determine the terrain in the observation area (the area inside the magenta square in Figure 11). The implementation area (the area inside the green square in Figure 11) consists of the grid next to the current node.

The area beyond observation area is the unknown (un-observed) area where the terrain is unknown to the vehicle (the gray area in Figure 11). At a node, the vehicle plans the path from the vehicle's current location to the destination. However, only the path planned in the implementation area is executed. This is similar to the strategies used in Receding Horizon (Model Predictive) Control algorithms. This process is repeated until the destination is reached or it turns out that no route can lead to destination or the vehicle reaches its end of life. The cost function is the weighted sum of three factors and is defined as:

$$\min_{s'} J = w_T (t_o(s, s') + t_u(s')) + w_{Tr} (d_o(s, s') + d_u(s')) + w_{Pr} (p_o(s, s') + p_u(s')) \quad (1)$$

where $t_o(s, s')$, $d_o(s, s')$, and $p_o(s, s')$ are normalized travel time, terrain, and fault costs on the path segment from point s to point s' , respectively, while $t_u(s')$, $d_u(s')$, and $p_u(s')$ are those corresponding costs on path segment from point s' to final destination, respectively; w_T , w_{Tr} , and w_{Pr} are the weighting factors on each cost factor and $w_T + w_{Tr} + w_{Pr} = 1$.

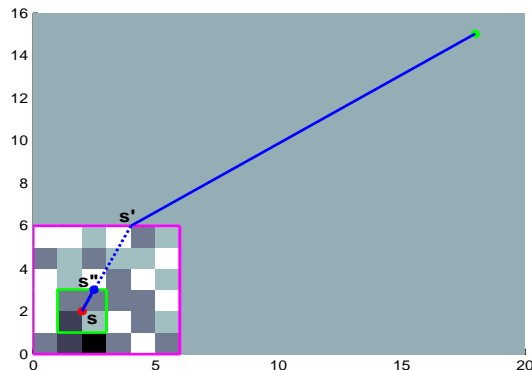


Figure 11: Receding horizon mission planning

To investigate how real-time prognostic information can be utilized in mission planning on the ACM testbed, we consider battery end-of-charge during the mission as a fault mode. In this case, the real-time battery prognosis routine described in section 4.2.2 is utilized to predict the state-of-charge of the battery. Note that battery remaining charge is a function of terrain difficulty and vehicle speed. Several laboratory experiments were conducted that illuminate the impact of battery prognostics on mission planning. Representative experimental results for four different optimality criteria are presented in Table 2. For example, it is apparent that when the mission plan is optimized for battery life (3rd column), the robot travels the longest distance and finishes the mission with the longest time, but consumes the least amount of battery life. In contrast, with the time optimal mission plan (1st column), the robot travels the shortest distance at a

higher speed, but consumes the most battery life. Both the mission planner and prognosis routine are implemented in real-time in the experiments.

Table 2: Representative Mission Planning Experiment Results for Four Different Optimality Criteria

	Optimality Criteria			
	Time	Terrain	Life	Weighted
$[w_T, w_{T_r}, w_{P_r}]$	[1,0,0]	[0,1,0]	[0,0,1]	[0.33,0.33,0.34]
Travelled distance (meters)	21.26	24.56	24.81	24.45
Travel time (seconds)	114.2	202.2	125.9	126.4
Remaining battery charge (%)	67.9	82.3	91	88.1

5. CONCLUSION

Real-time onboard PHM and ACM systems are needed to improve the reliability and survivability of autonomous vehicles engaged in critical missions. Research and development of enabling techniques have been conducted in recent years to achieve the required capabilities using relevant simulation programs with various levels of fidelity. This paper presents the development of a testbed that is built for the purpose of evaluating real-time vehicle PHM and ACM techniques on a real robotic platform. The testbed has been utilized to demonstrate the feasibility of a hierarchical ACM system that we have been developing over the past years highlighting the importance of using PHM information in control reconfiguration and mission planning. Several key software modules featuring real-time system level diagnosis, component fault prognosis and prognostics-enhanced mission planning have been successfully demonstrated on the testbed. Work continues on further development of the PHM-enabled control reconfiguration techniques and will eventually implement them on the testbed. More advanced situational awareness capabilities such as target tracking, localization and terrain classification will be developed so that the ACM techniques can be evaluated in more real world application scenarios. More importantly, efforts will be made to ensure the general applicability of the developed techniques to other types of autonomous vehicles such as UAVs and space exploration vehicles.

ACKNOWLEDGMENT

This work is partly supported by the NASA under SBIR contract NNX09CB61C. The authors gratefully

acknowledge the productive collaboration with Dr. Kai Goebel, Mr. Edward Balaban, Dr. Bhaskar Saha, Dr. Abhinav Saxena, Dr. Jose Celaya and the Prognostics Center of Excellence (PCoE) at Ames Research Center. The authors also thank Dr. George Vachtsevanos, Mr. Gregory Kacprzyński, and Mr. Brooks Przybyłek from Impact Technologies, Dr. Marcos Orchard from Universidad de Chile, and Mr. Brian Bole from Georgia Institute of Technology for their contributions on the development of the testbed and related techniques.

REFERENCES

- Army UAS Center of Excellence Fort Rucker AL (2010), *U.S. Army Roadmap for Unmanned Aircraft Systems 2010-2035: Eyes of the Army*, <http://www-rucker.army.mil/usaace/uas/US%20Army%20UAS%20RoadMap%202010%202035.pdf>
- DeCastro, J. A., Tang, L., Zhang, B. and Vachtsevanos, G. (2011), A Safety Verification Approach to Fault-Tolerant Aircraft Supervisory Control, in proceedings of *the AIAA Guidance, Navigation and Control Conference and Exhibit*.
- Edwards, D., Orchard, M., Tang, L., Goebel, K., and Vachtsevanos, G. (2010), Impact of Input Uncertainty on Failure Prognostic Algorithms: Extending the Remaining Useful Life of Nonlinear Systems, in proceedings of *the Annual Conference of the Prognostics and Health Management Society*, Portland, OR.
- Goebel, K., Saha, B., Saxena, A., Celaya, J., and Christophersen, J. (2008), Prognostics in Battery Health Management, *Instrumentation & Measurement Magazine, IEEE*, vol.11, no.4, pp.33-40.
- Navy's Program Executive Officer Littoral and Mine Warfare, Surface Warfare and Unmanned Maritime Vehicles Program Office (2007), *The Navy Unmanned Surface Vehicle (USV) Master Plan*, <http://www.navy.mil/navydata/technology/usvmprr.pdf>
- Narasimhan, S. and Brownston, L.(2007), HyDE – A General Framework for Stochastic and Hybrid Model-based Diagnosis, in proceedings of *the 18th International Workshop on Principles of Diagnosis (DX-07)*.
- Orchard, M., Kacprzyński, G., Goebel, K., Saha, B., and Vachtsevanos, G., (2008). Advances in Uncertainty Representation and Management for Particle Filtering Applied to Prognostics, in proceedings of *International Conference on Prognostics and Health Management*.
- Orchard, M., Tang, L., Saha, B., Goebel, K., and Vachtsevanos, G., (2010) Risk-Sensitive Particle-

- Filtering-based Prognosis Framework for Estimation of Remaining Useful Life in Energy Storage Devices, *Studies in Informatics and Control*, vol. 19, Issue 3, pp. 209-218.
- Orchard, M. (2009). *On-line Fault Diagnosis and Failure Prognosis Using Particle Filters. Theoretical Framework and Case Studies*, Publisher: VDM Verlag Dr. Müller Aktiengesellschaft & Co. KG, Saarbrücken, Germany.
- Saha, B., and Goebel, K. (2007). "Battery Data Set", NASA Ames Prognostics Data Repository, [<http://ti.arc.nasa.gov/project/prognostic-data-repository>], NASA Ames, Moffett Field, CA.
- Saha, B., Goebel, K., Poll, S. and Christophersen, J. (2009), Prognostics Methods for Battery Health Monitoring Using a Bayesian Framework, *IEEE Transactions on Instrumentation and Measurement*, vol.58, no.2, pp.291-296.
- Saxena, A., Celaya, J., Saha, B., Saha, S., and Goebel, K. (2010), Metrics for Offline Evaluation of Prognostics Performance, *International Journal of Prognostics and Health Management*, Vol.1(1), pp. 20.
- Schwabacher, M. and Goebel, K. (2007), A Survey of Artificial Intelligence for Prognostics, in *Proceedings of AAAI Fall Symposium*, Arlington, VA.
- Tang, L., Kacprzynski, G. J., Goebel, K., Saxena, A., Saha, B., and Vachtsevanos, G. (2008). Prognostics-enhanced Automated Contingency Management for Advanced Autonomous Systems, in *Proceedings of the International Conference on Prognostics and Health Management*.
- Tang, L., Kacprzynski, G., J., Goebel, K., and Vachtsevanos, G. (2010). Case Studies for Prognostics-Enhanced Automated Contingency Management for Aircraft Systems, in *Proceedings of IEEE Aerospace Conference*.
- Uckun, S., Goebel, K., and Lucas, P. J. F. (2008), Standardizing Research Methods for Prognostics, in proceeding of *the 1st International Conference on Prognostics and Health Management (PHM08)*, Denver CO.
- Vachtsevanos, G., Lewis, F. L., Roemer, M., Hess, A. & Wu, B. (2006). *Intelligent Fault Diagnosis and Prognosis for Engineering Systems*, 1st ed. Hoboken, New Jersey: John Wiley & Sons, Inc.,
- Zhang, B., Tang, L., Decastro, J. A., and Goebel, K. (2011), Prognostics-enhanced Receding Horizon Mission Planning for Field Unmanned Vehicles, in proceedings of *the AIAA Guidance, Navigation and Control Conference and Exhibit*.
- Liang Tang** is an Intelligent Control and Prediction Lead at Impact Technologies. Dr. Tang's career has been focused on the development of state estimation, health management and intelligent control solutions for a wide range of military and commercial systems. His recent work has also involved developing intelligent control, data fusion and navigation technologies for various unmanned autonomous platforms. He is currently the Principal Investigator of several NASA and DoD SBIR/STTR programs developing fault tolerant control strategies for autonomous vehicles, GPS-independent inertial navigation device, diagnostic algorithms for jet engine and distributed data fusion system for USVs. He received his Ph.D. degree from Shanghai Jiao Tong University, China, in 1999. Before he joined Impact Technologies in 2003, he was a postdoctoral research fellow with the School of ECE, Georgia Institute of Technology, where he conducted research on fault diagnosis and controls of UAVs. Dr. Tang has published more than 50 papers in his areas of expertise.
- Eric Hettler** is a Project Engineer at Impact Technologies, LLC. He earned BS and MEng degrees in Mechanical Engineering from the Rochester Institute of Technology concentrating in systems modeling and control systems. Under NASA funding, his recent work has focused on developing hardware and software code in support of the mobile robotic testbed for autonomous contingency management.
- Bin Zhang** is a Senior Project Engineer at Impact Technologies with over 10 years experience in intelligent systems, modeling and simulation, fault diagnosis, prognostics and health management. He received his Ph.D. degree from Nanyang Technological University, Singapore, where his research focused on intelligent learning control, system prototyping, and applications to power electronics and robotics. Before he joined Impact Technologies, he was a research fellow with the School of ECE, Georgia Institute of Technology, Atlanta GA, where he participated in some SBIR/STTR projects and developed novel signal processing algorithms, fault progression models, and fault diagnosis and failure prognosis algorithms. Dr. Zhang is a senior member of IEEE. He serves as an Associate Editor for IEEE Transactions on Industrial Electronics and International Journal of Fuzzy Logic and Intelligent Systems. He also serves in program committee for some international conferences. He has published more than 70 papers in his areas of expertise.
- Jonathan DeCastro** is Lead Engineer at Impact Technologies, LLC with over ten years of experience developing and maturing advanced control techniques,

high-fidelity control evaluation simulations, and control allocation/re-allocation for fault-tolerant and safety-critical aircraft propulsion systems. As a research scientist at the NASA Glenn Research Center, Mr. DeCastro had developed the commercial modular aero-propulsion system simulation (C-MAPSS) aircraft propulsion model and advanced control system, which has since become a NASA software product. His previous research accomplishments have centered on the development of development of advanced controls

technologies for aircraft, including receding-horizon control, fault-tolerant control, and distributed control. Mr. DeCastro earned the BS and MS degrees in Mechanical Engineering from Virginia Tech. He is author of over 20 publications and is the recipient of an AIAA Section Award for Best Young Professional Paper, a NASA Group Achievement Award for leadership in developing C-MAPSS, and two NASA Space Act Awards.

Adaptive Load-Allocation for Prognosis-Based Risk Management

Brian Bole¹, Liang Tang², Kai Goebel³, George Vachtsevanos¹

¹ *Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA.*

*bbole3@gatech.edu
gfv@ece.gatech.edu*

² *Impact Technologies, LLC, Rochester, NY 14623, USA.*

Liang.Tang@impact-tek.com

³ *NASA Ames Research Center, Moffett Field, CA 94035, USA.*

kai.goebel@nasa.gov

ABSTRACT

It is an inescapable truth that no matter how well a system is designed it will degrade, and if degrading parts are not repaired or replaced the system will fail. Avoiding the expense and safety risks associated with system failures is certainly a top priority in many systems; however, there is also a strong motivation not to be overly cautious in the design and maintenance of systems, due to the expense of maintenance and the undesirable sacrifices in performance and cost effectiveness incurred when systems are over designed for safety. This paper describes an analytical process that starts with the derivation of an expression to evaluate the desirability of future control outcomes, and eventually produces control routines that use uncertain prognostic information to optimize derived risk metrics. A case study on the design of fault-adaptive control for a skid-steered robot will illustrate some of the fundamental challenges of prognostics-based control design.

1. INTRODUCTION

Some form of risk management can be seen in virtually every decision that human beings make. Typically, the desirability of future outcomes can be objectively evaluated; however, evaluating the best present control decision is complicated by uncertainty in estimating the future effects of control actions. In the case of controlling a system with incipient faults, the design objective is to obtain a system with high performance, low maintenance cost, and low failure rates. The effects of decisions regarding the design, maintenance, and operation

of a system on its future performance, maintenance cost, and failure rates are commonly estimated by using empirical data or expert knowledge to assess probable outcomes.

The fault analysis process typically starts with the identification of potential failure modes and the quantification of the severity and likelihood of each, based on expert knowledge and historical data. The Failure Modes, Effects, and Critically Analysis (FMECA) is one of the most widely applied *a priori* fault analysis methods; it is currently nearly universally applied in industrial automation (Gergely, Spoiala, Spoiala, Silaghi, & Nagy, 2008), automotive (SAE, 1994), and aerospace (Saglimbene, 2009) industries. Fault Tree Analysis (FTA), Event Tree Analysis (ETA), Reliability Block Diagrams (RBD), and other fault analysis techniques that utilize historical failure rates will continue to play an ever more prominent role in the design of hardware specifications and contingency management policies.

In addition to the established practice of utilizing historical fault data to manage failure risks, there is also a growing push to develop technologies for online fault identification and fault growth prediction to improve system operation and maintenance. Online anomaly detection and diagnostic routines are enabling an increased use of condition based maintenance and control (CBMC) policies (Rao, 1998). Pseudo-inverse (Caglayan, Allen, & Wehmuller, 1988), model predictive control (MPC) (Monaco, Ward, & Bateman, 2004), and H_2 and H_∞ robust control theory (Doyle, Glover, Khar-gonekar, & Francis, 1987) are commonly used methods to recover controllability of a system after a known fault mode is detected. Further improvements in performance and safety are expected if the diagnostic information used by CBMC routines is supplemented with prognostic routines that predict the growth fault modes as a function of future use; however, the development and use of prognostic information is typi-

This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

cally an extremely challenging proposition due to significant uncertainty in predicting future fault growth. Prudent methods for modeling fault diagnostic and prognostic uncertainty should be selected on a case-by-case basis; particle Filtering and Bayesian Reasoning are commonly used for estimating fault magnitudes and predicting future growth based on uncertain measurements and physical modeling (Arulampalam, Maskell, Gordon, & Clapp, 2002; Orchard, Kacprzynski, Goebel, Saha, & Vachtsevanos, 2008; Saha & Goebel, 2008; Sheppard, Butcher, Kaufman, & MacDougall, 2006).

The analytical approach to fault-adaptive control design that is introduced in this paper will assume that a non-empty space of current control actions to maintain system stability is known, and the controller must attempt to select control actions from that space to best manage the risk posed by degrading components. The future effects of control actions will be represented by generic probability distributions and control actions for best risk management will be derived by attempting to optimize an objective function that quantifies the relative aversion to the risk posed by further degrading components and the risk of degrading future system performance. Candidate metrics for evaluating risk from uncertain prognostic estimates may be drawn from the growing body of publications on vehicle health management (IVHM) (Srivastava, Mah, & Meyer, 2008); although, nearly all current studies in this area consider only end of life predictions in risk calculations and ignore data regarding short term fault growth, which will not be ideal in many cases. Literature on risk management in finance and actuarial science contain a rich array of tools that facilitate flexible risk-reward analysis on a continuous scale over a finite horizon. For example, Black-Scholes models (Lauterbach & Schulz, 1990) and value at risk (VaR) (Venkataraman, 1997) are prolific financial risk management tools that are also promising candidates for analyzing prognostic predictions (Schreiner, Balzer, & Precht, 2010).

This paper will explore the fundamental principles behind the derivation, verification, and validation of controls for optimal risk management on systems with incipient faults that grow in severity with increased component loading. The utility of various VaR based risk metrics for evaluating risk over a prognostic horizon, will be explored in a case-study on the use of prognostics-based load-allocation control for an unmanned ground vehicle (UGV).

2. PROGNOSTICS FOR RISK MANAGEMENT

The risk analysis process should begin with the definition of an analytical expression to evaluate the desirability of future control outcomes. In practice some form of scenario analysis should be used to derive and validate evaluation metrics through empirical studies (Abhken, 2000). Evaluation functions for future control outcomes represent the relative value of preserving nominal system performance and minimizing component degradations or failures for given scenarios.

A general form of an outcome evaluation function is

$$J_M(x(t)) + J_d(d_i^T), \{t = t_0..T\}, \{i = 1, 2, ..N\} \quad (1)$$

where $x(t)$ represents the system state at time t , d_i^T represents the amount that component i has been degraded at the end of the mission, $J_M(x(t))$ evaluates how well the system conformed to mission specifications and mission priorities over a mission that starts at $t = t_0$ and ends at $t = T$, and $J_d(d_i^T)$ evaluates the cost associated with the final state of degradation for each of the N components. The problem of specifying control actions to maximize this evaluation function will be referred to as the intrinsic optimization problem.

Due to uncertainty in the way faults grow with component loading and uncertainty regarding external operating conditions, it is generally impossible to design a controller that solves the intrinsic optimization problem directly; however, any control technique that claims to manage or mitigate the risk posed by load dependent fault modes can be viewed as being implicitly derived based on the optimization of an intrinsic cost function. The development of analytical tools that utilize knowledge of the intrinsic cost function in the design of prognostics-based fault-adaptive controllers will facilitate an understanding of the benefits of proposed approaches, as well as their fundamental limitations.

3. COMPONENT LOAD-ALLOCATION

In a broad variety of systems the performance of the system and the growth of potential faults can be viewed as being direct functions of component loads. In general, the fault adaptive control problem can be fully understood in terms of a search for optimal performance and risk metrics that are evaluated on the space of allowable component loads over a given prognostic horizon. The space of allowable component load-allocations over a given prognostic horizon and the methods used to derive that space, will vary from one application to the next; however, many aspects of the fundamental searching problem will be invariant across a range of applications, facilitating the development of widely applicable analysis and control techniques.

The domain of allowable component load allocations over a given prognostic horizon will be defined at each control time-step using available system modeling and prognostic information to translate tolerances on performance degradations and fault growth risks into the component control domain. If the system is overactuated then the search for optimal component load allocations can be decomposed into two reduced-order sub-problems. An output control effort optimization will search for the optimal net system output control effort, and a restricted component load optimization will utilize any inherent overactuation in the system to find load allocations that minimize component degradations while providing the system output force requested by the output control effort optimization routine.

The separation of the component loading and system output regulation tasks will be shown for a generic nonlinear system,

$$\dot{x} = A(x) + B(x) \mathbf{u} \quad (2)$$

where $A(x) \in \mathbb{R}^n$, $B(x) \in \mathbb{R}^{n \times m}$, $\mathbf{x}(t) \in \mathbb{R}^n$, is the state, and $\mathbf{u}(t) \in \mathbb{R}^m$ is the control effort or load on each of the m components in the system. If $B(x)$ does not have full column rank, i.e., $\text{rank}\{B(x)\} = k < m \forall x$, then the system is overactuated, and $B(x)$ can be factorized as:

$$B(x) = B_\nu(x) B_u(x) \quad (3)$$

where $B_\nu(x) \in \mathbb{R}^{n \times k}$ and $B_u(x) \in \mathbb{R}^{k \times m}$ both have rank k . Now the system can be rewritten as:

$$\begin{aligned} \dot{x} &= A(x) + B_\nu(x) \boldsymbol{\nu} \\ \boldsymbol{\nu} &= B_u(x) \mathbf{u} \end{aligned} \quad (4)$$

where $\boldsymbol{\nu}(t) \in \mathbb{R}^k$ can be interpreted as the net control effort produced by the m system components.

Because $B_\nu(t)$ has full column rank, a desired system output will uniquely determine the net output control effort, $\boldsymbol{\nu}(t)$ (using the pseudo inverse); however, since $B_u(x)$ has a nullspace of dimension $m - k$ there are available degrees of freedom in assigning component loads, $\mathbf{u}(t)$, for a given $\boldsymbol{\nu}$. Then component loads for best risk management can effectively be expressed as a function of $\boldsymbol{\nu}$, where any inherent redundancies in actuation, identified by the null space of B_u , are used to minimize component damages while still resulting in the net control effort commanded.

Practical applications of control allocation are currently found in aerospace (Gokdere, Bogdano, Chiu, Keller, & Vian, 2006; Karpenko & Sepehri, 2005) and automotive vehicles (Hattori, Koibuchi, & Yokoyama., 2002). A survey of efficient methods for determining the optimal control allocation for general linear and nonlinear systems is discussed in (Oppenheimer, Doman, & Bolender, 2006). Proof of the equivalence of this type of control allocation and optimal control is given in (Harkegard & Glad, 2005), for nonlinear systems with quadratic cost functions.

3.1 Load-Allocation as a Bounded Optimization

The objective of the general fault-adaptive control problem is to select the current component load allocations in an attempt to optimize the system's intrinsic cost function. In this work, component loads are allocated at the current control time-step by attempting to optimize an objective function that uses system modeling and fault prognostic information to quantify the expected trade-off between system performance and fault risk over a specified prognostic horizon. Constraints on allowable system performance and fault growth risk over a prognostic horizon will be enforced in the domain of allowable component load allocations in an attempt to satisfy minimum remaining-useful-life requirements for failing components.

The analysis presented in this document will use a fault risk metric of the following generic form:

$$f(\tilde{d}_i(t + \tau)), \Pr(d_i(t + \tau) > \tilde{d}_i(t + \tau)) = \alpha, \quad \{i = 1, 2, \dots, N\}, \text{ given } \{\mathbf{u}_i(t) \dots \mathbf{u}_i(t + \tau)\} \quad (5)$$

where τ is the length of the prognostic horizon, $d_i(t)$ is the estimated degradation of component i at time t , $\tilde{d}_i(t + \tau)$ is a VaR estimate for component damage at the prognostic horizon, and $f(\tilde{d}_i(t + \tau))$ represents a risk metric that penalizes VaR estimates. VaR estimates are defined as the threshold damage such that the probability of the actual damage exceeding a given magnitude at a given future time equals α . Published literature contains relatively few examples of VaR being employed to manage the risk posed by incipient fault models; however, VaR is a standard risk assessment tool in finance, and it is powerful and widely applicable tool for risk management in systems with degrading components (Schreiner, Balzer, & Precht, 2008; Schreiner et al., 2010; Venkataraman, 1997).

A general form of the cost function used to represent the relative aversion to the risk of degrading future system performance and the risk posed by degrading components is

$$g(|\boldsymbol{\nu} - \mathbf{r}|) |_{t^{+\tau}} + f(\tilde{d}_i(t + \tau)) \quad (6)$$

where \mathbf{r} represents the desired net output control effort required for nominal performance and $g(|\boldsymbol{\nu} - \mathbf{r}|)$ penalizes performance degradation over a given prognostic horizon.

In published literature on prognostics for risk management there is a nearly ubiquitous use of expected remaining useful life (RUL) or expected time to failure (TTF) estimates to assess risk; however, in general, the methodology used to assess risk from fault prognosis information should be tailored to the system's expected use, its maintenance costs, the danger of potential failure modes, and the growth of uncertainty over a prognostic horizon. In this paper, the length of the prognostic horizon and the utility of various metrics for quantifying risk from prognostic predictions will be explored as a design choice. In cases where RUL or TTF based risk metrics are deemed most appropriate they can be realized as a special case of finite horizon prognosis, in which the prognostic horizon is extended until component failure is assured.

Constraints on allowable system performance are defined either in terms of a maximum deviation from commanded system states or a maximum deviation from the desired nominal system output force at a given time,

$$|\mathbf{y}_c - \mathbf{y}_o| \leq \Delta(t) \quad (7)$$

$$|\boldsymbol{\nu} - \mathbf{r}| \leq \tilde{\Delta}(t) \quad (8)$$

A finite horizon prognosis constraint will place an upper-bound on the probability that a component will become damaged by more than a specified amount over the prognostic horizon. This constraint is written as follows:

$$\Pr(d_i(t + \tau) > \gamma_i(t + \tau) | \mathbf{u}_i(t)) \leq \beta \quad (9)$$

where $\gamma_i(t + \tau)$ is the maximum allowable fault dimension at time $t + \tau$ and β is the upper bound on the probability that the fault dimension of component i is larger than its maximum allowable value at time $t + \tau$.

3.2 Verifying Constraint Feasibility

If the future performance requirements are known in advance, then the existence of feasible solutions to the optimal component load allocation problem can be verified by first finding the minimum allowable net output control effort needed to satisfy the performance constraints;

$$\tilde{\nu} = \min_{\mathbf{u}_i} \{\nu\}, \quad \text{s.t. } |\nu - \mathbf{r}| = \tilde{\Delta}(t), \quad \forall t \quad (10)$$

where $\tilde{\nu}$ is the minimum allowable net output control effort under the performance constraint. Feasible solutions to the optimal load-allocation problem exist if there exists a distribution of component loads that result in $\tilde{\nu}$ and do not violate the prognostic constraint at the end of the mission. This condition is written as follows:

$$\Pr(d_i(T) > \gamma_i(T) | \mathbf{u}_i(t)) \leq \beta, \\ \text{s.t. } \tilde{\nu} = B_u(x(t)) \mathbf{u}, \quad t \in [t_0, T] \quad (11)$$

4. UGV APPLICATION EXAMPLE

Simulation studies for optimal load-allocation on a skid-steered UGV will demonstrate some of the fundamental properties of the proposed control methods. As shown in Figure 1, each of the wheels in a skid-steered vehicle are fixed to the frame and are pointing straight forward. The system is over-actuated, as the four motors of the four-wheeled UGV are linked through their mutual contact with the ground. Assuming that all of the robot's wheels are getting approximately the same traction, then a skid-steered wheeled vehicle will behave much like a treaded vehicle. In the presented simulation studies the UGV's modeling is simplified by treating it as a treaded vehicle. The net output control effort of the modeled UGV is defined as follows:

$$\nu = \begin{bmatrix} \nu_f \\ \nu_\phi \end{bmatrix} = \begin{bmatrix} T_1 + T_2 + T_3 + T_4 \\ T_1 + T_2 - T_3 - T_4 \end{bmatrix} = \begin{bmatrix} T_L + T_R \\ T_L - T_R \end{bmatrix} \quad (12)$$

where ν_f represents the net motive torque applied in the direction of travel, ν_ϕ represents the net turning torque, T_L is the sum of the motor torques on the left side of the robot, and T_R is the sum of the motor torques on the right side of the robot.

The UGV model is

$$M\dot{x} = -C(x) + B \cdot u \\ y = \begin{bmatrix} \frac{r}{\alpha W} & \frac{r}{\alpha W} \\ -\frac{r}{\alpha W} & \frac{r}{\alpha W} \end{bmatrix} x \\ x = \begin{bmatrix} w_l \\ w_r \end{bmatrix} = \begin{bmatrix} w_1 \\ w_3 \end{bmatrix} = \begin{bmatrix} w_2 \\ w_4 \end{bmatrix} \quad (13) \\ u = [T_1 \quad T_2 \quad T_3 \quad T_4]^T, \quad y = \begin{bmatrix} v \\ \phi \end{bmatrix}$$

$$M = \begin{bmatrix} \frac{mr^2}{4} + \frac{r^2I}{\alpha W^2} & \frac{mr^2}{4} - \frac{r^2I}{\alpha W^2} \\ \frac{mr^2}{4} - \frac{r^2I}{\alpha W^2} & \frac{mr^2}{4} + \frac{r^2I}{\alpha W^2} \end{bmatrix} \\ B = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad (14)$$

where the coefficients in this model are defined in Table 1. Note that the UGV model is linear except for a possibly non-linear frictional force, $C(x)$, and the system is overactuated, because the B matrix does not have full column rank.

In simulations linear kinetic friction will be used,

$$C(\hat{x}) = \begin{bmatrix} k/2 & k/2 \\ k/2 & k/2 \end{bmatrix} \quad (15)$$

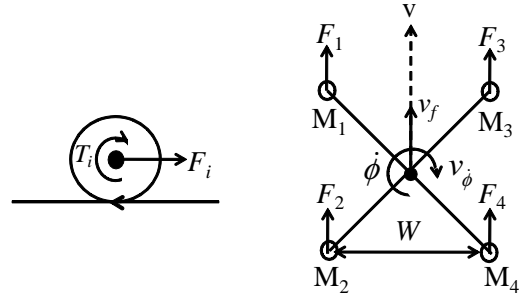


Figure 1. Visualization of motor torque allocation for a UGV

Symbol	Description	Units	Value
r	Wheel radius	m	0.1
W	Vehicle width	m	0.4
I	Wheel rotational inertia	kg·m ²	0.1
m	Vehicle mass	kg	1
$C(x)$	Frictional force	N	-
w_i	Wheel speed of motor i	rad/s	-
w_l	Left side wheel speed	rad/s	-
w_r	Right side wheel speed	rad/s	-
T_i	Torque produced by motor i	N·m	-
v	Vehicle speed	m/s	-
ϕ	Vehicle angular velocity	rad/s	-
α	Terrain-dependent parameter	-	-

Table 1. Definitions of symbols used in the UGV model

4.1 Prognostic Modeling

Winding insulation breakdown is a primary failure mechanism for the UGV's motors. The following model is used to estimate winding insulation lifetimes as a function of temperature,

$$L_N(t) = \alpha e^{-\beta T_W(t)} \quad (16)$$

where L_N is the expected remaining useful life (RUL) for new insulation in seconds and $T_W(t)$ ($^{\circ}\text{C}$) is the winding temperature at time t (Montsinger, 1930).

The RUL estimate for a motor winding at any given time is calculated using:

$$L(t) = L_N(t) \cdot \left(1 - \frac{d(t)}{100}\right) \quad (17)$$

where $d(t)$ is the percentage of insulation lifetime used prior to time t ,

$$d(t) = \int_0^t \frac{d\tau}{L(\tau)} \quad (18)$$

A probability distribution is added to the α coefficient in Eq. (16) to capture uncertainty in the prognostic model. Figure 2 shows the resulting probabilistic insulation life versus temperature model, where the pdf's mean corresponds to $\alpha = 10^5$ (s), and standard deviations are given by $\alpha = 10^5 \pm 1.5 \times 10^4$. The β coefficient in Eq. (16) is set to 0.035 ($^{\circ}\text{C}^{-1}$).

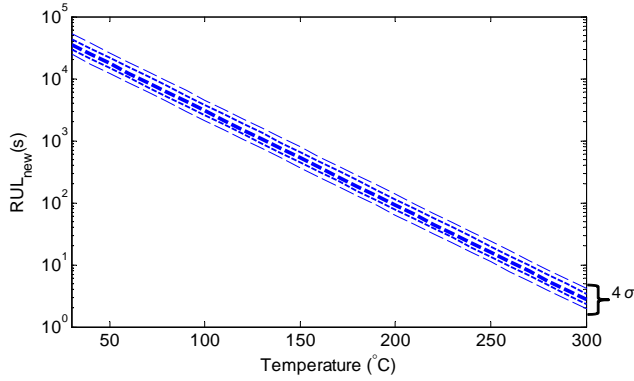


Figure 2. Addition of an uncertainty pdf to the insulation breakdown model.

Thermal Model

A first order thermo-electrical model, shown in Figure 3, is used to track the winding-to-ambient temperature as a function of copper losses,

$$\dot{T}_{wa} = -\frac{T_{wa}(t)}{R_{wa}C_{wa}} + \frac{P_{loss}(t)}{C_{wa}} \quad (19)$$

where T_{wa} is winding-to-ambient temperature, P_{loss} is power loss in the copper windings, C_{wa} is thermal capacitance, and R_{wa} is thermal resistance.

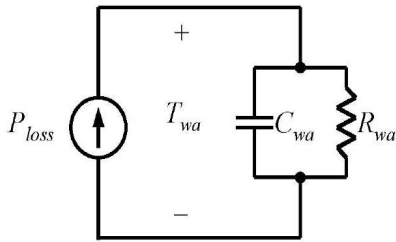


Figure 3. Thermal model for motor windings

5. SIMULATION STUDIES

In simulation studies of the UGV load-allocation problem, the intrinsic optimization problem, introduced in Eq. (1), is defined using the following performance and component degradation penalties:

$$J_M(\nu) = \frac{1}{T} \int_0^T \exp|\phi_c(t) - \phi(t)| + K_{p1} \quad (20)$$

$$J_d(\mathbf{u}) = \max_i \left[\exp\left(\tilde{d}_i(T)\right) \cdot \frac{4}{3} \right] + K_{p2} \quad (21)$$

where $\phi_c(t)$ and $\phi(t)$ represent waypoints for the desired and actual path followed by the UGV respectively. K_{p1} and K_{p2} are penalty functions that effectively enforce constraints on the maximum acceptable path error and the maximum acceptable VaR estimate at the end of a mission. Performance and component degradation constraints are defined as follows:

$$|\phi_c(t) - \phi(t)| < 1, \quad \forall t \in [0, \dots, T] \quad (22)$$

$$\tilde{d}_i(T) < 90\% \quad (23)$$

The performance and prognostic penalties introduced in Eq. (5) and Eq. (6) are defined as follows:

$$g(|\nu - \mathbf{r}|) = \int_t^{t+\tau} \left| \exp\left(\begin{bmatrix} r_f - \nu_f \\ r_\phi - \nu_\phi \end{bmatrix}\right) \right| dz \quad (24)$$

$$f\left(\tilde{d}_i(t+\tau)\right) = \lambda \cdot \sum_{i=1}^4 \left[\exp\left(\tilde{d}_i(t+\tau) - \gamma_i(t+\tau)\right) + \exp\left(\tilde{d}_i(t+\tau)\right) + C_p \right] \quad (25)$$

$$\Pr\left(d_i(t+\tau) > \tilde{d}_i(t+\tau)\right) = 2\% \quad (26)$$

where λ represents the relative value of maximizing performance and minimizing component degradations, $\gamma_i(t+\tau)$ is an upper-bound on the 98% confidence VaR estimates at time $t+\tau$, and C_p is an additional penalty that effectively disallows controls that cause the upper VaR bound to be exceeded (if other solutions exist). Simulation studies presented later in this paper will explore the effect of varying τ and λ on the system's intrinsic evaluation function. In the reported simulation studies, $\gamma_i(t+\tau)$ is defined using a linear interpolation from $\tilde{d}_i(t)$ to the maximum allowable degradation at the end of the mission. The effect of varying the formulation of $\gamma_i(t+\tau)$ on load-allocation in a triplex redundant electro-mechanical, was explored in a previous publication (Bole et al., 2010).

In simulations, the cost of possible motor load allocations is evaluated by assuming that the current demands on the system and the current component load allocations are constants over the prognostic horizon. The space of feasible motor-load allocations to be searched over is defined by the following performance constraint:

$$\begin{aligned} 0.8 \cdot r_L &\leq T_L \leq 1.2 \cdot r_L \\ 0.8 \cdot r_R &\leq T_R \leq 1.2 \cdot r_R \end{aligned} \quad (27)$$

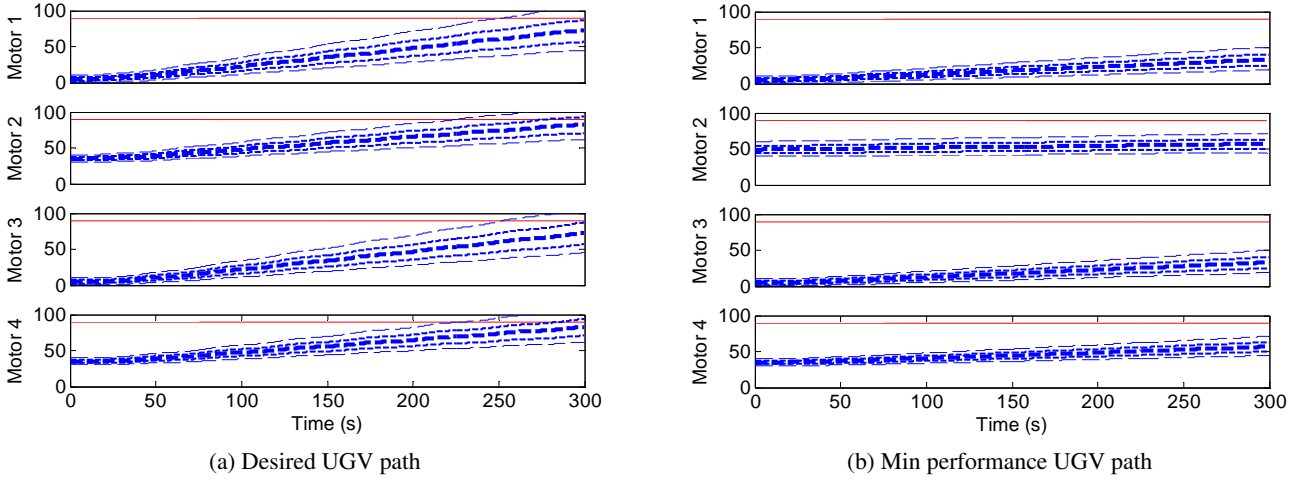


Figure 4. Plots of winding insulation degradation estimates (mean and ± 2 standard deviations) on the desired UGV path (a) and the minimum allowable performance path (b), using λ & τ such that $J_d(\mathbf{u}(t))|_{\lambda, \tau} = \min_{\lambda, \tau}[J_d(\mathbf{u}(t))]$

where r_L and r_R are desired net control effort outputs from the left-hand and right-hand motors respectively. The desired torque output from the UGV at a given time instant is defined by the following proportional control law:

$$\mathbf{r} = \begin{bmatrix} r_L + r_R \\ r_L - r_R \end{bmatrix} = \mathbf{r}_{\text{ref}}(t) + \begin{bmatrix} p_1 \cdot \cos(\phi_e) \cdot e_d \\ p_2 \cdot \sin(\phi_e) \end{bmatrix} \quad (28)$$

where $\mathbf{r}_{\text{ref}}(t)$ is output control effort that would be used at time t if the vehicle followed the reference path exactly, p_i are the proportional control coefficients, ϕ_e is the vehicle's heading error with respect to the reference path, and e_d is the vehicle's position error with respect to the reference path. Component load allocations for best risk management are found at each time-step by evaluating the objective function on a sufficiently dense uniform grid over the space of all component load allocations satisfying the performance constraints.

5.1 Verifying Mission Feasibility

In the simulation studies discussed here, the four-wheeled UGV is commanded to follow a figure-8 type path. By design, the commanded path is so demanding that following it exactly will yield no solutions to the load-allocation problem that satisfy the final VaR constraint (defined in Eq. (23)). The existence of solutions to the load-allocation problem that will not violate the performance and VaR constraints for the given mission is proven by verifying that using the minimum allowable UGV performance over the mission will allow all motors to end the mission with adequate health. Figure 4 shows simulation results for load-allocation controls that minimize the final VaR evaluation metric (defined in Eq. (21)) on the minimum allowable UGV performance path and the desired UGV path. As shown in the figure, the motors on each side of the vehicle are initialized at different levels of degrada-

tion in order to observe discrimination in the allocation motor loads based on their relative healths. The simulation results, shown in Figure 4, prove that although following the desired UGV path exactly is guaranteed to result in violation of the final VaR constraint, the load-allocation problem does have feasible solutions satisfying both the performance and VaR constraints.

5.2 Control with Foreknowledge of the Mission and the Fault Growth Model

Due to the fact that in simulation studies the desired path for the UGV and a fault growth model are known in advance, the optimal load allocations over the given mission can be approximated without the need for prognosis. Analysis of the direct optimization of the system's intrinsic cost function over a known mission will provide substantial insight into the development of prognostics-based risk-management controllers. Optimization routines will specify candidate UGV paths over a mission by defining a set of waypoints and using a third order spline to interpolate between those points.

The search space for the path planning routines is the set all adjustments to given waypoints that will not violate the performance constraint, given in Eq. (22). The net output control effort output required to follow a given path is found by inverting the modeled UGV dynamics given in Eq. (13),

$$\begin{bmatrix} T_L(t) \\ T_R(t) \end{bmatrix} = f^{-1}(\phi_p(t)), \quad \forall t \in [0, \dots, T] \quad (29)$$

where $\phi_p(t)$ is the (x, y) position of the UGV at time t .

Individual motor load allocations are derived using the following expression for splitting load proportionately among

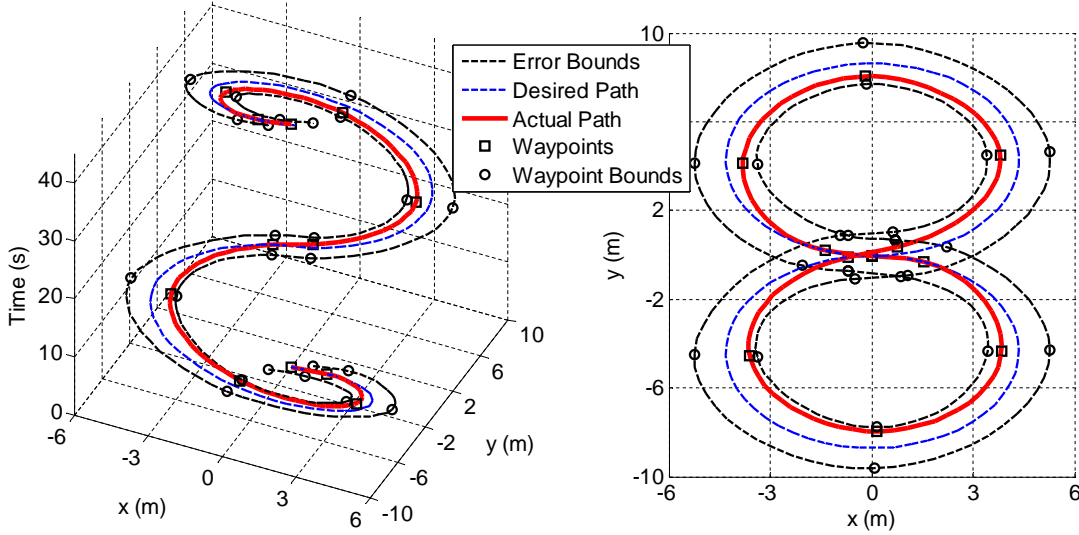


Figure 5. Results of search for optimal figure-8 path

	$\tilde{d}_1(T)$	$\tilde{d}_2(T)$	$\tilde{d}_3(T)$	$\tilde{d}_4(T)$	$J_d _0^T$	$J_M _0^T$	$(J_M + J_d) _0^T$
min allowable performance path	49%	73%	49%	73%	2.76	2.16	4.93
load-allocation with future knowledge	64%	64%	64%	64%	2.52	1.69	4.21
prognosis based load-allocation	70%	86%	70%	86%	3.17	1.58	4.76

Table 2. Results of simulation studies

the two motors on each side of the vehicle:

$$\begin{aligned} T_1(t) \cdot k_1 + T_2(t) &= T_L(t) \\ T_3(t) \cdot k_2 + T_4(t) &= T_R(t) \end{aligned} \quad (30)$$

Optimal motor load allocations for a given path are derived by evaluating Eq. (1) over sufficiently dense uniform grid on k_1 and k_2 , and selecting the value resulting in minimum cost. Figure 5 shows plots of the desired UGV path, the bounds on allowable path error, and an approximation of the optimal UGV path, for one cycle of the commanded figure-8 maneuver. The simulated mission consists of eight repetitions of this figure-8 maneuver. A nested optimization is used to estimate the optimal motor load allocations over the given mission. An outer-loop optimization routine uses a gradient descent search over the space of allowable adjustments to a set of waypoints, where the space of allowable adjustments to each waypoint is shown in Figure 5 as the linear region between the black circles. An inner loop optimization routine finds the net output torque from the left-hand and right-hand motors required to follow eight repetitions of a given figure-8 path, and then searches for the optimal proportional load split among the motors on each side of the vehicle using the uniform grid method described earlier.

Estimates of the optimal VaR metrics for the winding insulation degradations and the control evaluation costs for the given mission are shown in Table 2. Note that the estimated optimal motor load-allocations will result in final winding

VaR estimate being nearly equal for all four motors, due the fact that the control evaluation function is defined to penalize only the highest motor degradation. Also, note that the error between the commanded and the estimated optimal path is greatest in the extreme upper and lower regions of the figure-8 path because introducing an error in those regions results in the greatest reduction in the total distance traveled by the UGV. Both of these results are expected when the future commanded UGV path and the future fault grown model are known in advance; however, in general, it will be very difficult to match those results with controllers that rely on uncertain predictions of future states.

5.3 Prognostics-Based Control

At each control time-step, a prognostics-based controller will allocate motor loads to best manage the risk posed by uncertain estimates of future system performance and fault prognosis. In simulation, motor load-allocations for best risk management are derived by evaluating Eq. (6) on a sufficiently dense uniform grid over the space of all motor loads satisfying the performance constraint. Fundamentally, the prognostics-based control problem is to specify risk-reward evaluation metrics, of the form given in Eq. (6), that will result in the derived controls coming as close as possible to matching the minimum control evaluation metric achievable using foreknowledge. Figure 6 shows plots the intrinsic evaluation

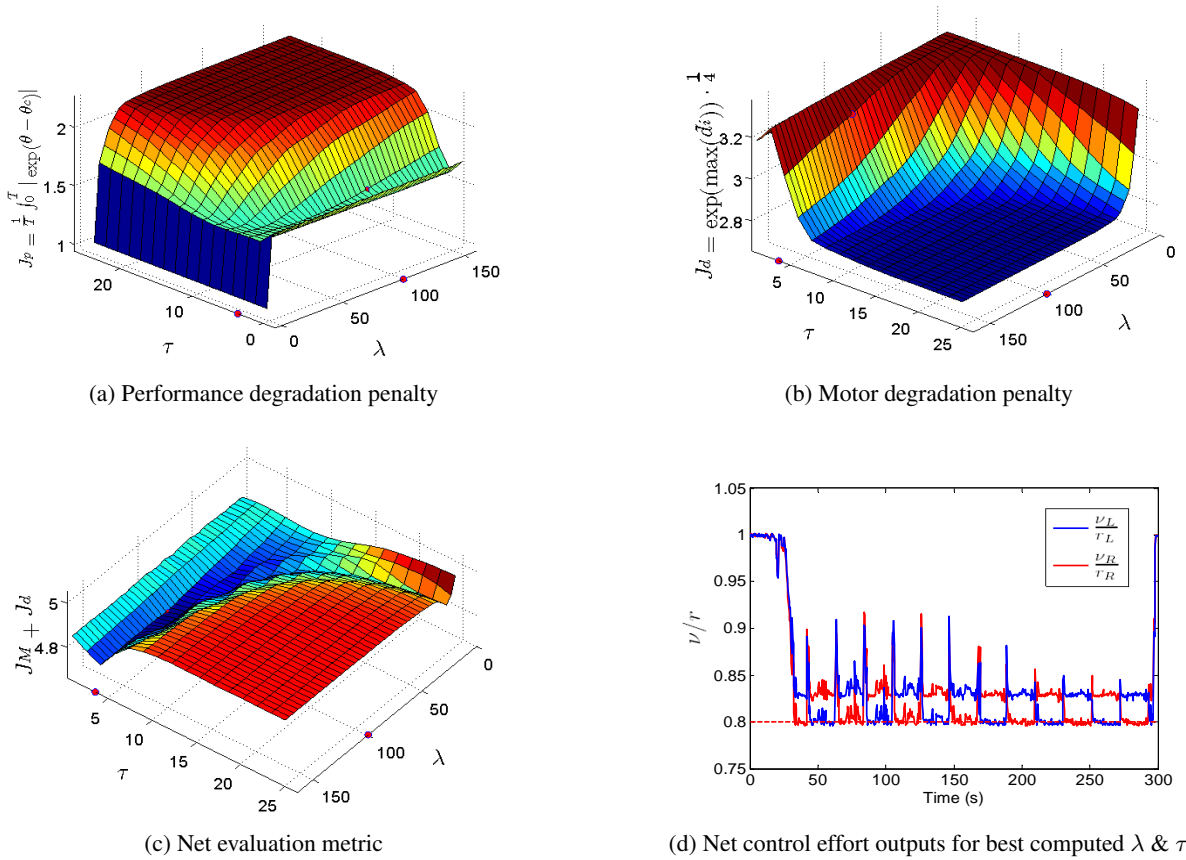


Figure 6. UGV simulation results for $\lambda = [0, 3, 6, \dots, 160]$ and $\tau = [.4, 1.6, 2.8, \dots, 24.4]$; optimal value at $\lambda = 100$ & $\tau = 3.6s$

metrics over a range of values for the prognostic horizon, τ , and the weighting factor, λ . The intrinsic evaluation metrics, shown in the plots, were obtained by computing the optimal motor load-allocation at each control time-step, after substituting τ and λ into the evaluation functions for predicted future component degradations and system performance, which were defined in Eq. (24) and Eq. (25). In general, as the prognostic horizon is increased the increased uncertainty in fault growth predictions will result in a greater perceived risk, and thus a more conservative control. Also, increasing the weighting factor, λ , on the prognostic penalty will tend to result in solutions with higher path errors and less component degradations. Plots of the intrinsic evaluation metrics versus λ and τ show these general trends. The trough seen in Figure 6 (c) indicates a domain of τ and λ values corresponding to controls that are neither overly conservative nor overly aggressive. The best computed intrinsic control evaluation costs and the corresponding winding insulation degradation VaR's at the end of the mission are given in Table 2. Future work will continue to explore the analytical relationships between the metrics used to evaluate risk from prognostic estimates and their resultant performance on example systems.

6. CONCLUSION

Any control technique that claims to manage or mitigate the risk posed by load dependent fault modes can be viewed as being implicitly derived based on the risk-reward optimization that was explicitly addressed in this work. The paper introduced a methodology for deriving and validating prognostics-based fault-adaptive control routines that began with the derivation of an expression for evaluating the desirability of future control outcomes, and eventually produced control routines that sought to optimize derived risk metrics using uncertain prognostic information. A case study on the design of fault-adaptive control for a skid-steered robot demonstrated some of the challenges associated with deriving risk metrics that will minimize the risk of component failures without becoming overly conservative and unnecessarily sacrificing performance. Future work will introduce more sophisticated methods for utilizing stochastic prognostic information to associate risk with a given distribution of component loads; more sophisticated methods for solving the resulting stochastic optimization problems will also be explored.

ACKNOWLEDGMENT

This work was supported by the United States National Aeronautics and Space Administration (NASA) under STTR Contract #NNX09CB61C. The program is sponsored by the NASA Ames Research Center (ARC), Moffett Field, California 94035. Dr. Kai Goebel, NASA ARC, Dr. Bhaskar Saha, MCT, NASA ARC and Dr. Abhinav Saxena, RIACS, NASA ARC, are the Technical POCs and Liang Tang is the Principal Investigator and Project Manager at Impact Technologies.

REFERENCES

- Abhken, P. (2000). An empirical evaluation of value at risk by scenario simulation. *Journal of Derivatives*, 7, 1074-1240.
- Arulampalam, S., Maskell, S., Gordon, N., & Clapp, T. (2002, Feb.). A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2), 174-188.
- Bole, B., Brown, D., Pei, H.-L., Goebel, K., Tang, L., & Vachtsevanos, G. (2010, Oct.). Fault adaptive control of overactuated systems using prognostic estimation. In *International conference on prognostics and health management (PHM)*.
- Caglayan, K., Allen, S., & Wehmuller, K. (1988). Evaluation of a second generation reconfiguration strategy for aircraft flight control systems subjected to actuator failure surface damage. In *Proceedings of the IEEE national aerospace and electronics conference* (p. 520-529).
- Doyle, J., Glover, K., Khargonekar, P., & Francis, B. (1987). State-space solutions to standard h_2 and h_∞ optimal control problems. *IEEE Transactions on Automatic control*, 33, 831-847.
- Gergely, E., Spoiála, D., Spoiála, V., Silaghi, H., & Nagy, Z. (2008). Design framework for risk mitigation in industrial PLC control. In *IEEE international conference on automation, quality and testing, robotics* (Vol. 2, p. 198-202).
- Gokdere, L., Bogdano, A., Chiu, S., Keller, K., & Vian, J. (2006). Adaptive control of actuator lifetime. In *IEEE aerospace conference*.
- Harkegard, O., & Glad, T. (2005, Jan.). Resolving actuator redundancy - optimal control vs. control allocation. *Automatica*, 41(1), 137-144.
- Hattori, Y., Koibuchi, K., & Yokoyama, T. (2002, Sept.). Force and moment control with nonlinear optimum distribution for vehicle dynamics. In *Proc. of the 6th international symposium on advanced vehicle control*.
- Karpenko, M., & Sepehri, N. (2005, January). Fault tolerant control of a servohydraulic positioning system with crossport leakage. *IEEE Transactions on Control Systems Technology*, 13(1), 155-161.
- Lauterbach, B., & Schulz, P. (1990). Pricing warrants: An empirical study of the black-scholes model and its alternatives. *Journal of Finance*, 45, 1181-1209.
- Monaco, J., Ward, D., & Bateman, A. (2004, Sept.). A retrofit architecture for model-based adaptive flight control. In *AIAA 1st intelligent systems technical conference*.
- Montsinger, V. M. (1930). Loading transformers by temperature. *Transactions of the American Institute of Electrical Engineers*, 32.
- Oppenheimer, M., Doman, D., & Bolender, M. (2006). Control allocation for over-actuated systems. In *14th mediterranean conference on control and automation*.
- Orchard, M., Kacprzynski, G., Goebel, K., Saha, B., & Vachtsevanos, G. (2008). Advances in uncertainty representation and management for particle filtering applied to prognostics. In *International conference on prognostics and health management PHM*.
- Rao, B. (1998). *Handbook of condition monitoring* (A. Davies, Ed.). Chapman and Hall.
- SAE. (1994). *Potential failure mode and effects analysis in design (design FMEA) and potential failure mode and effects analysis in manufacturing and assembly processes, reference manual* (Tech. Rep. No. J1739).
- Saglimbene, M. (2009). Reliability analysis techniques: How they relate to aircraft certification. In *Reliability and maintainability symposium* (p. 218-222).
- Saha, B., & Goebel, K. (2008). Uncertainty management for diagnostics and prognostics of batteries using bayesian techniques. In *IEEE aerospace conference*.
- Schreiner, A., Balzer, G., & Precht, A. (2008). Risk analysis of distribution systems using value at risk methodology. In *Proceedings of the 10th international conference on probabilistic methods applied to power systems*.
- Schreiner, A., Balzer, G., & Precht, A. (2010). Risk sensitivity of failure rate and maintenance expenditure: application of var metrics in risk management. In *15th IEEE mediterranean electrotechnical conference* (p. 1624-1629).
- Sheppard, J., Butcher, S., Kaufman, M., & MacDougall, C. (2006). Not-so-naïve bayesian networks and unique identification in developing advanced diagnostics. In *IEEE aerospace conference* (p. 1-13).
- Srivastava, A., Mah, R., & Meyer, C. (2008, Dec.). *Integrated vehicle health management automated detection, diagnosis, prognosis to enable mitigation of adverse events during flight* (Tech. Rep.). Version 2.02, National Aeronautics and Space Administration.
- Venkataraman, S. (1997). Value at risk for a mixture of normal distributions: the use of quasi-bayesian estimation techniques. *Economic Perspectives*, 2-13.

Brian M. Bole graduated from the FSU-FAMU School of Engineering in 2008 with a B.S. in Electrical and Computer Engineering and a B.S. in Applied Math. He received a M.S. degree in Electrical Engineering from the Georgia Institute

of Technology in 2011, and he is currently pursuing a Ph.D. Brian's research interests include stochastic optimization, robust control, fault prognosis, and risk management. Brian is currently instigating the use of risk management and stochastic optimization techniques for optimal adaptation of active component load allocations in robotic and aviation applications. In a previous project, Brian work with the Georgia Tech EcoCar team to develop a stochastic-optimization based controller for optimizing fuel economy on a charge sustaining hybrid electric vehicle.

Kai Goebel received the degree of Diplom-Ingenieur from the Technische Universität München, Germany in 1990. He received the M.S. and Ph.D. from the University of California at Berkeley in 1993 and 1996, respectively. Dr. Goebel is a senior scientist at NASA Ames Research Center where he leads the Diagnostics & Prognostics groups in the Intelligent Systems division. In addition, he directs the Prognostics Center of Excellence and he is the Associate Principal Investigator for Prognostics of NASA's Integrated Vehicle Health Management Program. He worked at General Electric's Corporate Research Center in Niskayuna, NY from 1997 to 2006 as a senior research scientist. He has carried out applied research in the areas of artificial intelligence, soft computing, and information fusion. His research interest lies in advancing these techniques for real time monitoring, diagnostics, and prognostics. He holds ten patents and has published more than 100 papers in the area of systems health management.

Liang Tang is a Lead Engineer at Impact Technologies LLC, Rochester, NY. His research interests include diagnostics, prognostics and health management systems (PHM), fault tolerant control, intelligent control, and signal processing. He obtained a Ph.D. degree in Control Theory and Engineering from Shanghai Jiao Tong University, China in 1999. Before he joined Impact Technologies, he worked as a post doctoral research fellow at Intelligent Control Systems Laboratory, Georgia Institute of Technology. At Impact Technologies he is responsible for multiple DoD and NASA funded research and development projects on structural integrity prognosis, prognostics and uncertainty management, automated fault accommodation for aircraft systems, and UAV controls. Dr. Tang has published more than 30 papers in his areas of expertise.

George J. Vachtsevanos is a Professor Emeritus of Electrical and Computer Engineering at the Georgia Institute of Technology. He was awarded a B.E.E. degree from the City College of New York in 1962, a M.E.E. degree from New York University in 1963 and the Ph.D. degree in Electrical Engineering from the City University of New York in 1970. He directs the Intelligent Control Systems laboratory at Georgia

Tech where faculty and students are conducting research in intelligent control, neurotechnology and cardiotechnology, fault diagnosis and prognosis of large-scale dynamical systems and control technologies for Unmanned Aerial Vehicles. His work is funded by government agencies and industry. He has published over 240 technical papers and is a senior member of IEEE. Dr. Vachtsevanos was awarded the IEEE Control Systems Magazine Outstanding Paper Award for the years 2002-2003 (with L. Wills and B. Heck). He was also awarded the 2002-2003 Georgia Tech School of Electrical and Computer Engineering Distinguished Professor Award and the 2003-2004 Georgia Institute of Technology Outstanding Interdisciplinary Activities Award.

An Adaptive Particle Filtering-based Framework for Real-time Fault Diagnosis and Failure Prognosis of Environmental Control Systems

Ioannis A. Raptis¹, George Vachtsevanos¹

¹ *Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA.*
iraptis@gatech.edu
gfv@ece.gatech.edu

ABSTRACT

Maintenance of critical or/complex systems has recently moved from traditional preventive maintenance to Condition Based Maintenance (CBM) exploiting the advances both in hardware (sensors / DAQ cards, etc.) and in software (sophisticated algorithms blending together the state of the art in signal processing and pattern analysis). Along this path, Environmental Control Systems and other critical systems/processes can be improved based on concepts of anomaly detection, fault diagnosis and failure prognosis. The enabling technologies borrow from the fields of modeling, data processing, Bayesian estimation theory and in particular a technique called particle filtering. The efficiency of the diagnostic approach is demonstrated via simulation results.

1. INTRODUCTION

Heating, Ventilating and Air Conditioning (HVAC) systems have a large span of applications ranging from industrial buildings, households to small scale units installed in aerial and ground vehicles operating as part of Environmental Control Systems (ECS). Defective or faulty operation of such systems have both environmental and economical impact. Typical drawbacks are high operating cost, maintenance cost and thermal discomfort.

A standard ECS system is composed of four main components that are encountered in subcritical vapor compression cycles: The evaporator, condenser, Thermostatic Expansion Valve (TEV) and compressor. The refrigerant enters the compressor as a superheated vapor at a low pressure. In the compressor, the refrigerant is compressed to a high pressure and it is routed to the condenser. At this higher pressure, the refrigerant has a higher temperature than the ambient conditions and the refrigerant condenses. The refrigerant exits the condenser as a subcooled liquid at a higher pressure and passes through the thermostatic expansion device. At the

exit of the expansion valve, the refrigerant is at low pressure and routed to the evaporator. At this lower pressure the refrigerant has a lower temperature than ambient conditions, therefore heat is transferred to the refrigerant, and the refrigerant evaporates. Finally the refrigerant re-enters the compressor and the cycle is repeated. The main components as well as the several phases of the thermo-fluid in a vapor compression cycle with two-phase heat exchangers are depicted in Figure 1.

From first principles modeling, the dynamic behavior of thermo-fluid systems is dictated by highly coupled, nonlinear partial differential equations. Such equations are both complicated to handle for analysis and conducting numerical simulations. The main difficulty in the dynamic modeling of vapor compression cycles is the representation of the thermo-fluid inside the two-phase heat exchanger. Wedekind's work (Wedekind, Bhatt, & Beck, 1978) indicated that two-phase transient flow problems can be converted into lumped-parameter systems of nonlinear ordinary differential equations assuming that the mean void fraction remains relatively invariant in the two-phase section of a heat exchanger. This approach has also been adopted in this paper following the work reported in (X. He, 1996; X.-D. He & Asada, 2003; Rasmussen, 2005).

The ECS systems are subjected to various fault conditions. A survey for the most common faults encountered in ECS systems is given in (Comstock, Braun, & Groll, 2002). In this paper, the fault under consideration is the refrigerant leakage that takes place in the evaporator. According to (Braun, 2003), refrigerant leakage accounts for about 12% of the total service calls in response to a loss of cooling. All refrigeration systems have the potential to leak because pressures in the system are usually many times higher than atmospheric. Loss of refrigerant from industrial and commercial refrigeration systems can occur: (a) due to gradual leakage from joints or seals (b) through accidental rupture of a pipe or joint takes place and results in a significant loss of refrigerant charge in a short period of time and (c) during servicing when some refrigerant can be accidentally vented to gain access to a section of pipe or a given piece of equipment for repair.

A statistical rule-based method for Fault Detection

This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

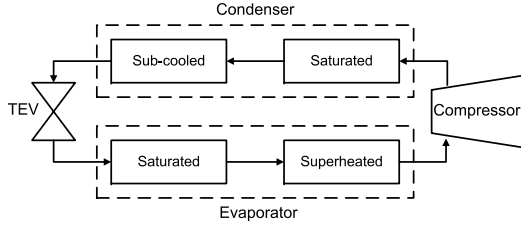


Figure 1: Main components of the vapor compression cycles and thermo-fluid phases

and Diagnosis (FDD) has been applied in for packaged air conditioning systems in (Li & Braun, 2003; Rossi & Braun, 1997) and for rooftop air conditioner in (Breuker & Braun, 1998). In (Stylianou & Nikanpour, 1996), an FDD method is employed for reciprocating chillers utilizing physical modeling, pattern recognition and expert knowledge. An on-line refrigerant leakage detection scheme is proposed in (Navarro-Esbri, Torrella, & Cabello, 2006) using adaptive algorithms.

This paper presents the implementation of an on-line particle-filtering-based framework for fault diagnosis and failure prognosis in a two-phase heat exchanger of an ECS. The methodology considers an autonomous module, and assumes the existence of fault indicators (for monitoring purposes) and the availability of real-time measurements. The fault detection and identification (FDI) module uses a hybrid state-space model of the plant, and a particle filtering algorithm to calculate the probability of leakage in the evaporator; simultaneously computing the state probability density function (pdf) estimates.

The failure prognosis module, on the other hand computes the remaining useful life (RUL) estimate of the faulty subsystem in real time, using the the detection algorithm current state estimates of the nonlinear state-space fault growth model and predicts the evolution in time of the probability distribution of the leaked mass.

The enabling technologies borrow from the fields of modeling, data processing, Bayesian estimation theory and in particular a technique called particle filtering. The proposed FDI framework is enhanced with an additional particle filtering routine that is executed in parallel with the state estimator, which estimates the unknown model parameters of the leakage progression model. The simulation result indicate that the proposed dual particle filtering scheme is highly adaptive and reliable even for abrupt crack that cause leakage.

This methodology allows the inclusion of customer specifications (statistical confidence in fault detection, minimum prediction window, etc.) in a simple and direct way. Moreover, all the outcomes are easily provided to plant operators through real-time updated graphs and may be easily coded and embedded in compact modules.

This paper is organized as follows: Section 2 presents the evaporator model which was used for the numerical simulations. In Section 3 the leakage flow rate progression model is given. The technical approach of the detection algorithm is presented in Section 4. The prognostic module is presented in Section 5. Results in the form of numerical simulations are given in Section 6. Finally, concluding remarks are given in Section 7.

2. EVAPORATOR MODELING

For the dynamic representation of the evaporator, this paper adopts the modeling approach introduced by (Grald & MacArthur, 1992; X.-D. He & Asada, 2003; Cheng, He, & Asada, 2004). This approach is based on the work reported in (Wedekind et al., 1978) where a mean void fraction is used in the two-phase region of the heat exchanger. The model converts the two-phase evaporating flow system into a type of lumped parameter system. The dynamics of the two heat exchangers use a moving boundary layer model to separate the distinct two-phase liquid from the single-phase superheated of the evaporator. Based on (X. He, 1996), the fundamental standing assumptions of the heat exchangers dynamic model are:

1. One dimensional fluid flow
2. Negligible heat conduction along the axial directions of heat exchangers
3. Invariant mean void fraction in the two-phase sections during a short transient
4. Negligible refrigerant pressure drop along the heat exchangers

Using the mean void fraction assumption the mass balance equation can be written as:

$$\frac{d}{dt} \{[\rho_l (1 - \bar{\gamma}) + \rho_g \bar{\gamma}] A_t l_e\} = \dot{m}_{in} - \dot{m}_{mid} \quad (1)$$

where $\bar{\gamma}$ is the mean void fraction, l_e is the tube length that corresponds to the two-phase section, A_t is the cross section area of the tube, \dot{m}_{in} is the inlet flow rate, \dot{m}_{mid} is the flow rate of the moving boundary, and ρ_l, ρ_g are the refrigerant's liquid and vapor density in the two-phase section. The energy balance equation can be written as:

$$\frac{d}{dt} \{[\rho_l h_l (1 - \bar{\gamma}) + \rho_g h_g \bar{\gamma}] A_t l_e\} = \underbrace{[h_l (1 - x) + h_g x] \dot{m}_{in}}_{h_{in}} - \underbrace{h_g \dot{m}_{mid} + l_e \pi U_w D_t (T_w - T_e)}_{Q(l_e, T_e, T_w)} \quad (2)$$

where x is the inlet vapor quality, D_t is the diameter of the tube, U_w is the heat transfer coefficient between the tube wall and the refrigerant, T_e is the temperature of the refrigerant in the two-phase section, T_w is the temperature of the tube wall, and h_l, h_g are the specific enthalpies of the refrigerant liquid and vapor, respectively. The first term in the right hand side of Eq. (2) represents the rate at which energy enters the two-phase region by the inlet mass flow rate, the second term represents the rate at which thermal energy exits the two-phase region, by the outlet mass flow rate. The last term represent the heat transfer rate from the tube wall to the refrigerant. Multiplying Eq. (1) with h_g and subtracting Eq. (2) one gets:

$$\frac{d}{dt} [\rho_l (1 - \bar{\gamma}) h_{lg} A_t l_e] = (1 - x) h_{lg} \dot{m}_{in} - Q \quad (3)$$

where $h_{lg} = h_g - h_l$. Assuming that the refrigerant properties remain constant over the time step, the moving boundary dynamics can be written as:

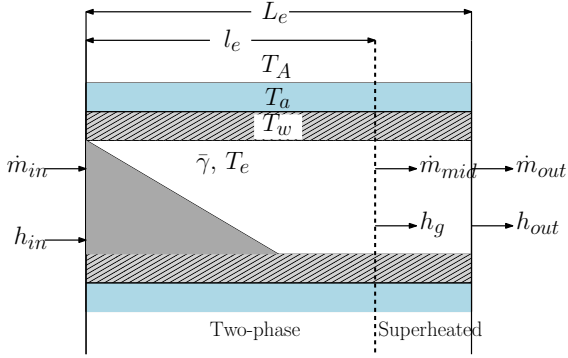


Figure 2: Schematic of the evaporator

$$\frac{dl_e}{dt} = -\frac{\pi U_w D_t (T_w - T_e)}{\rho_l (1 - \bar{\gamma}) h_{lg} A_t} l_e + \frac{1 - x}{\rho_l (1 - \bar{\gamma}) A_t} \dot{m}_{in} \quad (4)$$

The dynamic equation of the system's second state variable is produced by the vapor balance in the evaporator. The vapor mass flow rate entering the evaporator is \dot{m}_{inx} . The vapor mass flow rate exiting the evaporator is \dot{m}_{out} while the rate of vapor generated from liquid during the evaporation process in the two-phase section is $Q(l_e, T_e, T_w)/h_{lg}$. Assuming that in the evaporator the vapor volume is significantly larger than the liquid volume, the vapor balance equation is given by:

$$\frac{dm_e}{dt} = V_e \frac{d\rho_g}{dt} = \dot{m}_{inx} - \dot{m}_{out} + \frac{Q}{h_{lg}} \quad (5)$$

where m_e and V_e are the total vapor mass and total volume of the evaporator. Since ρ_g is the denotes the vapor saturated density, from the state equation one can easily obtain the one-to-one mapping $T_e(\rho_g(t))$.

It is assumed that the wall temperature is spatially uniform. The one dimensional energy balance equation for the tube wall is given by

$$c_w p_w A_w \frac{dT_w}{dt} = U_w A_w (T_e - T_w) + U_a A_a (T_a - T_w) \quad (6)$$

where c_w , p_w and A_w are the specific heat, density and cross sectional area of the tube wall, respectively. In addition, U_a denotes the heat transfer coefficient between the tube wall and the air, A_a the surface area between the tube wall and the air and T_a the air exiting temperature from the evaporator. Similarly, assuming that the air exit temperature is spatially uniform, the one dimensional energy balance equation can be written as

$$c_a p_a A_a \frac{dT_a}{dt} = U_w A_w (T_w - T_a) + \dot{m}_a (T_A - T_a) \quad (7)$$

where c_a , p_a and A_a are the specific heat, density and cross sectional area of the air tube, respectively. The variable T_A denotes the air temperature at the entrance of the evaporator and \dot{m}_a the air flow rate.

3. REFRIGERANT LEAKAGE MODEL

The refrigerant leakage is produced typically by a crack in the system pipes or by a faulty connection of the pipe system joints. Based on (Merritt, 1967), the mass flow rate of the leaked refrigerant is given by the following equation:

$$\frac{dm_{leak}}{dt} = c_d A_l \sqrt{2\rho(P - P_o)} \quad (8)$$

where m_{leak} is the mass of the leaked refrigerant, c_d is the discharge coefficient, A_l is the crack surface, ρ is the refrigerant density, P is the refrigerant pressure inside the pipe and P_o is the refrigerant pressure outside the pipe. Assuming that the outer pressure is significantly smaller than the refrigerant pressure inside the pipe, the refrigerant leakage growth model can be approximated by:

$$\frac{dm_{leak}}{dt} = C_r \sqrt{\rho_l P_e} \quad (9)$$

where P_e is the refrigerant pressure in the two-phase section of the evaporator $C_r = \sqrt{2}c_d A_l$.

4. TECHNICAL APPROACH - THE DIAGNOSTIC ALGORITHMS

A fault diagnosis procedure involves the tasks of fault detection and isolation (FDI), and fault identification (assessment of the severity of the fault). In general, this procedure may be interpreted as the fusion and utilization of the information present in a feature vector (measurements), with the objective of determining the operating condition (state) of a system and the causes for deviations from particularly desired behavioral patterns. Several ways to categorize FDI techniques can be found in literature. FDI techniques are classified according to the way that data is used to describe the behavior of the system: *data-driven* or *model-based* approaches.

Data-driven FDI techniques (Chen, Zhang, & Vachtsevanos, n.d.; Chen, Vachtsevanos, & Orchard, 2010) usually rely on signal processing and knowledge-based methodologies to extract the information hidden in the feature vector (also referred to as measurements). In this case, the classification/prediction procedure may be performed on the basis of variables that have little (or sometimes completely lack of) physical meaning. On the other hand, model-based techniques, as the name implies, use a description of a system (models based on first principles or physical laws) to determine the current operating condition.

A compromise between both classes of FDI techniques is often needed when dealing with complex nonlinear systems, given the difficulty of collecting useful faulty data (a critical aspect in any data-driven FDI approach) and the expertise needed to build a reliable model of the monitored system (a key issue in a model-based FDI approach).

From a nonlinear Bayesian state estimation standpoint, this compromise between data-driven and model-based techniques may be accomplished by the use of a Particle Filter (PF) based module built upon the dynamic state model describing the time progression or evolution of the fault (Orchard & Vachtsevanos, 2009; Chen, Brown, Sconyers, Vachtsevanos, & Zhang, 2010;

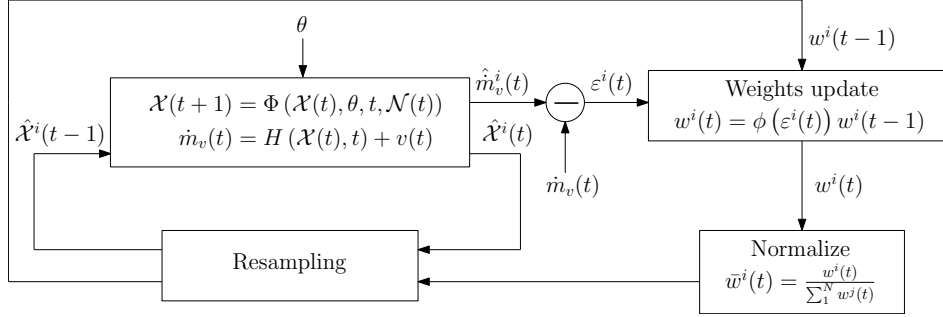
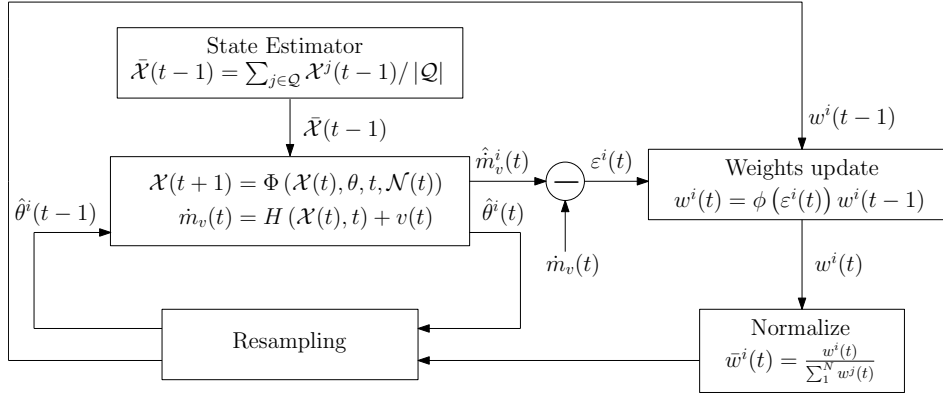


Figure 3: Block diagram of the PF algorithm for state estimation

Figure 4: Block diagram of the PF algorithm for parameter estimation where $|\mathcal{Q}|$ denotes the cardinality of the set $\mathcal{Q} = \{j \in 1, \dots, N : x_{d,2}^j(t-1) = 1\}$.

Orchard & Vachtsevanos, 2007). The fault progression is often nonlinear and, consequently, the model should be nonlinear as well. Thus, the diagnostic model is described by:

$$\begin{aligned} x_d(t+1) &= f_d(x_d(t), n(t)) \\ x_c(t+1) &= f_t(x_d(t), x_c(t), \omega(t)) \\ y(t) &= h_t(x_d(t), x_c(t), v(t)) \end{aligned} \quad (10)$$

where f_b , f_t , and h_t are nonlinear mappings, $x_d(t)$ is a collection of Boolean states associated with the presence of a particular operating condition in the system (normal operation, fault type #1, #2, etc.), $x_c(t)$ is a set of continuous-valued states that describe the evolution of the system given those operating conditions, $y(t)$ denotes the available measurements, $\omega(t)$ and $v(t)$ are non-Gaussian distributions that characterize the process and feature noise signals respectively. Since the noise signal $n(t)$ is a measure of uncertainty associated with Boolean states, it is advantageous to define its probability density through a random variable with bounded domain. For simplicity, $n(t)$ may be assumed to be uniform white noise (Orchard & Vachtsevanos, 2007). The PF approach using the above model allows statistical characterization of both Boolean and continuous-valued states, as new feature data (measurements) are received. As a result, at any given instant of time, this framework provides an estimate of the probability masses associated with each

fault mode, as well as a pdf estimate for meaningful physical variables in the system. Once this information is available within the FDI module, it is conveniently processed to generate proper fault alarms and to report on the statistical confidence of the detection routine.

One particular advantage of the proposed particle filtering approach is the ability to characterize the evolution in time of the above mentioned nonlinear model through modification of the probability masses associated with each particle, as new data from fault indicators are received. In addition, pdf estimates for the system continuous-valued states provide the capability of performing swift transitions to failure prognosis algorithms, one of the main advantages offered by the proposed approach.

The PF based FDI module is implemented accordingly using the non-linear time growth model given in Eq. (9) to describe the expected leaked mass flow rate. The goal is for the algorithm to make an early detection of the evaporator's leakage due to an unexpected crack or a faulty connection to the evaporators pipes. Two main operating conditions are distinguished: The normal condition reflects the fact that there is no leakage while a faulty condition indicating an unexpected crack in the evaporator which causes leakage. Denote by $x_{d,1}$ and $x_{d,2}$ two Boolean states that indicate normal and faulty conditions respectively. The nonlinear model is given by:

$$\begin{aligned} \begin{bmatrix} x_{d,1}(t+1) \\ x_{d,2}(t+1) \end{bmatrix} &= f_b \left(\begin{bmatrix} x_{d,1}(t) \\ x_{d,2}(t) \end{bmatrix} + n(t) \right) \\ \dot{m}_{leak}(t) &= \theta(t)x_{d,2}(t)\sqrt{p_l(T_e(t))P_e(t)} + \omega(t) \\ \dot{m}_v(t) &= h(\dot{m}_{leak}(t), t) + v(t) \end{aligned} \quad (11)$$

where

$$f_b(x) = \begin{cases} [1 & 0]^T & \text{if } \|x - [1 \ 0]^T\| \leq \|x - [0 \ 1]^T\| \\ [0 & 1]^T & \text{else} \end{cases}$$

$$[x_{d,1}(t_o) \ x_{d,2}(t_o) \ \dot{m}_{leak}(t_o)]^T = [1 \ 0 \ 0]^T \quad (12)$$

In the above equations $\theta(t)$, is a time-varying model parameter that represents the crack surface (the crack constant C_r) in the evaporator that causes the leak. The one-to-one mapping $h(\cdot)$ is also referred to as fault-to-feature mapping and it is obtained using the first principles model described in Section 2. A more practical approach is to approximate the fault-to-feature mapping by a neural network that assigns the operating conditions and the leakage to the valve flow rate. In particular set $h(\dot{m}_{leak}(t), t) \cong \Psi_{NN}(\dot{m}_{leak}(t), T_e(t), T_A(t))$.

The inlet flow rate of the evaporator is substituted by \dot{m}_v that denotes the TEV flow rate. It is assumed that \dot{m}_v can be measured. The above system can be written in a more compact form as

$$\mathcal{X}(t+1) = \Phi(\mathcal{X}(t), \theta, t, \mathcal{N}(t)) \quad (13)$$

$$\dot{m}_v(t) = H(\mathcal{X}(t), t) + v(t) \quad (14)$$

where $\mathcal{X}^T = [x_{d,1} \ x_{d,2} \ \dot{m}_{leak}]^T$ and $\mathcal{V}^T = [n \ \omega]^T$. The steps of the PF algorithm execution are described below:

1. From Eq. (13) generate N state estimates (particles) denoted by $\hat{\mathcal{X}}^i(t)$ where $i = 1, \dots, N$. To generate the state estimates, use a zero mean Gaussian distribution for $\omega(t)$ and uniform white noise for $n(t)$.
2. From Eq. (14) calculate the liquid side flow rate estimates denoted by \hat{m}_v^i , substituting the particles $\hat{\mathcal{X}}^i(t+1)$ to the mapping $H(\cdot)$.
3. Calculate the N errors $\varepsilon^i = \hat{m}_v^i - \dot{m}_v$, and assign to each particle $\hat{\mathcal{X}}^i(t)$ a weight $w^i(t) = \phi(\varepsilon^i)$, where $\phi(\cdot)$ denotes the standard normal distribution.
4. Normalize the weights $w^i(t)$. The normalized weights $\bar{w}^i(t)$ represent the discrete probability masses of each state estimate.
5. Calculate the final state estimate $\tilde{\mathcal{X}}(t)$ using the weighted sum of all the states $\hat{\mathcal{X}}^i(t)$.

An important part of the PF algorithm is the resampling procedure. Resampling is an action that takes place to counteract the degeneracy of the particles caused by estimates that have very low weights. A block diagram of

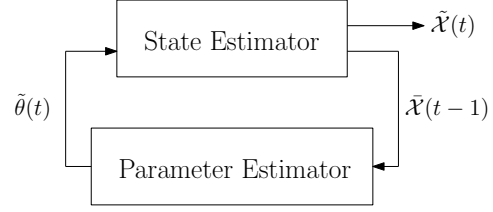


Figure 5: Block diagram of the complete FDI algorithm that utilizes a dual PF for state and parameter estimation.

the PF algorithm is given in Figure 3. An obvious shortcoming of the above procedure is that the crack coefficient C_r is unknown in a real life application. To this extent we augment to the standard FDI algorithm a parameter estimator module utilizing also PF with the objective to identify on-line the time-varying parameter $\theta(t)$. The parameter estimator is executed in parallel with the state estimator. The execution steps of the PF algorithm for parameter estimation are described below:

1. For each faulty particle ($x_{d,2}^i(t) = 1$) calculate the mean state $\bar{\mathcal{X}}(t)$. If there are not any faulty particles exit the parameter estimation module.
2. Using $\bar{\mathcal{X}}(t)$ from Eq. (13) generate N_θ parameter estimates (particles) denoted by $\hat{\theta}^i(t)$ where $i = 1, \dots, N_\theta$.
3. From Eq. (14) calculate the liquid side flow rate estimates denoted by \hat{m}_v^i , substituting the particles $\bar{\mathcal{X}}^i(t+1)$ to the mapping $H(\cdot)$.
4. Calculate the N errors $\varepsilon^i = \hat{m}_v^i - \dot{m}_v$, and assign to each particle $\hat{\theta}^i(t)$ a weight $w^i(t) = \phi(\varepsilon^i)$, where $\phi(\cdot)$ denotes the standard normal distribution.
5. Normalize the weights $w^i(t)$. The normalized weights $\bar{w}^i(t)$ represent the discrete probability masses of each state estimate.
6. Calculate the final state estimate $\tilde{\theta}(t)$ using the weighted sum of all the states $\hat{\theta}^i(t)$.

The resampling module takes place in an identical way as for the state estimator. A block diagram of the PF algorithm for parameter estimation is given in Figure 4. The interconnection of the two PF modules for both state and parameter estimation is given in Figure 5.

5. PROGNOSIS

Prognosis can be essentially understood as the generation of long-term predictions describing the evolution in time of a particular signal of interest or fault indicator. The goal of the prognostic algorithm is to use the evolution of the fault indicator in order to estimate the RUL of a failing component. Since prognosis intends to project the current condition of the indicator, it necessarily entails large-grain uncertainty. This paper adopts a prognosis scheme based on recursive Bayesian estimation techniques, combining the information from fault growth models and on-line data from sensors monitoring the plant.

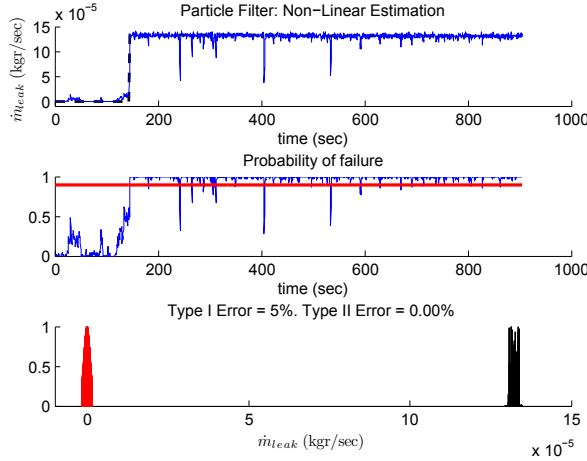


Figure 6: PF based FDI module

The prognosis algorithm is activated when a fault has been declared by the FDI module. Prognosis is a problem that goes beyond the scope of filtering applications, since it involves future time horizons. The main idea of the prognostic algorithm is to project the continuous state particle population in time, using the fault growth model, in order to estimate the time-to-failure (TTF) of each particle. Considering the nonlinear model given in Eq. (11) and using the notation of the diagnostic model introduced in Eq. (10), the progression in time of the continuous state can be written us:

$$x_c(t+1) = \psi(x_c(t), t) \quad (15)$$

The above equation represents the nonlinear mapping $f_t(\cdot)$, initially introduced in Eq. (10). From this mapping we have excluded the dependence of the noise $\omega(t)$ and the dependence of the boolean state $x_d(t)$, since in the prognostic mode a fault has already been detected. The inclusion of the time variable in the definition of the nonlinear mapping $\psi(\cdot)$ allows the investigation of time varying fault growth models. The execution of the prognostic algorithm at each time instant includes the following steps:

1. At each time instant, receive from the fault detection module N particles of the continuous state denoted by $\hat{x}_c^i(t)$, where $i = 1, \dots, N$. For each particle, using the fault growth model, iterate p^i steps in time, with $p^i \in \mathbb{N}$, such that the p^i -step ahead predictions given by:

$$\hat{x}_c^i(t+p^i) = \psi(\hat{x}_c^i(t+p^i-1), (t+p^i-1))$$

are such that $X_{hazard}^{low} \leq \hat{x}_c^i(t+p^i) \leq X_{hazard}^{high}$, where X_{hazard}^{low} , X_{hazard}^{high} denote the upper and lower bounds of a hazard zone that designate the limits of a critical failure. The initial estimates \hat{x}_c^i are taken directly by the PF detection algorithm described in the previous section.

2. Using the RUL estimates (p^i) and the normalized weights of the detection algorithm ($\bar{w}^i(t)$), the pdf and the weighted estimate of the RUL, denoted as \hat{t}_{RUL} , can be obtained for each time step.

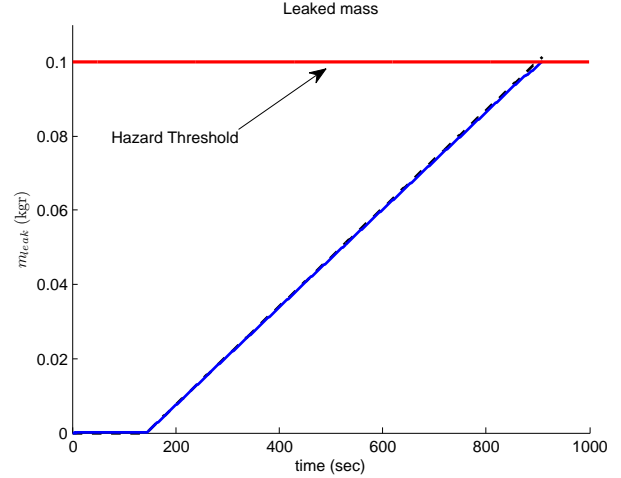


Figure 7: This figure illustrates the actual (dashed line) and the estimated (solid line) value of the leaked mass.

In many practical applications, the error that can be generated by considering the particle weights invariant for future time instants is negligible with respect to other sources, such as model inaccuracies or wrong assumptions about process/measurement noise parameters. Thus, from this standpoint, Eq. (15) is considered sufficient to extend the fault estimate trajectories, while the current particle weights are propagated in time without changes. The computational burden of this method is considerably reduced and, as it will be shown in simulation results, it can give a satisfactory view about how the system behaves in time for most practical applications.

The proposed fault diagnosis framework allows the use of the pdf estimates of the system continuous valued states (computed at the moment of fault detection) as initial conditions in the failure prognostic routine, giving excellent insight into the inherent uncertainty in the prediction problem. As a result, a swift transition between the two modules (fault diagnosis and failure prognosis) may be performed, and moreover, reliable prognosis can be achieved within a few cycles of operation after the fault is declared. This characteristic is, in fact, one of the main advantages offered by this particle-filter based framework.

6. SIMULATION RESULTS

The performance of the proposed FDI and prognostic algorithms was tested via numerical simulations. The evaporator dynamics are described by Eqs. (1)-(7). Regarding the inlet flow rates we set:

$$\dot{m}_{in} = \dot{m}_v - \dot{m}_{leak} \quad \text{and} \quad \dot{m}_{out} = \dot{m}_c$$

where \dot{m}_v and \dot{m}_c are the flow rates of the TEV and compressor, respectively. The leakage fault is seeded according to Eq. (9). The systems parameters are summarized in Table 1. The number of particles used for the two estimators (state and parameter) are $N = 100$ and $N_\theta = 150$. The crack constant is given by:

$$C_r(t) = 5 \cdot 10^{-9} \text{step}(t - 144) \quad (16)$$

Using the above representation we simulate the occurrence of an abrupt and unexpected crack that causes

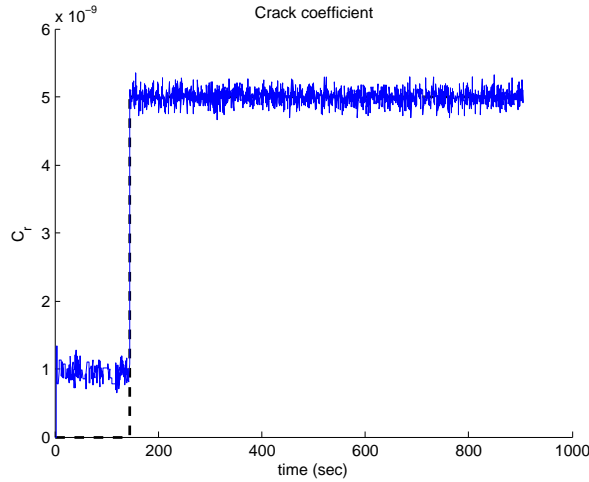


Figure 8: This figure illustrates the actual (dashed line) and the estimated (solid line) value of the crack coefficient.

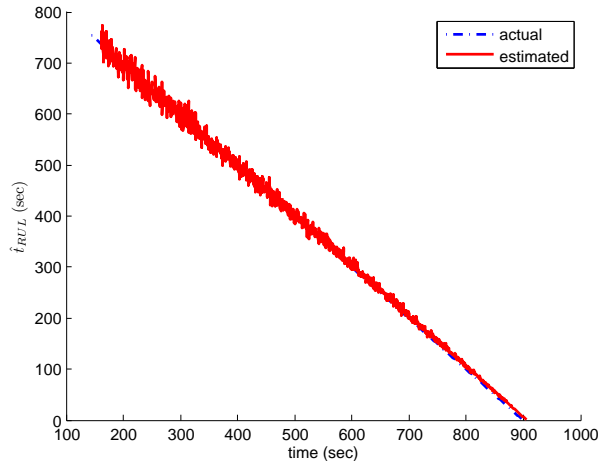


Figure 9: This figure illustrates the actual (dashed dotted line) and the estimated (solid line) value of RUL.

Table 1: Simulator parameters

$\bar{\gamma}$	0.8474	A_w	$6.6361 \cdot 10^{-5} m^2$
A_t	$3.14 \cdot 10^{-2} m^2$	$U_w A_w$	40.9962 W/K
U_w	592.9817 W/Km ²	\dot{m}_a	0.765 kgr/sec
D_t	0.02 m	c_a	$10^3 J/kg \cdot K$
V_e	$0.0057 m^3$	ρ_a	$1.1996 kgr/m^3$
c_w	$1.9552 \cdot 10^3 J/kg \cdot K$	A_a	$0.1518 m^2$
ρ_w	$7.8491 \cdot 10^3 kgr/m^3$	x	0

the nonlinear leakage growth model given in Eq. (9). Signal noise has been added to the available measurements. The saturated states are calculated based on the tables of R134a refrigerant. The operating conditions are $\dot{m}_v = \dot{m}_c = 0.0108 kgr/sec$ and $T_A = 26 C^\circ$. Besides detecting the faulty condition, it is desired to obtain some measure of the statistical confidence of the alarm signal. For this reason, two outputs will be extracted from the FDI module. The first output is the expectation of the Boolean state $x_{d,2}(t)$, which constitutes an estimate of the probability of fault. The second output is the statistical confidence needed to declare the fault via hypothesis testing (H_0 : ‘the evaporator is not leaking’ vs H_1 : ‘The evaporator is leaking’). The latter output needs another pdf to be considered as the baseline. In this case, a normal distribution $N(0, \sigma)$ is used to define this baseline data. This indicator is essentially equivalent to an estimate of type II error. Customer specifications are translated into acceptable margins for the type I and II errors in the detection routine.

The algorithm itself will indicate when the type II error (false negatives) has decreased to the desired level. Figure 6 shows two indicators that are simultaneously computed. The first indicator, depicted as a function of time, shows the probability of a determined failure mode, and it is based on the estimate of the Boolean state $x_{d,2}$. FDI alarms may be triggered whenever this indicator reaches a pre-determined threshold (in this case the threshold value is 0.9). If more information is needed, the type II detection error (second and third indicators, respectively) may be considered.

Figure 7 illustrates the actual and estimated leaked mass. Figure 8 illustrates the estimated crack coefficient. A small bias is evident in the crack coefficient estimate in the healthy condition. However, this bias has a very small value and a low probability of fault. Finally Figure 9 illustrates the actual RUL compared to \hat{t}_{RUL} that is estimated by the prognosis module. The results indicate that the enhanced FDI and prognostic algorithms provide very accurate estimates of the fault progression, the crack coefficient and the RUL estimate.

7. CONCLUSIONS

This paper is introducing an architecture for the development, implementation, testing and assessment of a particle-filtering-based framework for FDI and prognosis. The proposed framework for FDI has been successful and very efficient in pinpointing abnormal conditions in very complex and nonlinear processes, such as the detection of leakage in a two-phase evaporator of an ECS. The FDI algorithm is enhanced with an adaptive mod-

ule that provides estimates of the fault nonlinear model parameters. Regarding prognosis, it was shown that that the proposed approach is suitable for online implementation, providing acceptable results in terms of precision and accuracy. A successful case study has been presented, offering insights about how model inaccuracies and/or customer specifications (hazard zone or prediction window definitions) may affect the algorithm performance.

REFERENCES

- Braun, J. (2003). Automated Fault Detection and Diagnostics for Vapor Compression Cooling Equipment. *Transaction of the ASME*, 125, 266-274.
- Breuker, M., & Braun, J. (1998). Common faults and their impacts for rooftop air conditioners. *International Journal of HVAC&R Reserach*, 4, 303-318.
- Chen, C., Brown, D., Sconyers, C., Vachtsevanos, G., & Zhang, B. (2010). A .NET framework for an integrated fault diagnosis and failure prognosis architecture. In *IEEE AUTOTESTCON*.
- Chen, C., Vachtsevanos, G., & Orchard, M. (2010). Machine remaining useful life prediction based on adaptive neuro-fuzzy and high-order particle filtering. In *Annual Conference of the Prognostics and Health Management Society*.
- Chen, C., Zhang, B., & Vachtsevanos, G. (n.d.). *Prediction of machine health condition using neuro-fuzzy and Bayesian algorithms*. (To be published in IEEE Transactions on Instrumentation and Measurement)
- Cheng, T., He, X.-D., & Asada, H. (2004). Nonlinear observer design for two-phase flow heat exchangers of air conditioning systems. In *American Control Conference, 2004. Proceedings of the 2004* (Vol. 2, p. 1534 - 1539 vol.2).
- Comstock, M., Braun, J., & Groll, E. (2002). A survey of common faults for chillers. *ASHRAE Transactions*, 108, 819-825.
- Grald, E. W., & MacArthur, J. (1992). A moving-boundary formulation for modeling time-dependent two-phase flows. *International Journal of Heat and Fluid Flow*, 13(3), 266 - 272.
- He, X. (1996). *Dynamic Modeling and Multivariable Control of Vapor Compression Cycles in Air Conditioning Systems*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.
- He, X.-D., & Asada, H. (2003). A new feedback linearization approach to advanced control of multi-unit HVAC systems. In *American Control Conference, 2003. Proceedings of the 2003* (Vol. 3, p. 2311 - 2316 vol.3).
- Li, H., & Braun, J. (2003). An Improved Method for Fault Detection and Diagnosis Applied to Packaged Air Conditioners. *American Society of Heating, Refrigerating and Air Conditioning Engineers*, 109, 683-692.
- Merritt, H. (1967). *Hydraulic Control Systems*. John Wiley & Sons.
- Navarro-Esbri, J., Torrella, E., & Cabello, R. (2006). A vapour compression chiller fault detection technique based on adaptative algorithms. Application to on-line refrigerant leakage detection. *International Journal of Refrigeration*, 29, 716-723.
- Orchard, M., & Vachtsevanos, G. (2007). A particle filtering-based framework for real-time fault diagnosis and failure prognosis in a turbine engine. In *Control Automation, 2007. MED '07. Mediterranean Conference on* (p. 1 -6).
- Orchard, M., & Vachtsevanos, G. (2009). A particle-filtering approach for on-line fault diagnosis and failure prognosis. *Transactions of the Institute of Measurement and Control*, 31, 221-246.
- Rasmussen, B. (2005). *Dynamic Modeling and Advanced Control of Air Conditioning and Refrigeration Systems*. Unpublished doctoral dissertation, University of Illinois.
- Rossi, T., & Braun, J. (1997). A statistical, rule-based fault detection and diagnostic method for vapor compression air conditioners. *International Journal of HVAC&R Reserach*, 3, 19-37.
- Stylianou, M., & Nikanpour, D. (1996). Performance monitoring, fault detection, and diagnosis of reciprocating chillers. *ASHRAE Transactions*, 102, 615-627.
- Wedekind, G., Bhatt, B., & Beck, B. (1978). A System Mean void Fraction Model For Predicting Various Transient Phenomena Associated with Two-Phase Evaporating and Condensing Flows. *International journal of Multiphase Flow*, 4, 97-114.
- Ioannis A. Raptis** was born in Athens, Greece in 1979. He received his Dipl.-Ing. in Electrical and Computer Engineering from the Aristotle University of Thessaloniki, Greece and his Master of Science in Electrical and Computer Engineering from the Ohio State University in 2003 and 2006, respectively. From 2005 until 2006 he conducted research at the Locomotion and Biomechanics Laboratory of the Ohio State University. In 2006 he joined the Unmanned Systems Laboratory at the University of South Florida. In 2010 he received his Ph.D. degree in the department of Electrical Engineering at the University of South Florida. In 2010 he joined the Intelligent Control Systems Laboratory of the Georgia Institute of Technology. His research interests include nonlinear systems control theory, nonlinear control of electromechanical/robotic systems and rotorcraft/aircraft system identification and control.
- George J. Vachtsevanos** is a Professor Emeritus of Electrical and Computer Engineering at the Georgia Institute of Technology. He was awarded a B.E.E. degree from the City College of New York in 1962, a M.E.E. degree from New York University in 1963 and the Ph.D. degree in Electrical Engineering from the City University of New York in 1970. He directs the Intelligent Control Systems laboratory at Georgia Tech where faculty and students are conducting research in intelligent control, neurotechnology and cardiotechnology, fault diagnosis and prognosis of large-scale dynamical systems and control technologies for Unmanned Aerial Vehicles. Dr. Vachtsevanos was awarded the IEEE Control Systems Magazine Outstanding Paper Award for the years 2002-2003 (with L. Wills and B. Heck). He was also awarded the 2002-2003 Georgia Tech School of Electrical and Computer Engineering Distinguished Professor Award and the 2003-2004 Georgia Institute of Technology Outstanding Interdisciplinary Activities Award.

E2GK-pro: An Evidential Evolving Multimodeling Approach for Systems Behavior Prediction

Lisa Serir, Emmanuel Ramasso, Nouredine Zerhouni
FEMTO-ST Institute, UMR CNRS 6174 - UFC / ENSMM / UTBM,
Automatic Control and Micro-Mechatronic Systems Dep., 25000, Besançon, France
(lisa.serir, emmanuel.ramasso, noureddine.zerhouni)femto-st.fr

ABSTRACT

Nonlinear dynamic systems identification and nonlinear dynamic behavior prediction are important tasks in several areas of industrial applications. Multiple works proposed multimodel-based approaches to model nonlinear systems. Multimodeling permits to blend different model types together to form hybrid models. It advocates the use of existing, well known model types within the same model structure. Recently, a multimodeling strategy based on belief functions theory was developed based on a fuzzy rule based system. We propose a different approach of this latter taking advantage of new efficient evidential clustering algorithms for the determination of the local models and the assessment of the global model. In particular, we propose an online procedure based on the Evidential Evolving Gustafsson-Kessel (E2GK) algorithm that ensures an evolving partitioning of the data into clusters that correspond to operating regions of the global system. Thus the estimation of the local models is dynamically performed by upgrading and modifying their parameters while the data arrive. Each local model is weighted by a belief mass provided by E2GK, and the global model (multimodel) is a combination of all the local models.

1. INTRODUCTION

Dealing with nonlinear systems behavior identification and prediction is a widely encountered problem in real world applications in engineering, industry, time series analysis, prediction and fault diagnosis. Modeling an a priori unknown dynamic process from observed data is a hard task to perform. Among the large variety of proposed approaches taking into account nonlinearity, one can cite Fuzzy logic based models (Takagi & Sugeno, 1985) and especially neural network based approaches, which applications during the last decades are numerous in dynamical system modeling, and in particular in prognosis applications (El-Koujok, Gouriveau, & Zerhouni, 2011). Usually,

*This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

the models consist of a set of functional relationships between the elements of a set of variables. Multiple works propose multimodel-based approaches to avoid difficulties (modeling complexity) related to nonlinearity (P. Angelov, Lughofer, & Zhou, 2008; Madani, Rybnik, & Chebira, 2003).

Multimodeling permits to blend different model types together to form hybrid models, offering a unified view toward modeling with well known model types instead of promoting a singular model type which is insufficient to model large scale systems. In a general way, in multimodel-based approaches, a set of models, corresponding to a set of operating ranges of the system, contributes to identify the whole system. Such an approach can be seen as a weighted contribution of a set of models approximating the whole system's behavior, each of which is valid in a well defined interval which corresponds to operating region of the system or covers a part of the whole feature space of the problem to be solved. The description of the global system's behavior is made by combination of all the local models. The contribution of each local model in the assessment of the multimodel's output is quantified by an *activation degree*.

One of the most popular models is the TSK fuzzy model that showed great performances in many applications on prediction (El-Koujok et al., 2011). A first order Takagi-Sugeno model can be seen as a multimodel structure consisting of linear models. It is based on a fuzzy decomposition of the input space to describe the inherent structure for a concrete problem by partitioning each input variable range into fuzzy sets. For each part of the state space, a fuzzy rule can be constructed to make a linear approximation of the input, and the global output is a combination of all the rules. Then, the parameters of the models (non-linear parameters of membership degrees and linear parameters for the consequent of each rule) are tuned in an appropriate learning procedure. Usually, the identification of the linear parameters is addressed by some gradient descent variant, e.g., the least squares algorithm, whereas non-linear parameters are determined by some clustering method on the input space. This kind of approach has been applied to build a Neuro-Fuzzy predictor in the context of prognosis application by (El-Koujok et al., 2011).

It was based on the evolving extended Takagi-Sugeno system (exTS) proposed by Angelov (P. P. Angelov & Filev, 2004).

Recently, a multimodeling strategy based on belief functions theory was developed based on a TSK fuzzy model (Ramdani, Mourot, & Ragot, 2005). The basic idea was to consider a fuzzy rule based system with a belief structure as output. The focal elements of each rule were formed by a subset of a collection of functional models each of which was constructed based on a fuzzy model of Takagi-Sugeno type. In this paper we investigate this method and we introduce some modification taking advantage of new efficient evidential clustering algorithms for the determination of the local models and the assessment of the global model. In particular, we propose an online procedure using the Evidential Evolving Gustafsson-Kessel (E2GK) (Serir, Ramasso, & Zerhouni, 2011) algorithm that ensures an evolving partitioning of the data into clusters that correspond to operating regions of the global system. Thus the estimation of the local models is dynamically performed by upgrading and modifying their parameters while the data arrive. Each local model is weighted by a belief mass provided by E2GK, and the global model (multimodel) is a combination of all the local models.

The paper is organized as follows: Section 2 is dedicated to the necessary background for our approach. In section 3, the existing approach will be first presented (Section 3.1), analyzed (Section 3.2) in order to introduce the proposed model (Section 3.3). Results will finally be presented in Section 4.

2. BACKGROUND

A brief description of belief functions is first given. Then, ECM algorithm is presented followed by E2GK algorithm as the basis of the prediction algorithm.

2.1 Belief Functions

Dempster-Shafer theory of evidence, also called belief functions theory, is a theoretical framework for reasoning with partial and unreliable information. Ph. Smets proposed the *Transferable Belief Model* (TBM) (Smets & Kennes, 1994) as a general framework for uncertainty representation and combination of various pieces of information without additional priors. In particular, TBM offers the possibility to explicitly emphasize doubt, that represents ignorance, and conflict, that emphasizes the contradiction within a fusion process. We give here some of the basic notions of the theory and refer the reader to (Smets & Kennes, 1994) for a more complete description.

The central notion of the theory of belief functions is the *basic belief assignment* (BBA), also called *belief mass assignment* that represents the *belief* of an agent in subsets of a finite set Ω , called the frame of discernment. It is defined by:

$$m : 2^\Omega \rightarrow [0, 1] \quad (1)$$

$$A \mapsto m(A),$$

with $\sum_{A \subseteq \Omega} m(A) = 1$. A belief mass can not only be assigned to a singleton ($|A| = 1$), but also to a *subset* ($|A| > 1$) of variables *without assumption concerning additivity*. This property permits the explicit modelling of

doubt and conflict, and constitutes a fundamental difference with probability theory. The subsets A of Ω such that $m(A) > 0$, are called the *focal elements* of m . Each focal element A is a set of possible values of ω . The quantity $m(A)$ represents a fraction of a unit mass of belief allocated to A . Complete ignorance corresponds to $m(\Omega) = 1$, whereas perfect knowledge of the value of ω is represented by the allocation of the whole mass of belief to a unique singleton of Ω , and m is then said to be *certain*. In the case of all focal elements being singletons, m boils down to a probability function and is said to be *bayesian*.

A positive value of $m(\emptyset)$ is considered if one accepts the *open-world assumption* stating that the set Ω might not be complete, and thus ω might take its value outside Ω . The conflict is then interpreted as a mass of belief given to the hypothesis that ω might not lie in Ω . This interpretation is useful in clustering for outliers detection (Masson & Denoeux, 2008).

2.2 Evidential C-Means

In 2008, Masson and Denoeux (Masson & Denoeux, 2008) proposed a clustering algorithm based on the concept of *Credal Partition* (Masson & Denoeux, 2008). Similar to the concept of fuzzy partition, but more general, it particularly permits a better interpretation of the data structure and makes it possible to code all situations, from certainty to total ignorance. Considering a set of N data x_1, \dots, x_n to be grouped in c clusters, a credal partition is constructed by assigning a BBA to each possible subset of clusters. Partial knowledge regarding the membership of an data point i to a class j is represented by a BBA m_{ij} on the set $\Omega = \{\omega_1, \dots, \omega_c\}$. ECM is an optimization based clustering algorithm whose objective function is given by:

$$J_{ECM}(M, V) = \sum_{i=1}^N \sum_{\{j/A_j \neq \emptyset, A_j \subseteq \Omega\}} |A_j|^\alpha m_{ij}^\beta d_{ij}^2 + \sum_{i=1}^N \delta^2 m_i(\emptyset)^\beta \quad (2)$$

subject to

$$\sum_{\{j/A_j \neq \emptyset, A_j \subseteq \Omega\}} m_{ij} + m_i(\emptyset) = 1 \quad \forall i = 1, \dots, N, \quad (3)$$

where:

- α is used to penalize the subsets of Ω with high cardinality,
- $\beta > 1$ is a weighting exponent that controls the fuzziness of the partition,
- d_{ij} denotes the Euclidean distance between object i and prototype v_j ,
- δ controls the amount of data considered as outliers.

The $N \times 2^c$ partition matrix M is derived by determining, for each object i , the BBAs $m_{ij} = m_i(A_j)$, $A_j \subseteq \Omega$ such that m_{ij} is low (resp. high) when the distance d_{ij} between data i and focal element A_j is high (resp. low). The matrix M is computed by the minimization of criterion (2) and was shown to be (Masson & Denoeux, 2008), $\forall i = 1 \dots n$,

$\forall j/A_j \subseteq \Omega, A_j \neq \emptyset$:

$$m_{ij} = \frac{|A_j|^{-\alpha/(\beta-1)} d_{ij}^{-2/(\beta-1)}}{\sum_{A_k \neq \emptyset} |A_k|^{-\alpha/(\beta-1)} d_{ik}^{-2/(\beta-1)} + \delta^{-2/(\beta-1)}} , \quad (4)$$

and $m_i(\emptyset) = 1 - \sum_{A_j \neq \emptyset} m_{ij}$. Centers of clusters are optimized by minimizing criterion (2). The distance between an object and any non empty subset $A_j \subseteq \Omega$ is then defined by computing the center of each subset A_j . This latter is the barycenter \bar{v}_j of the centers of clusters composing A_j .

From the credal partition, the classical clustering structures (possibilistic, fuzzy and hard partitions) can be recovered (Masson & Denoeux, 2008). One can also summarize the data by assigning each object to the set of clusters with the highest mass. One then obtains a partition of the points in at most 2^c groups, where each group corresponds to a set of clusters. This makes it possible to find the points that unambiguously belong to one cluster, and the points that lie at the boundary of two or more clusters. Moreover, points with high mass on the empty set may optionally be rejected as outliers.

2.3 E2GK: Evidential Evolving Gustafson-Kessel Algorithm

In (Serir et al., 2011), an online clustering method, the evidential evolving Gustafson-Kessel algorithm (E2GK), was introduced in the theoretical framework of belief functions. The algorithm enables an online partitioning of data streams based on two existing and efficient algorithms: Evidential c -Means (ECM) and Evolving Gustafson-Kessel (EGK) (Georgieva & Filev, 2009). E2GK makes it possible to compute, online, a credal partition as data gradually arrive. We summarize in the following the main steps of the algorithm:

Step 1 – Initialization: At least one cluster center should be provided. Otherwise, the first point is chosen as the first prototype. If more than one prototype is assumed in the initial data, the Gustafsson-Kessel (Gustafson & Kessel, 1978) or ECM algorithm can be applied to identify an initial partition matrix. The result of the initialization phase is a set of c prototypes v_i and covariance matrices F_i .

Step 2 – Decision making: The boundary of each cluster is defined by the cluster radius r_i , defined as the median distance between the cluster center v_i and the points belonging to this cluster with membership degree larger or equal to a given threshold u_h :

$$r_i = \text{median}_{\forall x_j \in i\text{-th cluster and } P_{ij} > u_h} \|v_i - x_j\|_{A_i} . \quad (5)$$

where P_{ij} is the confidence degree that point j belongs to $\omega_i \in \Omega$ and can be obtained by three main process: either by using the belief mass $m_j(\omega_i)$, or the pignistic transformation (Smets & Kennes, 1994) that converts a BBA into a probability distribution, or by using the plausibility transform (Cobb & Shenoy, 2006). We propose to choose the belief mass for which the computation is faster.

The minimum membership degree u_h - initially introduced in (Georgieva & Filev, 2009) and required to decide whether a data point belongs or not to a cluster - can be difficult to assess. It may depend on the density of the data

as well as on the level of cluster overlapping. Thus u_h is automatically set to $1/c$ in order to reduce the number of parameters while ensuring a natural choice for its value.

Step 3 – Computing the partition matrix: Starting from the resulting set of clusters at a given iteration, the partition matrix M is built as in ECM. The Mahalanobis-like distance d_{ik} is considered assuming that each cluster volume ρ_i is one as in standard GK algorithm:

$$d_{ik}^2 = \|x_k - v_i\|_{A_i}^2 = (x_k - v_i)A_i(x_k - v_i)^T , \quad (6a)$$

$$A_i = [\rho_i \cdot \det(F_i)]^{1/n} F_i^{-1} , \quad (6b)$$

$$F_i = \frac{\sum_{k=1}^N (m_{ik})^\beta (x_k - v_i)^T (x_k - v_i)}{\sum_{k=1}^N (m_{ik})^\beta} . \quad (6c)$$

where F_i is the fuzzy covariance matrix.

Storing the whole partition is not efficient. Indeed, only the belief masses on singletons need to be stored in order to make the decision concerning the radius. As shown in Eq. 4, values on singletons are easy to compute but the problem is to estimate the normalization factor. To overcome this problem, all values of masses have to be computed but not stored. This little trick exponentially decreases memory consumption.

Step 4 – Adapting the structure: Given a new data point x_k , two cases are considered:

- *Case 1: x_k belongs to an existing cluster, thus a clusters' update has to be performed.* Data point x_k is assigned to the closest cluster p if the distance d_{pk} is less or equal to the radius r_p . Then, an update of the p -th cluster has to be performed as follows:

$$v_{p,new} = v_{p,old} + \theta \cdot \Delta , \quad (7)$$

where

$$\Delta = x_k - v_{p,old} , \quad (8)$$

and

$$F_{p,new} = F_{p,old} + \theta \cdot (\Delta^T \Delta - F_{p,old}) , \quad (9)$$

where θ is a learning rate (and can be set in $[0.05, 0.3]$), $v_{p,new}$ and $v_{p,old}$ denote respectively the new and old values of the center, and $F_{p,new}$ and $F_{p,old}$ denote the new and old values of the covariance matrix.

- *Case 2: x_k is not within the boundary of any existing cluster (i.e. $d_{pk} > r_p$), thus a new cluster may be defined and a clusters' update has to be performed.* The number of clusters is thus incremented: $c = c + 1$. Then, the incoming data x_k is accepted as a center v_{new} of the new cluster and its covariance matrix F_{new} is initialized with the covariance matrix of the closest cluster $F_{p,old}$. In order to quantify the credibility of the estimated clusters, a parameter P_i has been introduced in (Georgieva & Filev, 2009) to assess the number of points belonging to the i -th cluster. The authors suggested a threshold parameter P_{tol} to guarantee the validity of the covariance matrices and to improve the robustness. This parameter corresponds to the desired minimal amount of points falling within the boundary of each cluster. The threshold value is context determined due to the specificity of the considered data set. The new created cluster is then rejected if it contains less than P_{tol} data points.

After creating a new cluster, the data structure evolves. However, the new cluster may contain data points previously assigned to another cluster. Thus, the number of data points in previous clusters could change. After the creation of a new cluster, E2GK verifies that all clusters have at least the required minimum amount of data points (P_{tol} or more). If clusters don't satisfy this condition, the cluster with the minimum number of points is removed.

The overall algorithm is presented in Alg. 1.

Algorithm 1 E2GK algorithm

- 1: **Initialization:** Take the first point as a center or apply the off-line GK algorithm to get the initial number of clusters c and the corresponding matrices V and F_i , $i = 1 \dots c$
 - 2: **Calculate** \bar{v}_j , the barycenter of the cluster centers composing $A_j \subseteq \Omega$
 - 3: **Calculate the credal partition** M using Eq. 4 (store only singletons and normalize)
 - 4: **for all** new data point x_k **do**
 - 5: Find the closest cluster p
 - 6: Calculate the radius r_p of the closest cluster (Eq. 5)
 - 7: **if** $d_{pk} \leq r_p$ **then**
 - 8: Update the center v_p (Eq. 7)
 - 9: Update the covariance matrix F_p (Eq. 9)
 - 10: **else**
 - 11: Create a new cluster: $v_{c+1} := x_k$ and $F_{c+1} := F_p$
 - 12: Keep it if the number of points in this cluster is $\geq P_{tol}$
 - 13: **end if**
 - 14: **Recalculate** the credal partition M
 - 15: **Check the new structure:** estimate the number of points within each cluster and remove one cluster for which the latter is $\leq P_{tol}$
 - 16: **end for**
-

3. MODELING DYNAMICS

In this section, the existing approach is first presented (Section 3.1) and then analyzed (Section 3.2). Finally we introduce the proposed model based on belief functions (Section 3.3).

3.1 The existing approach

In (Ramdani et al., 2005), a multi-modeling strategy based on belief function theory was developed for modeling complex nonlinear mappings by combination of simpler functional models. It was based on the TSK fuzzy model. The basic idea was to consider a fuzzy rule based system with a belief structure as output. The focal elements of each rule were formed by a subset of a collection of functional models. Each functional model is constructed based on a fuzzy model of Takagi-Sugeno type in two steps: structure identification and parameters estimation. In the first step, the antecedent and consequent variables of the model are determined. From the available training data that contain input-output samples, a regression matrix and an output vector are constructed. In the second step, the number of rules K , the antecedent fuzzy sets, and the parameters of the rule consequents are identified. The system behaviour is approximated by local linear models of the different

operating regions that are represented by clusters. The Gustafsson-Kessel fuzzy clustering algorithm (Gustafsson & Kessel, 1978) is applied on the product-space of input and output variables to discover the potential regions of the rules and capture the interaction between the input and output variables. Thus, a certain number c of functional relationships between input and output variables, denoted by $f^j(x)$, $j = 1, \dots, c$, are assumed and form the frame of discernment Ω :

$$\Omega = \{ \{f^1\}, \dots, \{f^c\} \} , \quad (10)$$

where $\{f^j\}$ is the hypothesis that corresponds to the functional model $f^j(x)$. The authors consider the case where the number of input prototypes (or rules) is equal to the number of functional prototypes ($K = c$). In order to predict an output value y for a given input vector x , each of the K rules (determined in the second step (Ramdani et al., 2005)) provides a piece of evidence concerning the value of the unknown output y , which can be represented by a belief mass m^i , $i = 1, \dots, K$:

$$\begin{cases} m^i(\{f^j\}|x) &= \phi_i(x), j = 1, \dots, J(i) \\ m^i(\Omega|x) &= 1 - \phi_i(x) \\ m^i(A|x) &= 0 \quad \forall A \in F^\Omega - F^i \end{cases} \quad (11)$$

where F^Ω is the power set of Ω , F^i are the focal sets of m^i , and The function $\phi_i(x)$ is related to the input domain (domain of expertise) of the i th rule. We refer the reader to (Ramdani et al., 2005) for more details. This method of constructing belief masses is based on a method proposed by T. Denoeux (Denoeux, 2000) in the context of classification.

In order to make a decision, the outputs of the different rules which are belief structures, are combined using the Dempsters rule of combination giving the overall belief structure m , which is a vector of $c + 1$ elements:

$$m = \bigoplus_{i=1}^K m^i, \quad (12)$$

It is then normalized providing a belief structure: $m_j^* = \frac{m_j}{\sum_{q=1}^{c+1} m_q}$ $j = 1, \dots, c + 1$.

The overall multimodel is then defined as a combination of the functional prototypes with an additional model representing the frame of discernment, denoted by f^Ω :

$$\hat{y} = \sum_{i=1}^c m^* (\{f^i\}) f^i(x) + m^*(\Omega) f^\Omega(x) . \quad (13)$$

Here, the authors associate the mass of total ignorance to a general model $f^\Omega(x)$, which is a convex combination of local linear functions whose parameters are identified globally by a single least squares equation (Eq.18, (Ramdani et al., 2005)). This formulation emphasizes the doubt concerning the model. On the other hand, the linear models $f^i(x)$ are identified by the weighted least squares (Eq.19 in (Ramdani et al., 2005)).

3.2 Analysis of the existing approach

Problem 1 – Determining the belief masses: In (Ramdani et al., 2005), the approach relies on fuzzy modeling using belief functions based on two existing approaches. The first

approach was proposed by Yager (Yager & Filev, 1995) in the context of fuzzy modeling. This strategy allows the integration of probabilistic uncertainty in fuzzy rule based systems. The output of the rules is a belief structure whose focal elements are fuzzy sets among the output variable linguistic terms. The second approach is the evidential k-nearest neighbours proposed by Denoeux (Denoeux, 2000) in the context of classification and later applied in regression analysis (Petit-Renaud & Denoeux, 1999). For a given input query vector, the output variable is obtained in the form of a fuzzy belief assignment (FBA), defined as a collection of fuzzy sets of values with associated masses of belief. In (Petit-Renaud & Denoeux, 1999), the output FBA is computed nonparametrically on the basis of the training samples in the neighbourhood of the query point. In this approach, the underlying principle is that the neighbours of the query point are considered as sources of partial information on the response variable; the bodies of evidence are discounted as a function of their distance to the query point, and pooled using the Dempster's rule of combination.

Contribution 1: We propose a simpler and more efficient method than the one described in section 3.1 to generate the masses of belief directly from the data at the clustering step. Indeed, in 2004, the authors of (Ramdani et al., 2005) couldn't yet benefit from new efficient clustering algorithms exclusively based on belief functions. In 2008, the first clustering algorithm, the Evidential c-Means algorithm (ECM) (described in section 2.2), based on belief functions was proposed by M-H. Masson and T. Denoeux (Masson & Denoeux, 2008). In the approach proposed by (Ramdani et al., 2005), applying ECM to the set of learning data would directly provide the BBA.

Problem 2 – Modeling doubt regarding the global model: In (Ramdani et al., 2005), the authors define the overall model as a combination of the functional prototypes with a single model representing the frame of discernment denoted by $f^\Omega(x)$. This particular model is associated to the mass of total ignorance (Eq.13). Doing so, the authors claimed to emphasize the doubt concerning the global model. We believe that the global model $f^\Omega(x)$ as proposed in (Ramdani et al., 2005), which is a convex combination of local linear functions, doesn't bring significant additional information to the model. Indeed, it is very similar to the local linear models as shown in their experiments.

Contribution 2: ECM assigns masses of belief to singletons but also to unions of clusters representing doubt regarding the general model. As a unit mass is distributed among all possible subsets of Ω , the masses on singletons are computed taking into account the doubt regarding the global model. Thus, we propose a different formulation of the overall multimodel, where we no longer have to combine the functional prototypes with a model representing the frame of discernment:

$$\hat{y} = \sum_{i=1}^c m^* (\{f^i\}) f^i(x) . \quad (14)$$

Contrary to the original approach where the doubt concerning the global model is emphasized by taking into account an additional model representing the frame of discernment, we develop an approach where doubt is emphasized directly based on E2GK.

Problem 3 – Evolving Modeling: The previous approach is suitable for a fixed set of data supplied in batch mode and under the assumption that the model structure remains unchanged. When the training data are collected continuously, some of them will reinforce and confirm the information contained in the previous data, while others could bring new information. This new information could concern a change in operating conditions, development of a fault or simply more significant change in the dynamic of the process. They may provide enough new information to form a new local model or to modify or even delete an existing one. Thus an adaptation of the model structure is necessary. To do so, an on-line clustering of the input-output data space with gradually evolving regions of interest should be used.

Contribution 3: We propose to use the recently proposed online clustering method (Serir et al., 2011) E2GK (Evidential Evolving Gustafson-Kessel) that enables online partitioning of data streams and adapts the clusters' parameters along time. As presented in section 2.3, E2GK uses the concept of credal partition of ECM, offering a better interpretation of the data structure. The resulting BBAs can then be used in Eq.14.

3.3 The proposed model (E2GK-pro)

Based on the same general idea, we propose to construct a model for approximating nonlinear functional mappings. As in (Ramdani et al., 2005), the system behaviour is approximated by local linear models of the different operating regions that are represented by clusters.

Compared to the original approach, we propose the following methodology :

1. Use the online evidential clustering algorithm E2GK that is capable of generating the belief masses and adapt the clusters' parameters along time;
2. For each cluster discovered by E2GK, construct a linear local model and update with the new incoming data;
3. Predict the new output \hat{y} by the linear combination of the local models as in Eq.14.

Initialization: The first data point x_1 is chosen as the first prototype. At the moment not enough data are available to construct the first model. As discussed in section (2.3), a new cluster is created if it contains at least P_{tol} data points. We will consider the same threshold for the necessary amount of data to construct a new model. Basically, the initialization step is the same as in E2GK.

Adapting the structure: At each new incoming data point x_k , an update of the clusters' parameters is performed by E2GK. Either x_k belongs to an existing cluster, thus a clusters' update has to be performed. Or, x_k is not within the boundary of any existing cluster, thus a new cluster may be defined and a clusters' update has to be performed. After the creation of a new cluster, E2GK verifies if all clusters have at least the required minimum amount of data points (P_{tol} or more) and suppresses the clusters that fail to satisfy this condition. To each cluster i corresponds a local model f^i such that:

$$f^i(x) = \theta_{i0} + \theta_{i1}x_1 + \dots + \theta_{ir}x_r \quad (15)$$

where $x = [x_1, \dots, x_r]^T$ is the vector of data composing the i th cluster and $\theta_i^T = [\theta_{i0}, \theta_{i1}, \dots, \theta_{ir}]$ is the vector of the parameters of f^i .

Then either a simple or a weighted recursive least squares estimation (RLS) (P. Angelov et al., 2008) could be used to identify the parameters of these linear sub-models.

Three cases are considered: 1) *a new cluster is created*, then the partition matrix changes and a learning step has to be performed for both the new local model and the previously generated local models; 2) *a cluster is removed*, then the partition matrix also changes and a new learning step has to be performed to update the existing local models; 3) Nothing happens.

Predicting the new output: Once the parameters of the local models are identified, the new output (prediction at $t + 1$) is estimated by Eq. 14.

We summarize the general approach in Alg. 2:

Algorithm 2 General Approach

Require: x_k a new data point and E2GK parameters

Ensure: \hat{x}^{k+1} and E2GK parameters update

- 1: **if** a new prototype is created **then**
 - 2: Add a new model
 - 3: Estimate parameters of the new model
 - 4: Update parameters of existing models
 - 5: **end if**
 - 6: Predict the new output x_{k+1} (Eq.14)
-

4. EXPERIMENT

The proposed EG2K-pro algorithm is designed for the prediction at $t + 1$. Further predictions requires other developments which are under study. So we are in the same case as in (Ramdani et al., 2005) where we assess the algorithm for the prediction of signals at $t + 1$.

Experiments were conducted on three applications:

- the 1-D Mackey-Glass chaotic time series,
- a multidimensional case: the PHM 2008 challenge data,
- a multidimensional case: the PRONOSTIA platform.

4.1 A benchmark 1-D problem

As a first example of application, we consider the Mackey-Glass chaotic time series:

$$x(t) = \frac{a \cdot x(t - \tau)}{1 + x^{10}(t - \tau)} - b \cdot x(t), \quad (16)$$

with $a = 0.3, b = 0.1, \tau = 20, x_0 = 1.2$ and 100 points.

E2GK parameters were set to $\delta = 10, \alpha = 1, \beta = 2, \theta = 0.01$ and $P_{\text{tol}} = 10$, and inputs were composed of $[t \ x(t-2) \ x(t-1) \ x(t)]$.

Figure 1 depicts the prediction at $t+1$, with a mean-squared error (MSE) of 2.10^{-2} .

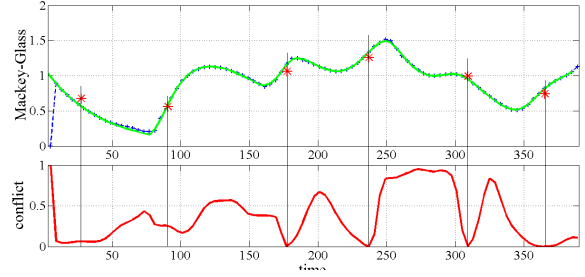


Figure 1. Top: Predictions (continuous bold line) and segmentation (stars). Bottom: Conflict evolution.

One interesting feature of the proposed algorithm is the degree of conflict (Fig. 1). As expected the degree of conflict is low around prototypes and increases when data points are far from the latter. The most interesting observation concerns the increasing of conflict in non-linear parts. For example, around $t = 200$ and $t = 275$, the increasing is much more important than in interval $[1, 150]$.

Let consider the maximum degree of belief generated by the clustering phase at instant (data-point) t :

$$s_t = \max_k m_{tk} \quad (17)$$

If the maximum is low, then the confidence in the prediction should be also low. This quantity is illustrated in Figures 2 (E2GK) and 3 (EGK).

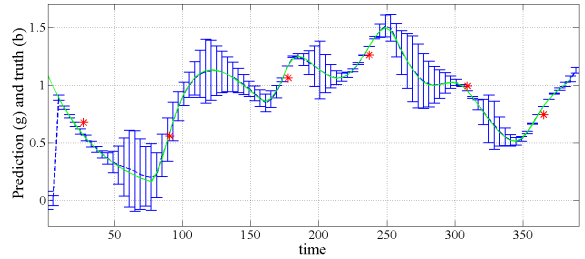


Figure 2. Illustration of $1 - s_t$ (Eq. 17) to quantify uncertainty around predictions by E2GK. These uncertainties appear on the figure with error bars around predictions (continuous line).

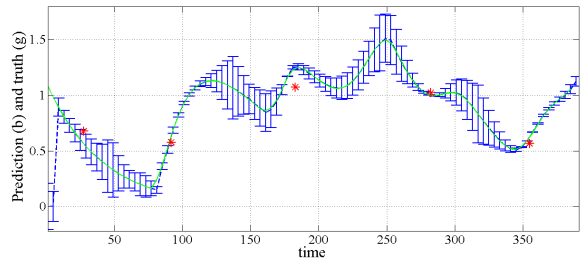


Figure 3. Illustration of $1 - s_t$ (Eq. 17) to quantify uncertainty around predictions by EGK. These uncertainties appear on the figure with error bars around predictions (continuous line).

At each instant, this value appears as an error bar. In both methods (EGK and E2GK), the maximum degree is close to 1 when points are located near clusters and therefore the values of $1 - s_t$ are close to 0. The main difference is that values for E2GK present more contrast than the ones for EGK. For example, in $[175, 275]$, the error is almost constant for EGK while in E2GK it increases as the distance to clusters increase. Moreover, for E2GK, high values are generally encountered in non-linearities that is not the case for EGK. In these areas, the value of conflict is generally high because points can belong to several clusters. These two figures also show the predictions (continuous line).

4.2 A multi-dimensional case: the PHM 2008 challenge data

We considered the challenge dataset concerning diagnostic and prognostics of machine faults from the first Int. Conf. on Prognostics and Health Management (2008) (Saxena, Goebel, Simon, & Eklund, 2008). The dataset is a multiple multivariate time-series (26 variables) with sensor noise. Each time series was from a different engine of the same fleet and each engine started with different degrees of initial wear and manufacturing variation unknown to the user and considered normal. The engine was operating normally at the start and developed a fault at some point. The fault grew in magnitude until system failure. The first experiment (*train_FD001.txt*) with five preselected features (3, 4, 5, 7, 9) was considered.

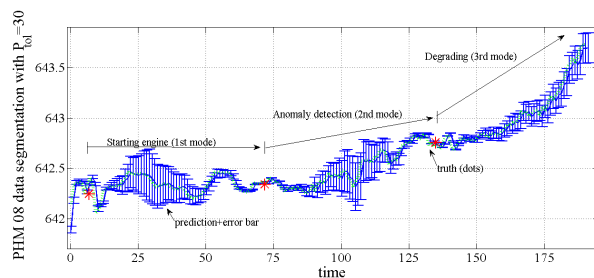


Figure 4. Segmentation (stars) by E2GK with $P_{\text{tot}} = 30$ obtained on PHM 2008 data, and prediction (continuous line) with error bars representing the opposite of the degree of support in each model. The real data appear with dots.

The automatic segmentation obtained by E2GK is given in Figure 4. This figure also depicts the prediction (with MSE equal to $1.6 \cdot 10^{-4}$) and error bars representing the difference between 1 and the maximum degree of belief for each data point (Eq. 17). E2GK parameters were the same as in the previous section, except $P_{\text{tot}} = 30$. In comparison, Figure 5 is the result for $P_{\text{tot}} = 10$ where the segmentation is finer as expected.

4.3 A multi-dimensional case: the PRONOSTIA platform

Description of PRONOSTIA

PRONOSTIA is an experimentation platform (Figure 6) dedicated to the test and validation of the machinery prognosis approaches, focusing on bearing prognostics. It was

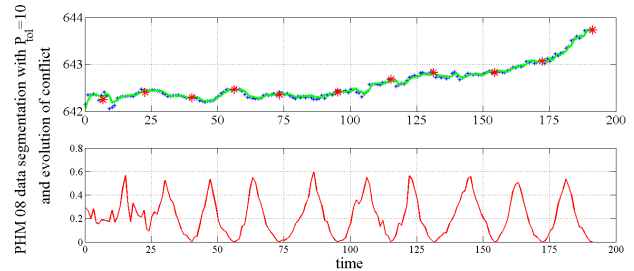


Figure 5. Top: Segmentation (stars) by E2GK with $P_{\text{tot}} = 10$ obtained on PHM 2008 data (to be compared to Figure 4) and prediction (continuous line). Bottom: The opposite of the maximum degree of belief (the lower the conflict the higher the confidence in predictions)

developed at FEMTO-ST institute (“Franche-Comté Electronics, Mechanics, Thermal Processing, Optics - Science and Technology”) in particular in AS2M department (Automatic control and Micro-Mechatronic Systems).

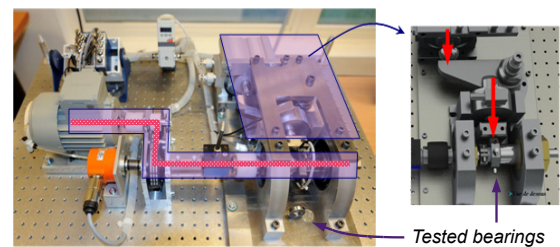


Figure 6. PRONOSTIA platform.

The main objective of PRONOSTIA is to provide real experimental data that characterise the degradation of a ball bearing along its whole operational life (until fault/failure). The collected data are vibration and temperature measurements of the rolling bearing during its functioning mode.

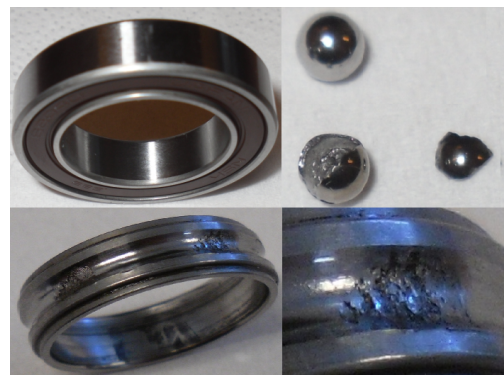


Figure 7. Bearing degradation.

The internal bearing ring is put in rotation, while the external bearing ring is maintained fixed. A radial load is applied on the external bearing ring in order to simulate its functioning. To speed up the degradation, the load exceeds the maximal load recommended by the supplier. The

originality of this experimental platform lies not only in the conjunction of the characterization of both the bearing functioning (speed, torque and radial force) and its degradation (vibrations and temperature), but also in the possibilities, offered by the platform, to make the operating conditions of the bearing vary during its useful life. Figure 7 depicts a bearing before and after the experiment.

The bearing operating conditions are determined by instantaneous measures of the radial force applied on the bearing, the rotation speed of the shaft handling the bearing, and of the torque inflicted on the bearing. During a test, the rolling bearing starts from its nominal mode until the fault state. The bearing behavior is measured using different types of sensors (Figure 8) such as miniaturized acceleration sensors and temperature probe.

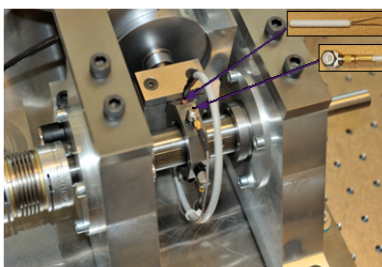


Figure 8. Sensors for degradation measurement.

The raw signals provided by the sensors are processed in order to extract relevant information concerning bearings states. Several techniques have been implemented and gathered in a signal processing toolbox with Matlab (Fig. 9): time-domain methods (RMS, skewness and kurtosis, crest factor, K-factor, Peak-to-Peak), frequency-domain methods (spectral and cepstrum analysis, envelope detection), time-frequency domain (short-time Fourier transform) and wavelets (discrete transform).

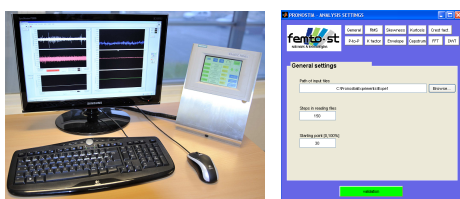


Figure 9. (left) Labview VI for raw signal visualization and (right) the graphical user interface to set the optional parameters (if required) of the signal processing algorithms.

Application of E2GK-pro on PRONOSTIA

E2GK parameters were the same as in the first section, except $P_{\text{tol}} = 20$. The prediction results are given in Figure 10 with a MSE equal to 6.10^{-5} .

The obtained segmentation is provided in Figure 11.

In comparison, Figure 12 is the result of segmentation for $P_{\text{tol}} = 10$. For this value, EGK was not able to provide a segmentation. As expected, the number of clusters is

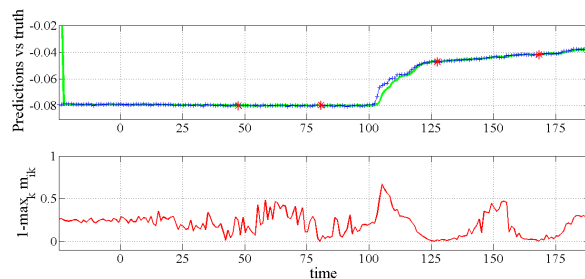


Figure 10. Top: The predictions for PRONOSTIA's data (continuous line). Bottom: The opposite of the maximum degree of belief.

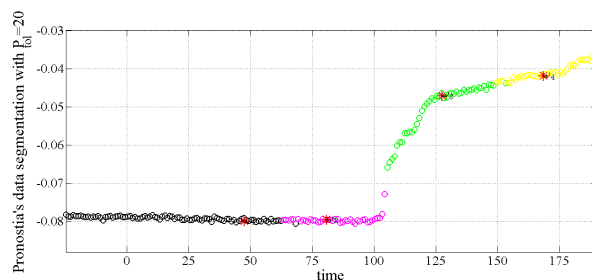


Figure 11. Segmentation (stars) by E2GK with $P_{\text{tol}} = 20$ obtained for a bearing in PRONOSTIA's platform.

greater for this latter value and “over”-segmentation appears mainly in areas with changes.

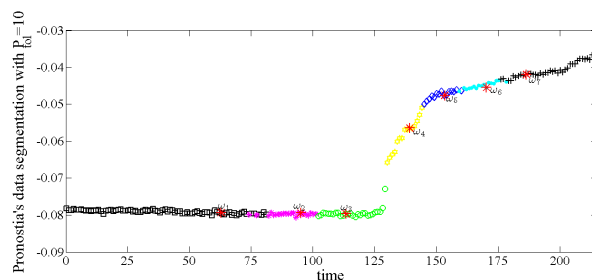


Figure 12. Segmentation (stars) by E2GK with $P_{\text{tol}} = 10$ for a bearing in PRONOSTIA's platform.

5. CONCLUSION

E2GK-pro is an evidential approach proposed for detecting, adapting and combining local models in order to analyse complex systems behavior. The approach relies on three main processes: 1) an online clustering called E2GK that generates belief functions and adapts its structure gradually, 2) the creation, adaptation or removing of models which are locally computed for each cluster, and 3) prediction of the future evolution.

Experiments were done on three datasets: one simulated and two real-world problems, in particular the PRONOSTIA platform. Results demonstrate the ability of the proposed method for online segmentation of multi-dimensional time-series and to build provide predictions

for the next iteration. We also proposed a confidence value attached to predictions.

Future work is mainly focused on the validation of the proposed methodology for long term prediction and to its comparison to Angelov's methodology (P. Angelov et al., 2008).

REFERENCES

- Angelov, P., Lughofer, E., & Zhou, X. (2008). Evolving fuzzy classifiers using different model architectures. *Fuzzy Sets and Systems*, 3160-3182.
- Angelov, P. P., & Filev, D. P. (2004). An approach to online identification of Takagi-Sugeno fuzzy models. *IEEE Trans Syst Man Cybern B Cybern*, 34, 484-98.
- Cobb, B. R., & Shenoy, P. P. (2006). On the plausibility transformation method for translating belief function models to probability models. *International Journal of Approximate Reasoning*, 41, 314-330.
- Denoeux, T. (2000). A Neural network classifier based on Dempster-Shafer theory. *IEEE Trans. Syst., Man, Cybern*, 30, 131-150.
- El-Koujok, M., Gouriveau, R., & Zerhouni, N. (2011). Reducing arbitrary choices in model building for prognostics: An approach by applying parsimony principle on an evolving neuro-fuzzy system. *Microelectronics Reliability*, 51, 310-330.
- Georgieva, O., & Filev, D. (2009). Gustafson-Kessel Algorithm for Evolving Data Stream Clustering. In *International Conference on Computer Systems and Technologies - CompSysTech 09*.
- Gustafson, E., & Kessel, W. (1978). Fuzzy clustering with a fuzzy covariance matrix. In *IEEE Conference on Decision and Control*.
- Madani, K., Rybnik, M., & Chebira, A. (2003). Non Linear Process Identification Using a Neural Network Based Multiple Models Generator. *LNCS series*, 647-654.
- Masson, M.-H., & Denoeux, T. (2008). ECM: An evidential version of the fuzzy c-means algorithm. *Pattern Recognition*, 41(4), 1384 - 1397.
- Petit-Renaud, S., & Denoeux, T. (1999). Regression analysis using fuzzy evidence theory. *Proceedings of FUZZ-IEEE*, 3, 1229-1234.
- Ramdani, M., Mourot, G., & Ragot, J. (2005). A Multi-Modeling Strategy based on Belief Function Theory. In *CDC-ECC '05*.
- Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008). Damage Propagation Modeling for Aircraft Engine Run-to-Failure Simulation. In *IEEE Int. Conf. on Prognostics and Health Management*.
- Serir, L., Ramasso, E., & Zerhouni, N. (2011). E2GK: Evidential Evolving Gustafsson-Kessel Algorithm For Data Streams Partitioning Using Belief Functions. In *11th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*.
- Smets, P., & Kennes, R. (1994). The Transferable Belief Model. *Artificial Intelligence*, 66, 191-234.
- Takagi, T., & Sugeno, M. (1985). Fuzzy identification of systems and its application to modeling and control. *IEEE Trans. On Systems Man and Cybernetics*, 15, 116-132.
- Yager, R. R., & Filev, D. P. . (1995). Including probabilistic uncertainty in fuzzy logic controller modeling using Dempster-Shafer theory. *IEEE Trans. Syst., Man, Cybern.*, 25, 1221-1230.

Bayesian fatigue damage and reliability analysis using Laplace approximation and inverse reliability method

Xuefei Guan¹, Jingjing He², Ratneshwar Jha¹, Yongming Liu²

¹ Department of Mechanical & Aeronautical Engineering, Clarkson University, Potsdam, NY, 13699, USA

guanx@clarkson.edu

rjha@clarkson.edu

² Department of Civil & Environmental Engineering, Clarkson University, Potsdam, NY, 13699, USA

jihe@clarkson.edu

yliu@clarkson.edu

ABSTRACT

This paper presents an efficient analytical Bayesian method for reliability and system response estimate and update. The method includes additional data such as measurements to reduce estimation uncertainties. Laplace approximation is proposed to evaluate Bayesian posterior distributions analytically. An efficient algorithm based on inverse first-order reliability method is developed to evaluate system responses given a reliability level. Since the proposed method involves no simulations such as Monte Carlo or Markov chain Monte Carlo simulations, the overall computational efficiency improves significantly, particularly for problems with complicated performance functions. A numerical example and a practical fatigue crack propagation problem with experimental data are presented for methodology demonstration. The accuracy and computational efficiency of the proposed method is compared with simulation-based methods.

1. INTRODUCTION

Efficient inference on reliability and responses of engineering systems has drawn attention to the prognostics and health management society due to the increasing complexity of those systems (Melchers, 1999; Brauer & Brauer, 2009). For high reliability demanding systems such as aircraft and nuclear facilities, time-dependent reliability degradation and performance prognostics must be quantified to prevent potential system failures. Reliable predictions of system reliability and system responses are usually required for decision-making in a time and computational resource constrained situation. The basic idea of time-independent component reliability analysis involves computation of a multi-dimensional integral over the failure domain of the performance function (Madsen, Krenk, & Lind, 1986; Ditlevsen & Madsen, 1996; Rackwitz, 2001). For many practical problems with high-dimensional parameters, the exact evaluation of this integral is either analytically intractable or computationally infeasible with a given time constraint. Analytical approximations and numerical simulations are two major computational methods to solve this problem (Rebba & Mahadevan, 2008).

The simulation-based method includes direct Monte Carlo (MC) (Kalos & Whitlock, 2008), Importance Sampling (IS) (Gelman & Meng, 1998; Liu, 1996), and other MC simulations with different sampling techniques. Analytical approximation methods, such as first- and second-order reliability methods (FORM/SORM) have been developed to estimate the reliability without large numbers of MC simulations. FORM and SORM computations are based on linear (first-order) and quadratic (second-order) approximations of the limit-state surface at the *most probable point* (MPP) (Madsen et al., 1986; Ditlevsen & Madsen, 1996). Under the condition that the limit-state surface at the MPP is close to its linear or quadratic approximation and that no multiple MPPs exist in the limit-state surface, FORM/SORM are sufficiently accurate for engineering purposes (Bucher et al., 1990; Cai & Elishakoff, 1994; Zhang & Mahadevan, 2001; Zhao & Ono, 1999). If the final objective is to calculate the system response given a reliability index, the inverse reliability method can be used. The most well-known approach is inverse FORM method proposed in (Der Kiureghian, Yan, & Chun-Ching, 1994; Der Kiureghian & Dakessian, 1998; Li & Foschi, 1998). Several studies for static failure using the inverse FORM method have been reported in the literature. (Du, Sudjianto, & Chen, 2004) proposed an inverse reliability strategy and applied it to the integrated robust and reliability design of a vehicle combustion engine piston. (Saranyasontorn & Manuel, 2004) developed an inverse reliability procedure for wind turbine components. (Lee, Choi, Du, & Gorsich, 2008) used the inverse reliability analysis for reliability-based design optimization of nonlinear multi-dimensional systems. (Cheng, Zhang, Cai, & Xiao, 2007) presented an artificial neural network based inverse FORM method for solving problems with complex and implicit performance functions. (Xiang & Liu, 2011) applied the inverse FORM method to time-dependent fatigue life predictions.

Conventional forward and inverse reliability analysis is based on the existing knowledge about the system (e.g., underlying physics, distributions of input variables). Time-dependent reliability degradation and system response changing are not reflected. For many practical engineering problems, usage monitoring or inspection data are usually available at a regular time interval either via structural health monitoring system or non-destructive inspections. The new information can be used to update the initial estimate of sys-

Xuefei Guan et.al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

tem reliability and responses. The critical issue is how to incorporate the existing knowledge and new information into the estimation. Many methodologies have been proposed to handle reliability updating problems. Bayesian updating is the most common approach to incorporate these additional data. By continuous Bayesian updating, all the variables of interest are updated and the inference uncertainty can be significantly reduced, provided the additional data are relevant to the problem and they are informative. (Hong, 1997) presented the idea of reliability updating using inspection data. (Papadimitriou, Beck, & Katafygiotis, 2001) reported a reliability updating procedure using structural testing data. (Graves, Hamada, Klamann, Koehler, & Martz, 2008) applied the Bayesian network for reliability updating. (Wang, Rabiei, Hurtado, Modarres, & Hoffman, 2009) used Bayesian reliability updating for aging airframe. A similar updating approach using Maximum relative Entropy principles has also been proposed in (Guan, Jha, & Liu, 2009). In those studies, MCMC simulations have been extensively used to draw samples from posterior distributions. The Convergence Theorem ensures the resulting Markov chain converges to the target distribution (Gilks, Richardson, & Spiegelhalter, 1996) and it becomes almost a standard approach for Bayesian analysis with complex models. For practical problems with complicated performance functions, simulations are time-consuming and efficient computations are critical for time constrained reliability evaluation and system response prognostics. Some of the existing analytical methods includes variational methods (Ghahramani & Beal, 2000) and expectation maximization methods (Moon, 1996). Those methods usually focus on the approximation of distributions and does not provide a systematical procedure for inverse reliability problems. In structural health management settings, simulation-based method may be infeasible because updating is frequently performed upon the arrival of sensor data. All these application require efficient and accurate computations. However, very few studies are available on the investigation of complete analytical updating and estimation procedure without using simulations.

The objective of the proposed study is to develop an efficient analytical method for system reliability and response updating without using simulations. Three computational components evolved in this approach are Bayesian updating, reliability estimation, and system response estimation given a reliability or a confidence level. For Bayesian updating, Laplace method is proposed to obtain an analytical representation of the Bayesian posterior distribution and avoid MCMC simulations. Once the analytical posterior distribution is obtained, FORM method can be applied to update system reliability or probability of failure. In addition, predictions of system response associated with a reliability or a confidence level can also be updated using inverse FORM method to avoid MC simulations.

The paper is organized as follows. First, a general Bayesian posterior model for uncertain variables is formulated. Relevant information such as response measures and usage monitoring data are used for updating. Then an analytical approximation to the posterior distribution is derived based on Laplace method. Next, FORM method is introduced to estimate system reliability levels and a simplified algorithm based on inverse FORM method is formulated to calculate system response given a reliability level or a confidence level. Following that, numerical and application examples are presented to demonstrate the proposed method. The efficiency and accuracy of the proposed method are compared with simulation results.

2. PROBABILISTIC MODELING AND LAPLACE APPROXIMATION

In this section, a generic posterior model for uncertain parameters is formulated using Bayes' theorem to incorporate additional data such as measurements. Uncertainties from model parameters, measurement, and model independent variables are systematically included. To avoid MCMC simulations as in classical Bayesian applications, Laplace approximation is derived to obtain an analytical representation of the posterior distribution. The updated reliability and system responses can readily be evaluated using this posterior approximation.

2.1 Bayesian modeling for uncertain parameters

Consider a generic parameterized model $\mathcal{M}(y; x)$ describing an observable event d , where x is an uncertain model parameter vector and y is model independent variable. If the model is perfect, one obtains $\mathcal{M}(y; x) = d$. In reality, such a perfect model is rarely available due to uncertainties such as the simplification of the actual complex physical mechanisms, statistical error in obtaining the parameter x , and the measurement error in d . Using probability distributions to describe those uncertainties is a common practice.

Given the prior probability distribution of x , $p(x|\mathcal{M})$, and the known relationship (conditional probability distribution or likelihood function) between d and x , $p(d|x, \mathcal{M})$, the posterior probability distribution $p(x|d, \mathcal{M})$ is expressed using Bayes' theorem as

$$p(x|d, \mathcal{M}) = p(x|\mathcal{M})p(d|x, \mathcal{M}) \frac{1}{\mathcal{Z}} \propto p(x|\mathcal{M})p(d|x, \mathcal{M}), \quad (1)$$

where $\mathcal{Z} = \int_X p(x|\mathcal{M})p(d|x, \mathcal{M})dx$ is the normalizing constant.

The model \mathcal{M} is assumed to be the only feasible model and \mathcal{M} is omitted hereafter for simplicity. Let m be the model prediction and e the error term (for example, the measurement error of d). The variable d reads

$$d = m + e. \quad (2)$$

The probability distribution for m is represented by the function $p(m|x) = f_M(m)$ and the probability distribution for e is by the function $p(e|x) = f_E(e)$. The conditional probability distribution of $p(d|x)$ can be obtained by marginalizing the joint probability distribution of $p(d, m, e|x)$ as follows:

$$p(d|x) = \int_M \int_E p(m|x)p(e|x)p(d, m, e|x)dedm. \quad (3)$$

Because $d = m + e$,

$$p(d, z, e|x) = \delta(d - m - e). \quad (4)$$

Substitute Eq. (4) into Eq. (3) to obtain

$$p(d|x) = \int_M f_M(m)f_E(d - m)dm. \quad (5)$$

Next, terms $f_M(m)$ and $f_E(e)$ need to be determined. Consider a general case where the model prediction m has a statistical noise component $\epsilon \in \mathcal{E}$ with a distribution function $p(\epsilon|x) = f_{\mathcal{E}}(\epsilon)$ due to the modeling error $m = \mathcal{M}(y; x) + \epsilon$. Equation (2) is revised as

$$d = \mathcal{M}(y; x) + \epsilon + e. \quad (6)$$

Marginalizing $p(m|\epsilon, \theta) = \delta(m - \mathcal{M}(y; x) - \epsilon)$ over ϵ to obtain

$$f_M(m) = \int_{\mathcal{E}} p(\epsilon|x)p(m|x, \theta)d\epsilon = f_{\mathcal{E}}(m - \mathcal{M}(y; x)). \quad (7)$$

For the purpose of illustration, ϵ and e are assumed to be two independent Gaussian variables with standard deviations of σ_ϵ and σ_e , respectively. This assumption is usually made when no other information about the uncertain variables is available (Gregory, 2005). Equation (5) is the convolution of two Gaussians and it can be further reduced to another Gaussian distribution as

$$p(d|x) = \frac{1}{\sqrt{2\pi(\sigma_\epsilon^2 + \sigma_e^2)}} \exp\left[-\frac{(d - \mathcal{M}(y;x))^2}{2(\sigma_\epsilon^2 + \sigma_e^2)}\right]. \quad (8)$$

Substituting Eq. (3) into Eq. (1) yields the posterior probability distribution of the uncertain parameter x incorporating the observable event d . The reliability or system state variables can readily be updated with Eq. (1). For problems with high dimensional parameters, the evaluation of Eq. (1) is rather difficult because the exact normalizing constant \mathcal{Z} , which is a multi-dimensional integral, is either analytically intractable or computationally expensive. Instead of evaluating this equation directly, the most common approach is to draw samples from it using MCMC simulations. For applications where performance functions are computationally expensive to evaluate, this approach is time-consuming and hence not suitable for online updating and prognostics. To improve the overall computational efficiency, Laplace method is proposed to approximate the non-normalized Bayesian posterior distribution of $p(x|d)$. The derivation of Laplace approximation is presented below.

2.2 Laplace approximation for Bayesian posterior distributions

Consider the above non-normalized multivariate distribution $p(x|d)$ in Eq. (1) and its natural logarithm $\ln p(x|d)$. Expanding $\ln p(x|d)$ using Tylor series around an arbitrary point x^* yields

$$\begin{aligned} \ln p(x|d) = & \ln p(x^*|d) + (x - x^*)^T \nabla \ln p(x^*|d) + \\ & \frac{1}{2!} (x - x^*)^T [\nabla^2 \ln p(x^*|d)] (x - x^*) + \\ & O((x - x^*)^3), \end{aligned} \quad (9)$$

where $\nabla \ln p(x^*|d)$ is the gradient of $\ln p(x|d)$ evaluated at x^* , $\nabla^2 \ln p(x^*|d)$ is the Hessian matrix evaluated at x^* , and $O(\cdot)$ are higher-order terms. Assume that the higher-order terms are negligible in computation with respect to the other terms. We obtain

$$\begin{aligned} \ln p(x|d) \approx & \ln p(x^*|d) + \underbrace{(x - x^*)^T \nabla \ln p(x^*|d)}_{(*)} + \\ & \frac{1}{2!} (x - x^*)^T [\nabla^2 \ln p(x^*|d)] (x - x^*). \end{aligned} \quad (10)$$

The term $(*)$ is zero at local maxima (denoted as x_0) of the distribution since $\nabla \ln p(x_0|d) = 0$. Therefore, if we choose to expand $\ln p(x|d)$ around x_0 , we can eliminate term $(*)$ in Eq. (10) to obtain

$$\ln p(x|d) \approx \ln p(x_0|d) + \frac{1}{2} (x - x_0)^T [\nabla^2 \ln p(x_0|d)] (x - x_0). \quad (11)$$

Exponentiating $\ln p(x|d)$ of Eq. (11) yields

$$e^{\ln p(x|d)} \approx p(x_0|d) \exp\left\{-\frac{1}{2} (x - x_0)^T [-\nabla^2 \ln p(x_0|d)] (x - x_0)\right\}. \quad (12)$$

The last term of Eq. (12) resembles remarkably a multivariate Gaussian distribution with a mean vector of x_0 and a covariance matrix $\Sigma = [-\nabla^2 \ln p(x_0|d)]^{-1}$. The normalizing constant is

$$\mathcal{Z} = \int_X e^{\ln p(x|d)} dx \approx p(x_0|d) \sqrt{(2\pi)^n |\Sigma|}, \quad (13)$$

where $\Sigma = [-\nabla^2 \ln p(x_0|d)]^{-1}$, n is the dimension of the variable x , and $|\Sigma|$ is the determinant of Σ .

The non-normalized Bayesian posterior distribution $p(x|d)$ is now approximated as

$$p(x|d) \approx \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left\{-\frac{1}{2} (x - x_0)^T [\Sigma^{-1}] (x - x_0)\right\}, \quad (14)$$

which is a multivariate Gaussian distribution with a mean vector of x_0 and a covariance matrix Σ . To compute x_0 and Σ , the first step is to find the local maxima of $\ln p(x|d)$ and evaluate the Hessian of $\ln p(x|d)$ at the local maxima. Numerical root-finding algorithms can be used to find local maxima, such as Gauss-Newton algorithm (Dennis Jr, Gay, & Walsh, 1981), Levenberg-Marquardt algorithm (More, 1978), trust-region dogleg algorithm (Powell, 1970), and so on. Laplace method can yield accurate results given the target distribution is approximately Gaussian distributed, which is quite common for practical problems (Gregory, 2005).

With the analytical representation of the posterior distribution $p(x|d)$, the updated reliability index can be calculated using the FORM method. In addition, updated system response predictions associated with a reliability index or a confidence level can also be calculated using inverse FORM method. For the sake of completeness, the basic concept of the FORM and inverse FORM methods are introduced briefly.

3. FORM AND INVERSE FORM METHODS

The time-invariant reliability analysis entails computation of a multi-dimensional integral over the failure domain of the performance function.

$$P_F \equiv P[g(x) < 0] = \int_{g(x) < 0} f_X(x) dx, \quad (15)$$

where $x \in \mathcal{R}^n$ is a real-valued n -dimensional uncertain variable, $g(x)$ is the performance function, such that $g(x) < 0$ represents the failure domain, P_F is the probability of failure, and $f_X(x)$ is the joint probability distribution of x . The surface $g(x) = 0$ is usually called limit-state surface. In FORM/SORM methods, the uncertain variable is usually transformed from the standard probability space to the standard Gaussian space, also referred to as *reduced variable space*. Denote the transformed performance function as $g(z)$, where $z \in \mathcal{R}^n$ is an n -dimensional standard Gaussian variable, also called *reduced variable*. The distance between the closest point (most probable point (MPP), labeled as MPP in Figure 1) on the limit-state surface $g(z) = 0$ to the origin in the reduced variable space is the Hasofer-Lind reliability index (Madsen et al., 1986), denoted as β_{HL} in Figure 1. MPP is also known as the *design point*.

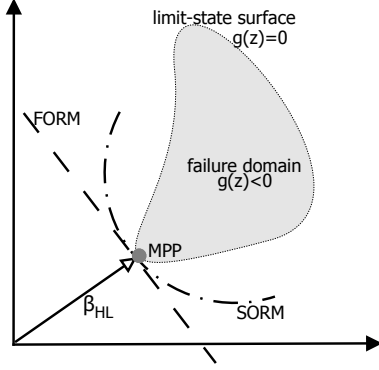


Figure 1: Linear (FORM) and quadratic (SORM) approximations of the performance function at MPP on the limit-state surface.

Reliability analysis entails the computation of β_{HL} and the design point, which is a standard constrained optimization problem defined as

$$\text{minimize: } \|z\| \quad \text{subject to } g(z) = 0, \quad (16)$$

where $\|z\|$ denotes the distance between the point z and the origin in the reduced variable space.

The design point is generally not known a priori, hence an iterative process is required to find the design point z^* in the reduced variable space such that $\beta_{HL} \equiv \|z^*\|$ corresponds to the shortest distance between z^* and the origin of the reduced variable space. Because reduced variables are based on the mean and standard deviation of a normal distribution, the non-normal variables must be transferred to its equivalent normal distribution. Rackwitz-Fiessler (Madsen, 1977) procedure is usually adopted for this purpose. The idea requires the cumulative density function (CDF) and the probability density function (PDF) of the target distribution be equal to a normal CDF and PDF at the value of variable x on the limit-state surface. This procedure finds the mean μ_{eq} and standard deviation σ_{eq} of the equivalent normal distribution and thus the variable x can be reduced to a standard Gaussian variable $z = (x - \mu_{eq})/\sigma_{eq}$. Several algorithms are available to locate the design point z^* , for example the Hasofer & Lind - Rackwitz & Fiessler (HL-RF) algorithm (Hasofer & Lind, 1974; Rackwitz & Fiessler, 1978). With an initial guess of z_0 on the limit-state surface, the basic procedure computes the new location for z^* iteratively according to

$$z_{k+1} = \frac{1}{|\nabla g(z_k)|^2} [\nabla g(z_k) z_k - z_k] \nabla g(z_k)^T. \quad (17)$$

A reasonable guess can be fixing the first $n-1$ components of z_0 to its distribution means and solving for the last component on the limit-state surface. The iterative procedure terminates based on some criteria such as $|\beta_{k+1} - \beta_k| < \epsilon_\beta$, where ϵ_β is a small control parameter assigned by users. Usually a value of $\epsilon_\beta = 10^{-4}$ to 10^{-3} yields accurate results for β_{HL} and the design point (Cheng et al., 2007).

After finding the design point and β_{HL} by solving Eq. (16) using the iterative formula of Eq. (17), FORM or SORM can approximate the probability of failure using a linear or quadratic approximation of the performance function, respectively. Both of them are based on Taylor series expansion of the performance function around the design point truncated

to linear and quadratic terms. For example, using FORM method yields the probability of failure as

$$P_F^{\text{FORM}} \approx \Phi(-\beta_{HL}), \quad (18)$$

where Φ is the standard Gaussian CDF. The precision of this approximation depends on the non-linearity of the limit-state surface. Experience shows that FORM method yields accurate results for general engineering purposes (Cheng et al., 2007). FORM is a widely used computational model in reliability index approach (RIA) for reliability-based design optimization (RBDO) since it finds the reliability index β_{HL} . The advantage of RIA is that the probability of failure is forwardly calculated for a given design. However, inverse reliability analysis in performance measure approach (PMA) is known to be more robust and informative than the reliability analysis in RIA (Tu, Choi, & Park, 1999; Youn, Choi, & Du, 2005). The idea of inverse reliability analysis in PMA is to investigate whether a given design satisfies the probabilistic constraint with a target reliability index β_t . The inverse reliability analysis can also be expressed as an optimization problem such that

$$\text{minimize: } g(z) \quad \text{subject to } \|z\| = \beta_t. \quad (19)$$

In inverse reliability analysis, among the different values of performance function $g(z)$ taking on z that pass through the β_t curve in the reduced variable space, the one z^* that minimizes the performance function is sought. Figure 2 illustrates the inverse reliability analysis. The point z^* is also called MPP and the corresponding minimal value of $g(z^*)$ is called probabilistic performance measure (PPM). Both reliability analysis and inverse reliability analysis search for MPPs. The difference is that the former search for the MPP on the limit-state surface $g(z) = 0$ while the latter search for MPP on the β_t curve. Based on the idea of inverse FORM procedure proposed in (Der Kiureghian et al., 1994), an efficient and simplified iterative formula in the reduced variable space is formulated as:

$$z_{k+1} = z_k + \lambda \left[-\beta_t \frac{\nabla g(z_k)}{|\nabla g(z_k)|} - z_k \right], \quad (20)$$

where ∇ is the gradient vector with respect to z and λ is the step size at the k th iteration (a small constant is used in this formula instead of an adaptive value). The initial value z_0 is usually assigned to the distribution mean value. The iterative procedure proceeds until a convergence is achieved, i.e., when

$$\frac{|z_{k+1} - z_k|}{|z_{k+1}|} \leq \epsilon, \quad (21)$$

where ϵ is a small quantity assigned by the user. For practical problems, $\epsilon_t = 10^{-4}$ to 10^{-3} usually yields satisfactory estimates (Cheng et al., 2007). Based on the iterative formula, an algorithm locating MPP in inverse reliability problems is given as Algorithm 1.

Algorithm 1 Inverse FORM algorithm solving MPP given a target reliability index β_t

- 1: Initiate z_0 and λ , set $k = 0$
- 2: **repeat**
- 3: calculate z_{k+1} according to Eq. (20)
- 4: calculate $d = \frac{|z_{k+1} - z_k|}{|z_{k+1}|}$
- 5: $k \leftarrow k + 1$
- 6: **until** $d \leq \epsilon_t$

For a given confidence level, the reliability indexes associated with that level should be first calculated using inverse Gaussian CDF then the MPPs associated with these indexes can be calculated using Algorithm 1. System responses are readily evaluated with these MPPs.

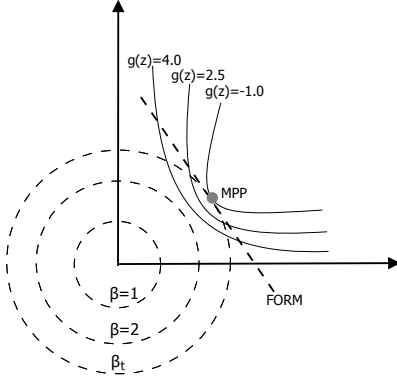


Figure 2: Inverse reliability analysis and MPP for target probability of failure of β_t and the linear approximation of the performance function at MPP labeled as FORM. Values of $g(z)$ are for illustration purposes only.

Both iterative formulae in Eq. (17) and Eq. (20) implicitly assumes that the components of z are uncorrelated. For correlated component variables in z , the correlated components need to be transformed into uncorrelated components via the orthogonal transformation of $z' = L^{-1}(z^T)$, where L is the lower triangular matrix obtained by Cholesky factorization of the correlation matrix R such that $LL^T = R$, where L^T is the transpose of L .

The overall computational procedure according to the proposed method is summarized as follows:

1. Formulate Bayesian posterior distributions according to Eq. (8).
2. Compute the posterior approximation according to Eq. (14).
3. Reliability or probability of failure estimation is calculated using iterative formula of Eq. (17) and Eq. (18).
4. To estimate system responses associated with a reliability level or confidence level, calculate MPPs using Algorithm 1 and then calculate system responses with the obtained MPPs.

Prior estimations are evaluated according to Steps 3 and 4 using prior distributions. To illustrate the proposed method, several examples are presented in the next section.

4. EXAMPLES

A numerical example is given first to illustrate the overall procedure, and a practical fatigue crack propagation problem with experimental data and a beam example with finite element analysis data are demonstrated. Comparisons with traditional simulation-based methods are made to investigate the accuracy and computational efficiency of the proposed method.

4.1 A numerical example with two uncertain variables

Consider a performance function $f(x, y) = x + y$ describing an observable event $z = f(x, y) + \epsilon$, where x and y are two uncertain variables and ϵ is an Gaussian error term with zero mean and a standard deviation of $\sigma_\epsilon = 0.5$. Variable x is normally distributed with a mean of $\mu_x = 2$ and a standard deviation of $\sigma_x = 0.5$ and variable y is also normally distributed with a mean $\mu_y = 5$ and a standard deviation $\sigma_y = 1.5$. Variables x and y are correlated with a correlation coefficient of $\rho_{xy} = -0.5$. The covariance matrix is $\Sigma_{xy} = \begin{bmatrix} \sigma_x^2 & \sigma_x \sigma_y \rho_{xy} \\ \sigma_x \sigma_y \rho_{xy} & \sigma_y^2 \end{bmatrix}$. $f(x, y) > 9$ is defined as failure event and the limit-state surface is $f(x, y) - 9 = 0$. The likelihood function can be expressed according to Eq. (8) as

$$p(z|x) = \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \exp \left\{ -\frac{1}{2} \left[\frac{z - f(x, y)}{\sigma_\epsilon} \right]^2 \right\}. \quad (22)$$

Assume that the evidence of $z = 8$ is observed. The posterior distribution that encodes this information is formulated according to Eq. (1) as,

$$p(x|z) = \frac{1}{\sqrt{2\pi}|\Sigma_{xy}|} \exp \left\{ -\frac{1}{2} \begin{pmatrix} x - \mu_x \\ y - \mu_y \end{pmatrix}^T \Sigma_{xy}^{-1} \begin{pmatrix} x - \mu_x \\ y - \mu_y \end{pmatrix} \right\} \times \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \exp \left\{ -\frac{1}{2} \left[\frac{z - f(x)}{\sigma_\epsilon} \right]^2 \right\}. \quad (23)$$

Based on the information given above, the prior estimate of the probability of failure for event $f(x, y) > 9$ and the prediction of system response z associated with a given reliability or confidence level can be calculated using FORM and inverse FORM methods. After obtaining the additional data $z = 8$, those estimates can be updated using the proposed analytical procedure. The updating process firstly involves the Laplace approximation for the posterior of Eq. (23). Then the iterative formula of Eq. (17) is employed to find the design point (x^*, y^*) and β , and the probability of failure P_F can be estimated using FORM according to $\Phi(-\beta)$.

To calculate the confidence bound (e.g., $[lo, up] = [0.025, 0.975]$ bound) of z , reliability indexes associated with the upper and lower limits are first calculated according to $\beta_{lo} = \Phi^{-1}(lo)$ and $\beta_{up} = \Phi^{-1}(up)$. The iterative inverse FORM formula of Eq. (20) solves the required design point for β_{lo} and β_{up} . Finally the confidence bound of z can be computed using these two design points. To compare the efficiency and the accuracy, MC and MCMC simulations serve as benchmark solutions to this example. Table 1 presents results for this example. The prior estimates for probability of failure (PoF) and interval prediction are calculated using FORM and inverse FORM methods, respectively. For this simple example, just a few function evaluations ensure obtaining converged results. A crude Monte Carlo simulation with 10^6 samples yields very close results. For the posterior estimate with Bayesian updating, the proposed analytical solutions using Laplace, FORM, and inverse FORM (results are labeled as iFORM in all the tables hereafter) methods are very close to the solution obtained using MCMC simulation with a chain length of 10^6 . By comparing the number of function evaluations between the analytical and MC or MCMC solutions, it is observed that the proposed analytical method can reduce the computational cost by several orders of magnitude. It would be significantly advantageous to use the proposed analytical procedure for time constrained or online prognosis systems.

Table 1: Probability of Failure (PoF), confidence interval (CI) estimates, and the number of function evaluations (NFE) for $f(x, y) = x + y$. Both x and y are normally distributed with means of 2 and 5 and standard deviations of 0.5 and 1.5, respectively. The correlation coefficient between x and y is -0.5 . The failure is defined as $f(x, y) > 9$.

	Method	PoF	95 CI	NFE
prior	FORM,iFORM	0.030181	4.2570~9.7430	21
	MC	0.030417	4.2493~9.7455	10^6
posterior	Laplace,FORM,iFORM	0.0045455	6.6536~8.9852	47
	MCMC	0.0046910	6.6506~8.9898	10^6

4.2 Reliability updating and response prognostics of a fatigue crack damaged system with experimental data

In this section, a practical fatigue crack damage problem is presented with experimental data. As a typical damage mode in many structures, the reliability of a system with possible fatigue cracks must be accurately quantified in order to avoid severe failure events. Because fatigue crack propagation is a time-dependent process, crack growth prognosis provides valuable information for system maintenance or unit replacement. Due to the stochastic nature of fatigue crack propagation, fatigue crack growth is not a smooth and stable process. Therefore additional information such as usage information from health monitoring systems and crack size measures from inspections can be used to update various quantities of interest. By performing continuous updating, uncertainties associated with system reliability and crack size prognosis can be reduced for decision-making. Because crack growth equations are usually in the forms of differential equations or finite element models, simulation-based methods are relatively more expensive in terms of computational cost. To demonstrate the updating procedure with the proposed method and validate its effectiveness and efficiency, experimental data are incorporated in this example. A portion of the experimental data is used to obtain the parameter distributions of the crack growth equation and one from the rest of the dataset is arbitrarily chosen to represent the "actual" target system. First we estimate PoF and crack growth prognosis with the prior parameter distributions. Then we choose a few points from the "actual" target system to represent measurements from crack size inspections. These measures are used to perform Bayesian updating with the analytical methods proposed in previous sections. Both system reliability and crack growth prognosis are updated. Results are compared with simulation-based methods in terms of accuracy and efficiency.

(Virkler, Hillberry, & Goel, 1979) reported a large set of fatigue crack propagation data on aluminum alloy 2024-T3. The dataset consists of fatigue crack propagation trajectories recorded from 68 center-through crack specimens, each of which has the same geometry, loading, and material configurations. Each specimen has a width of $w = 154.2\text{mm}$ and a thickness of $d = 2.54\text{mm}$. The initial crack size is $a_0 = 9.0\text{mm}$. A constant cyclic loading with a stress range of $\Delta\sigma = 48.28\text{MPa}$ was applied. Without loss of generality, the classical Paris' equation (Paris & Erdogan, 1963) is chosen as the crack growth rate governing equation. Other crack growth equations can also be applied with the same procedure. Paris' equation describes the crack growth rate per one constant cyclic load as

$$\frac{da}{dN} = c(\Delta K)^m, \quad (24)$$

where ΔK is the stress intensity range in one loading cycle.

For this particular crack and geometry configuration, $\Delta K = \sqrt{\pi a}[\sec(\pi a/w)]\Delta\sigma$. Terms c and m are uncertainty model parameters that are usually obtained via statistical regression analysis of experimental testing data. For convenience, lnc is usually used instead of c . Given a specific number of loading cycles, solving the ordinary differential equation in Eq. (24) gives the corresponding crack size.

The first fifteen crack growth trajectories from Virkler's dataset identifies these two parameters using Maximum Likelihood Estimation as a joint Gaussian distribution of (lnc, m) with a mean vector of $\mu_0 = [-26.7084, 2.9687]$ and a covariance matrix of $\Sigma_0 = \begin{bmatrix} 0.5435 & -0.0903 \\ -0.0903 & 0.0150 \end{bmatrix}$.

$$p_0(\text{lnc}, m) = \frac{1}{2\pi\sqrt{|\Sigma_0|}} \times \exp\left\{-\frac{1}{2}[(\text{lnc}, m) - \mu_0]\Sigma_0^{-1}[(\text{lnc}, m) - \mu_0]^T\right\} \quad (25)$$

As we mentioned earlier in this section, another specimen from the rest of the dataset is arbitrarily chosen to represent the target system. The reliability and crack growth prognosis of this target system are of interest. The prior estimate of reliability and fatigue crack growth prognosis of the target system can then be estimated using this joint distribution and the model in Eq. (24). Let $\mathcal{M}(N; \text{lnc}, m)$ denotes the model output (crack size) given a number of loading cycles N and parameters lnc and m . Three crack size measures a_i with corresponding numbers of loading cycles N_i at the early stage of the target system are chosen to represent the actual inspection data. They are $(a_1, N_1) = (10, 33062)$, $(a_2, N_2) = (11, 55101)$, and $(a_3, N_3) = (12, 75569)$. The standard deviation of Gaussian likelihood is also estimated as $\sigma_a = 0.304\text{mm}$. The failure event is defined as the crack size exceeding 40.0mm given the number of loading cycles as 220,000. With these additional measurement data, the posterior distribution of (lnc, m) (with r response measures) reads

$$p_n(\text{lnc}, m) \propto p_0(\text{lnc}, m) \times \exp\left\{-\frac{1}{2}\sum_{i=1}^r \left[\frac{a_i - \mathcal{M}(N_i; \text{lnc}, m)}{\sigma_a}\right]^2\right\} \quad (26)$$

Following the proposed analytical procedure, we obtain updated results of reliability and crack size prognosis. Table 2 shows the prior and posterior (updated) results of PoF and 95% interval predictions of crack size at 220,000 loading cycles. We can observe from this table that the simulation method requires 200,000 function evaluations while the analytical method requires less than 200 function evaluations to produce similar results.

Figure 3 presents crack growth prognosis results obtained by the proposed analytical method. MCMC simulation results are displayed in the same figure for comparison. Several

Table 2: Prior and updated estimates of Probability of Failure (PoF), confidence interval (CI) for crack size, and the number of function evaluations (NFE) for fatigue crack problem. The failure is defined as the crack size exceeds $a_c = 40\text{mm}$ at the number of loading cycles $N_c = 220,000$. 95% CI predictions are calculated at the number of loading cycles equal to N_c .

Measures	Method	PoF	95% CI	NFE
0(prior)	FORM,iFORM	0.0467	28.8290~41.3095	60
	MCMC	0.0498	28.9417~41.4685	2×10^5
1	Laplace,FORM,iFORM	0.0225	28.3694~39.8084	96
	MCMC	0.0186	28.3563~39.5466	2×10^5
2	Laplace,FORM,iFORM	0.0042	27.7926~37.4207	105
	MCMC	0.0039	27.6989~37.3537	2×10^5
3	Laplace,FORM,iFORM	0.0002	27.2484~34.9112	111
	MCMC	0.0001	27.0817~34.6913	2×10^5

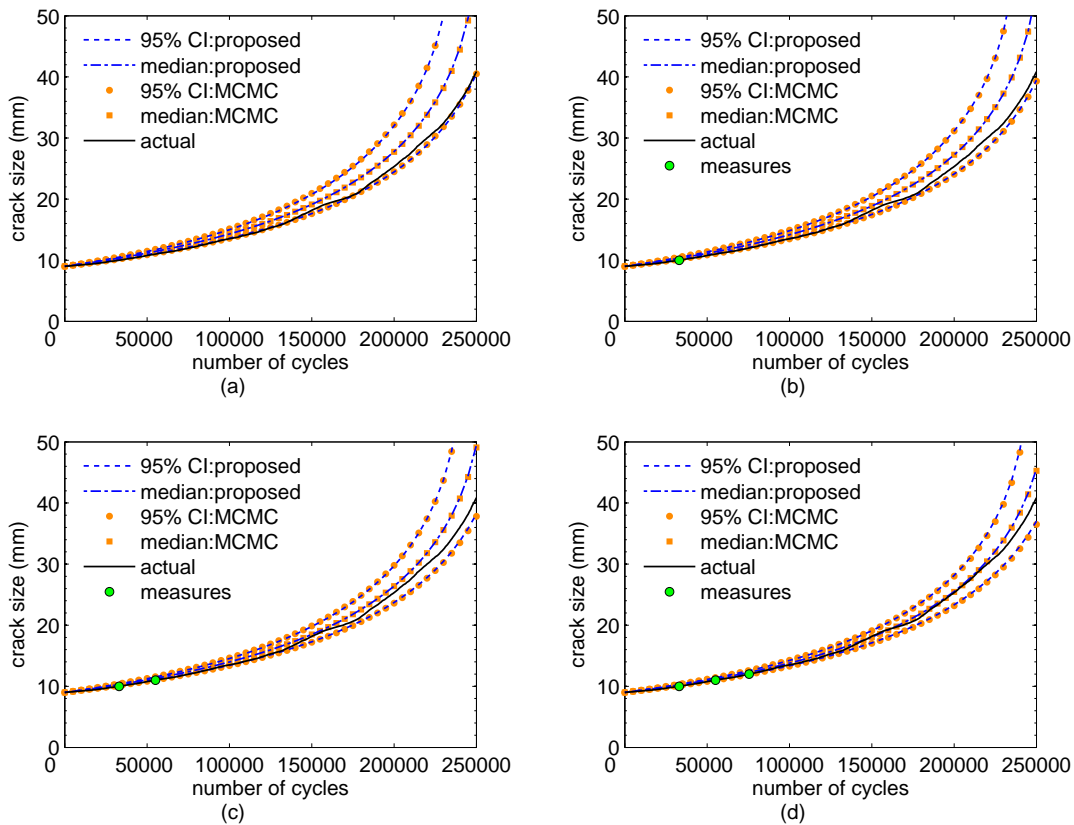


Figure 3: Prior and posterior prognostics of fatigue crack growth trajectory using the proposed method (Laplace,iFORM) and the traditional simulation-based methods (MCMC). Median and 95% interval predictions are presented: (a) prior estimation; (b) updated with 1 measure; (c) updated with 2 measures; (d) updated with 3 measures.

aspects can be observed and interpreted: 1) The proposed analytical Laplace and (inverse) FORM method yields almost identical prognostic results to those obtained using traditional MCMC simulations, which can be confirmed by observing Figure 3(a-d). 2) In Figure 3(a), the prior median and interval prediction of the crack growth is far from the actual target system because of various uncertainty associated with the crack propagation process such as the material uncertainty, modeling uncertainty, as well as measurement uncertainty. These uncertainties are finally encoded into the model parameter ($\ln c$, m) in form of distributions through statistical regression.

These uncertainties cause the prior estimation deviates from the actual target system. 3) Inspection data, or crack size measurement in this example, is critical to improve the accuracy for time-dependent nonlinear system prognostics. With inspection data, uncertainties can be greatly reduced. As shown in Figure 3(b-d), both the median and interval predictions for crack growth trajectories become closer to the actual trajectories as more measurements are integrated into the Bayesian updating process.

4.3 A cantilever beam example

A beam example is used to examine the proposed method through finite element analysis (FEA). Data from FEA provide representative sensor output. By analyzing the sensor output data, frequency information of the beam is extracted to update the finite element model and also the reliability level. For the sake of illustration and simplification, we use a simple cantilever beam. More complex full-scale structural finite element model analysis follows the same procedure as presented here.

A cantilever aluminum beam is divided into ten elements using finite element modeling, as shown in Figure (4). The beam is 1m long, 0.1m wide and 0.01m thick. The design cross section area is $A = 0.001\text{m}^2$. Assume the cross section area of the first segment of the beam (attached to the wall) is modeled by $A_1 = \alpha A$ due to manufacturing uncertainty, where term α is a Gaussian variable with a mean of 1 and a standard deviation of 0.5. Because of usage (aging) and material degradation, α may vary along time. Other segments have deterministic cross section dimensions that are equal to the design value of A . The material has a Young's modulus of $E = 6.96 \times 10^{10}\text{Pa}$ and a density of $2.73 \times 10^3\text{kg/m}^3$.

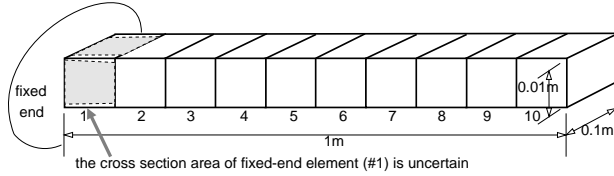


Figure 4: The cantilever beam finite element model. The cross section area of the first element (attached to the wall) is uncertain due to manufacture and usage and is modeled by $A_1 = \alpha A$, where $A = 0.001\text{m}^2$ is design cross section area and $\alpha \sim \text{Norm}(1, 0.02^2)$.

The failure event is defined as the first natural frequency is less than 8Hz due to the degradation of the stiffness of the beam. The sensor data are synthesized by setting $\alpha = 0.95$ and solving the dynamical equation of the beam under a free vibration. After adding 5 percent of Gaussian white noise, the first four mode frequency data are extracted from the sensor data using Fast Fourier Transformation (FFT). They are $(f_1, f_2, f_3, f_4) = (8.03, 50.5, 142, 280)\text{Hz}$.

Based on the above information, the Bayesian posterior for uncertain variable α given the frequency information extracted from the sensor data is

$$p(\alpha) \propto \exp\left\{-\frac{1}{2} \frac{(\alpha - 1)^2}{0.5^2}\right\} \times \exp\left\{-\sum_{j=1}^F \sum_{i=1}^N \left(\{\omega\}_i [(-2\pi f_i)^2 \mathbf{M}(\alpha) + \mathbf{K}(\alpha)] \{\phi_i\}_j\right)^2\right\}, \quad (27)$$

where N is the number of measured mode and F is the number of measured mode shape coordinates. Term $\{\omega\}_i$ is the i th weighting factor for i th frequency component in the likelihood function. For the purpose of illustration, $\{\omega\}$ is configured such that each frequency component has a coefficient of variation of 0.1. Terms $\mathbf{M}(\alpha)$ and $\mathbf{K}(\alpha)$ are the mass and stiffness matrices, respectively. Because α is a variable, actual values for $\mathbf{M}(\alpha)$ and $\mathbf{K}(\alpha)$ depends on each realization of α . Term $\{\phi\}_i$ is the i th mode shape. For the current data, $N = 4$ and $F = 20$.

Using the proposed method we obtain results shown in Table (3). Simulation-based results are also listed in this table for comparison.

Table 3: Prior and posterior estimates of probability of failure (PoF) in the beam example. Frequency data (first four natural frequency extracted from synthesized noisy data via FFT) are used to perform Bayesian updating. Statistics of α (mean, standard deviation (SD)) and computational cost in term of number of function evaluations (NFE) are shown.

	Method	PoF	NFE	α (mean,SD)
prior:	FORM	0.004435	11	1.0, 0.02
	MC	0.004428	10^6	1.0, 0.02
posterior:	Laplace,FORM	0.0182	57	0.9757, 0.0134
	MCMC	0.0164	10^6	0.9760, 0.0135

Results of the proposed method are similar to those obtained using traditional simulation-based methods. However, the computational cost is much smaller. Finite element models in practical problems are usually more sophisticated than this beam example, and simulation-based methods are not feasible for such computationally extensive problems. The proposed method provides an alternative to solving such problems and it yields accurate results under the condition that uncertain variables are approximately Gaussian-like.

In this section, three examples are presented to demonstrate and validate the proposed analytical method. Some important aspects of the proposed method are closely revealed, including the computational benefits in terms of efficiency and accuracy. Appropriate conditions to assure these benefits are also analyzed.

5. CONCLUSIONS

In this paper, an efficient analytical Bayesian method for reliability and system response updating is developed. The method is capable of incorporating additional information such as inspection data to reduce uncertainties and improve the estimation accuracy. One major difference between the proposed work and the traditional approach is that the proposed method performs all the calculations including Bayesian updating without using MC or MCMC simulations. A numerical example, a practical fatigue crack propagation problem with experimental data, and a finite element beam problem with FEA data are presented to demonstrate the proposed method. Comparisons are made with traditional simulation-based methods to investigate the accuracy and efficiency. Based on the current study, several conclusions are drawn.

1. The proposed method provides an efficient analytical computational procedure for computing and updating system reliability responses. No MC or MCMC simulation is required therefore it provides a feasible and practical solution to time constrained or online prognostics. The method is also beneficial for structural health monitoring problems where Bayesian updating and system response predictions are frequently performed upon the arrival of sensor data.

2. The proposed method is capable of incorporating additional information such as the inspection data and usage data from health monitoring system by way of Bayesian updating. This property is beneficial for highly stochastic time-dependent nonlinear system where prior estimates for reliability and system response may become unreliable along with

system developing. By continuous Bayesian updating, estimation uncertainties can be reduced.

3. The proposed method yields almost identical results to those produced by traditional simulation-based methods given that uncertain variables are approximately Gaussian distributed. This is true for most of the engineering problems where the uncertain parameters are normal or log-normal variables (which can be transformed and truncated into normal variables). When these conditions are not assured, the results need careful interpretations. The efficiency and accuracy of the proposed method is demonstrated and verified using three examples. The proposed method provides an alternative for time-constrained prognostics problems. If the problem involves too many random variables, traditional simulation-based method may be more appropriate. Systematical comparisons of the method with other approaches such as variational method will be conducted in the future.

ACKNOWLEDGMENTS

The research reported in this paper was supported by the NASA ARMD/AvSP IVHM project under NRA NNX09AY54A. The support is gratefully acknowledged.

REFERENCES

- Brauer, D., & Brauer, G. (2009). Reliability-centered maintenance. *Reliability, IEEE Transactions on*, 36(1), 17–24.
- Bucher, U., et al. (1990). A fast and efficient response surface approach for structural reliability problems. *Structural Safety*, 7(1), 57–66.
- Cai, G., & Elishakoff, I. (1994). Refined second-order reliability analysis. *Structural Safety*, 14(4), 267–276.
- Cheng, J., Zhang, J., Cai, C., & Xiao, R. (2007). A new approach for solving inverse reliability problems with implicit response functions. *Engineering structures*, 29(1), 71–79.
- Dennis Jr, J., Gay, D., & Walsh, R. (1981). An adaptive nonlinear least-squares algorithm. *ACM Transactions on Mathematical Software (TOMS)*, 7(3), 348–368.
- Der Kiureghian, A., & Dakessian, T. (1998). Multiple design points in first and second-order reliability. *Structural Safety*, 20(1), 37–49.
- Der Kiureghian, A., Yan, Z., & Chun-Ching, L. (1994). Inverse reliability problem. *Journal of engineering mechanics*, 120(5), 1154–1159.
- Ditlevsen, O., & Madsen, H. (1996). *Structural reliability methods* (Vol. 315). Citeseer.
- Du, X., Sudjianto, A., & Chen, W. (2004). An integrated framework for optimization under uncertainty using inverse reliability strategy. *Journal of Mechanical Design*, 126, 562.
- Gelman, A., & Meng, X. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2), 163–185.
- Ghahramani, Z., & Beal, M. (2000). Variational inference for Bayesian mixtures of factor analysers. *Advances in neural information processing systems*, 12, 449–455.
- Gilks, W., Richardson, S., & Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC.
- Graves, T., Hamada, M., Klamann, R., Koehler, A., & Martz, H. (2008). Using simultaneous higher-level and partial lower-level data in reliability assessments. *Reliability Engineering & System Safety*, 93(8), 1273–1279.
- Gregory, P. (2005). *Bayesian logical data analysis for the physical sciences: a comparative approach with Mathematica support*. Cambridge Univ Pr.
- Guan, X., Jha, R., & Liu, Y. (2009). Probabilistic fatigue damage prognosis using maximum entropy approach. *Journal of Intelligent Manufacturing*, 1-9. (10.1007/s10845-009-0341-3)
- Hasofer, A., & Lind, N. (1974). Exact and invariant second-moment code format. *Journal of the Engineering Mechanics Division*, 100(1), 111–121.
- Hong, H. (1997). Reliability analysis with nondestructive inspection. *Structural Safety*, 19(4), 383–395.
- Kalos, M., & Whitlock, P. (2008). *Monte carlo methods*. Wiley-VCH.
- Lee, I., Choi, K., Du, L., & Gorsich, D. (2008). Inverse analysis method using MPP-based dimension reduction for reliability-based design optimization of nonlinear and multi-dimensional systems. *Computer Methods in Applied Mechanics and Engineering*, 198(1), 14–27.
- Li, H., & Foschi, R. (1998). An inverse reliability method and its application. *Structural Safety*, 20(3), 257–270.
- Liu, J. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, 6(2), 113–119.
- Madsen, H. (1977). Some experience with the Rackwitz-Fiessler algorithm for the calculation of structural reliability under combined loading. *DIALOG-77, Danish Engineering Academy, Lyngby, Denmark*, 73–98.
- Madsen, H., Krenk, S., & Lind, N. (1986). *Methods of structural safety*. Prentice-Hall, Inc., Englewood Cliffs, NJ.
- Melchers, R. (1999). *Structural reliability analysis and prediction*. John Wiley & Son Ltd.
- Moon, T. (1996). The expectation-maximization algorithm. *Signal Processing Magazine, IEEE*, 13(6), 47–60.
- More, J. (1978). The Levenberg-Marquardt algorithm: implementation and theory. *Numerical analysis*, 105–116.
- Papadimitriou, C., Beck, J., & Katafygiotis, L. (2001). Updating robust reliability using structural test data. *Probabilistic Engineering Mechanics*, 16(2), 103–113.
- Paris, P., & Erdogan, F. (1963). A critical analysis of crack propagation laws. *Journal of Basic Engineering*, 85(4), 528–534.
- Powell, M. (1970). A FORTRAN subroutine for solving systems of nonlinear algebraic equations. *Numerical methods for nonlinear algebraic equations*, 115.
- Rackwitz, R. (2001). Reliability analysis—a review and some perspectives. *Structural Safety*, 23(4), 365–395.
- Rackwitz, R., & Flessler, B. (1978). Structural reliability under combined random load sequences. *Computers & Structures*, 9(5), 489–494.
- Rebba, R., & Mahadevan, S. (2008). Computational methods for model reliability assessment. *Reliability Engineering & System Safety*, 93(8), 1197–1207.
- Saranyasoontorn, K., & Manuel, L. (2004). Efficient models for wind turbine extreme loads using inverse reliability. *Journal of Wind Engineering and Industrial Aerodynamics*, 92(10), 789–804.
- Tu, J., Choi, K., & Park, Y. (1999). A new study on reliability-based design optimization. *Journal of Mechanical Design*, 121, 557.
- Virkler, D., Hillberry, B., & Goel, P. (1979). The Statistical Nature of Fatigue Crack Propagation. *Journal of Engineering Materials and Technology*, 101, 148.

- Wang, X., Rabiei, M., Hurtado, J., Modarres, M., & Hoffman, P. (2009). A probabilistic-based airframe integrity management model. *Reliability Engineering & System Safety*, 94(5), 932–941.
- Xiang, Y., & Liu, Y. (2011). Application of inverse first-order reliability method for probabilistic fatigue life prediction. *Probabilistic Engineering Mechanics*, 26(2), 148–156.
- Youn, B., Choi, K., & Du, L. (2005). Enriched performance measure approach for reliability-based design optimization. *AIAA journal*, 43(4), 874–884.
- Zhang, R., & Mahadevan, S. (2001). Integration of computation and testing for reliability estimation. *Reliability Engineering & System Safety*, 74(1), 13–21.
- Zhao, Y., & Ono, T. (1999). A general procedure for first/second-order reliability method (form/sorm). *Structural Safety*, 21(2), 95–112.

Xuefei Guan is a graduate research assistant in department of mechanical engineering at Clarkson University. He received his B.S. degree in Reliability Engineering and M.S. degree in Aeronautical Engineering from Beihang University in China in 2005 and 2008, respectively. His research interests are probabilistic analysis, Bayesian and entropy-based methods and applications, Markov Chain Monte Carlo simulation, and reliability analysis.

Jingjing He is a graduate research assistant in department of civil engineering at Clarkson University. Her research interests are fatigue analysis, structural dynamics, diagnosis and prognosis. She received her B.S. degree in Reliability Engineering and M.S. degree in Aerospace System Engineering from Beihang University in China in 2005 and 2008, respectively.

Ratneshwar Jha is an Associate Professor in the Department of Mechanical and Aeronautical Engineering at Clarkson University. His research interests include structural health monitoring, modeling of composite and smart structures, adaptive control of structural vibrations, intelligent flight controls, and multidisciplinary design optimization. Dr. Jha is an Associate Fellow of AIAA and a member of ASME and ASEE. Dr. Jha earned PhD in Mechanical Engineering from Arizona State University in 1999, MS in Aerospace Engineering from Georgia Institute of Technology in 1983, and B. Tech in Aeronautical Engineering from Indian Institute of Technology in 1981. Dr. Jha worked in the aerospace industry from 1983 to 1995 where he led a team of engineers working on conceptual and preliminary designs of combat aircraft.

Yongming Liu is an assistant Professor in the department of civil and environmental engineering. His research interests include fatigue and fracture analysis of metals and composite materials, probabilistic methods, computational mechanics, and risk management. He completed his PhD at Vanderbilt University, and obtained his Bachelors' and Masters' degrees from Tongji University in China. Dr. Liu is a member of ASCE and AIAA and serves on several technical committees on probabilistic methods and advanced materials.

BAYESIAN SOFTWARE HEALTH MANAGEMENT FOR AIRCRAFT GUIDANCE, NAVIGATION, AND CONTROL

Johann Schumann¹, Timmy Mbaya², Ole Mengshoel³

¹ SGT, Inc. NASA Ames, Moffett Field, CA 94035
Johann.M.Schumann@nasa.gov

² University of Massachusetts, Boston
timstim@mail.com

³ Carnegie Mellon University, NASA Ames, Moffett Field, CA 94035
Ole.Mengshoel@sv.cmu.edu

ABSTRACT

Modern aircraft—both piloted fly-by-wire commercial aircraft as well as UAVs—more and more depend on highly complex safety critical software systems with many sensors and computer-controlled actuators. Despite careful design and V&V of the software, severe incidents have happened due to malfunctioning software.

In this paper, we discuss the use of Bayesian networks to monitor the health of the on-board software and sensor system, and to perform advanced on-board diagnostic reasoning. We focus on the development of reliable and robust health models for combined software and sensor systems, with application to guidance, navigation, and control (GN&C). Our Bayesian network-based approach is illustrated for a simplified GN&C system implemented using the open source real-time operating system OSEK/Trampoline. We show, using scenarios with injected faults, that our approach is able to detect and diagnose faults in software and sensor systems.

1. INTRODUCTION

Modern aircraft depend increasingly on the reliable operation of complex, yet highly safety-critical software systems. Fly-by-wire commercial aircraft and UAVs are fully controlled by software. Failures in the software or a problematic software-hardware interaction can have disastrous consequences.

Although on-board diagnostic systems nowadays exist for most aircraft (hardware) subsystems, they are mainly working independently from each other and are not capable of reliably determining the root cause or causes of failures, in particular when software failures are to blame. Clearly, a powerful FDIR (Fault Detection, Isolation, Recovery) or ISHM

(Integrated System Health Management) system for software has a great potential for ensuring safety and operational reliability of aircraft and UAVs. This is particularly true, since many software problems do not directly manifest themselves but rather exhibit *emergent behavior*. For example, when the F-22 Raptors crossed the international date line, a software problem in the guidance, navigation, and control (GN&C) system did not only shut down that safety-critical component but also brought down communications, so the F-22s had to be guided back to Hawaii using visual flight rules.¹

An on-board software health management (SWHM) system monitors the flight-critical software while it is in operation, and thus is able to detect faults, such as the F-22 problems, as soon as they occur. In particular, an SWHM system

- *monitors the behavior of the software and interacting hardware during system operation.* Information about operational status, signal quality, quality of computation, reported errors, etc., is collected and processed on-board. Since many software faults are caused by problematic hardware/software interactions, status information about software components must be collected and processed, in addition to that for hardware.
- *performs diagnostic reasoning in order to identify the most likely root cause(s) for the fault(s).* This diagnostic capability is extremely important. In particular, for UAVs, the available bandwidth for telemetry is severely limited; a “dump” of the system state and analysis by the ground crew in case of a problem is not possible.

For manned aircraft, an SWHM can reduce the pilot’s workload substantially. With a traditional on-board diagnostic system, the pilot can get swamped by diagnostic errors and warnings coming from many different subsystems. Recently, when one of the engines exploded on a

Johann Schumann et.al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

¹<http://www.af.mil/news/story.asp?storyID=123041567>

Qantas A380, the pilot had to sort through literally hundreds of diagnostic messages in order to find out what happened. In addition, several diagnostic messages contradicted each other.²

In this paper, we describe our approach of using Bayesian networks as the modeling and reasoning paradigm to achieve SWHM. With a properly developed Bayesian network, detection of faults and reasoning about root causes can be performed in a principled way. Also, a proper probabilistic treatment of the diagnosis process, as we accomplish with our Bayesian approach (Pearl, 1988; Darwiche, 2009), can not only merge information from multiple sources but also provide a posterior distribution for the diagnosis and thus provide a metric for the quality of this result. We note that this approach has been very successful for electrical power system diagnosis (Ricks & Mengshoel, 2009, 2010; Mengshoel et al., 2010).

It is obvious that an SWHM system that is supposed to operate on-board an aircraft, in an embedded environment, must satisfy important properties: first, the implementation of the SWHM must have a small memory and computational footprint and must be certifiable. Second, the SWHM should exhibit a low number of false positives and false negatives. False alarms (false positives) can produce nuisance signals; missed adverse events (false negatives) can be a safety hazard. Our approach of using SWHM models, that have been compiled into arithmetic circuits, are amenable to V&V (Schumann, Mengshoel, & Mbaya, 2011).

The remainder of the paper is structured as follows: Section 2. introduces Bayesian networks and how they can be used for general diagnostics. In Section 3. we demonstrate our approach to software health management with Bayesian networks and discuss how Bayesian SWHM models can be constructed. Section 4. illustrates our SHWM approach with a detailed example. We briefly describe the demonstration architecture and the example scenario, discuss the use of a Bayesian health model to diagnose such scenarios, and present simulation results. Finally, in Section 5. we conclude and identify future work.

2. BAYESIAN NETWORKS

Bayesian networks (BNs) represent multivariate probability distributions and are used for reasoning and learning under uncertainty (Pearl, 1988). They are often used to model systems of a (partly) probabilistic nature. Roughly speaking, random variables are represented as nodes in a directed acyclic graph (DAG), while conditional dependencies between variables are represented as graph edges (see Figure 1 for an example). A key point is that a BN, whose graph structure often

reflects a domain's causal structure, is a compact representation of a joint probability table if the DAG is relatively sparse. In a discrete BN (as we are using for SWHM), each random variable (or node) has a finite number of states and is parameterized by a conditional probability table (CPT).

During system operation, observations about the software and system (e.g., monitoring signals and commands) are mapped into states of nodes in the BN. Various probabilistic queries can be formulated based on the assertion of these observations to yield predictions or diagnoses for the system. Common BN queries of interest include computing posterior probabilities and finding the most probable explanation (MPE). For example, an observation about abnormal behavior of a software component could, by computing the MPE, be used to identify one or more components that are most likely in faulty states.

Different BN inference algorithms can be used to answer the queries. These algorithms include join tree propagation (Lauritzen & Spiegelhalter, 1988; Jensen, Lauritzen, & Olesen, 1990; Shenoy, 1989), conditioning (Darwiche, 2001), variable elimination (Li & D'Ambrosio, 1994; Zhang & Poole, 1996), and arithmetic circuit evaluation (Darwiche, 2003; Chavira & Darwiche, 2007). In resource-bounded systems, including real-time avionics systems, there is a strong need to align the resource consumption of diagnostic computation with resource bounds (Musliner et al., 1995; Mengshoel, 2007) while also providing predictable real-time performance. The compilation approach—which includes join tree propagation and arithmetic circuit evaluation—is attractive in such resource-bounded systems.

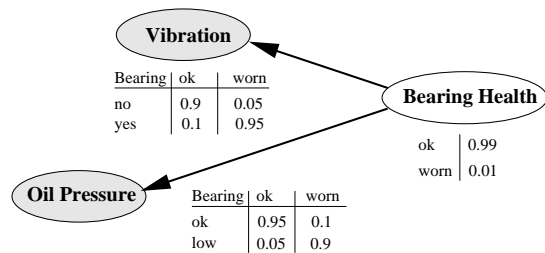


Figure 1. Simple Bayesian network. CPT tables are shown near each node.

Let us consider a very simple example of a Bayesian network (Figure 1) as it could be used in diagnostics. We have a node Bearing Health (*BH*) representing the health of a ball bearing in a diesel engine, a sensor node Vibration (*V*) representing whether vibration is measured or not, and a node Oil Pressure (*OP*) representing oil pressure. Clearly, the sensor readings depend on the health status of the ball bearing, and this is reflected by the directed edges. The degrees of influence are defined in the two CPTs depicted next to the sensor nodes. For example, if there is vibration, the probability that $p(BH = \sim \text{ok})$ increases. To obtain the health of the ball bearing, we input (or clamp) the states of the BN

²<http://www.aerosocietychannel.com/aerospace-insight/2010/12/exclusive-qantas-qf32-flight-from-the-cockpit/>

sensor nodes and compute the posterior distribution (or belief) over BH . The prior distribution of failure, as reflected in the CPT shown next to BH , is also taken into account in this calculation.

Our example network in Figure 1 represents the joint probability $p(BH, V, OP)$ and is shown in Table 1. For simplicity, we replace all CPT entries with θ_x (i.e., $\theta_{ok} \leftrightarrow BH$ is ok, and $\theta_{\sim ok} \leftrightarrow BH$ is worn). Let λ_i indicate whether evidence of a specific state is observed (i.e., $\lambda_v = 1$ means evidence of vibration is observed, and $\lambda_v = 0$ means no evidence of vibration is observed). The probability distribution $p(BH, V, OP)$ captured by the Bayesian network above is shown in Table 1.

BH	V	OP	$p(BH, V, OP)$
ok	v	op	$\lambda_{ok} \lambda_v \lambda_{op} \theta_{v ok} \theta_{ok} \theta_{op ok}$
ok	v	$\sim op$	$\lambda_{ok} \lambda_v \lambda_{\sim op} \theta_{v ok} \theta_{ok} \theta_{\sim op ok}$
ok	$\sim v$	$\sim op$	$\lambda_{ok} \lambda_{\sim v} \lambda_{\sim op} \theta_{\sim v ok} \theta_{ok} \theta_{\sim op ok}$
ok	$\sim v$	op	$\lambda_{ok} \lambda_{\sim v} \lambda_{op} \theta_{\sim v ok} \theta_{ok} \theta_{op ok}$
$\sim ok$	v	op	$\lambda_{\sim ok} \lambda_v \lambda_{op} \theta_{v \sim ok} \theta_{\sim ok} \theta_{op \sim ok}$
$\sim ok$	$\sim v$	op	$\lambda_{\sim ok} \lambda_{\sim v} \lambda_{op} \theta_{\sim v \sim ok} \theta_{\sim ok} \theta_{op \sim ok}$
$\sim ok$	v	$\sim op$	$\lambda_{\sim ok} \lambda_v \lambda_{\sim op} \theta_{v \sim ok} \theta_{\sim ok} \theta_{\sim op \sim ok}$
$\sim ok$	$\sim v$	$\sim op$	$\lambda_{\sim ok} \lambda_{\sim v} \lambda_{\sim op} \theta_{\sim v \sim ok} \theta_{\sim ok} \theta_{\sim op \sim ok}$

Table 1. Probability distribution for $p(BH, V, OP)$.

According to this joint probability distribution table, the first row ($\lambda_{ok} \lambda_v \lambda_{op} \theta_{v|ok} \theta_{ok} \theta_{op|ok}$) is representing the probability that the health of the ball bearing is okay ($\lambda_{ok} = 1$), and that vibrations and good oil pressure are observed (λ_v and $\lambda_{op} = 1$) would be 9.4% indicating a very low degree of belief in such a state. Given the corresponding numerical CPT entries this number is calculated as $\theta_{v|ok} \theta_{ok} \theta_{op|ok} = 0.1 * 0.99 * 0.95 = 0.09405$. On the other hand, the fourth row ($\lambda_{ok} \lambda_{\sim v} \lambda_{op} \theta_{\sim v|ok} \theta_{ok} \theta_{op|ok}$) representing the probability that the ball bearing is okay ($\lambda_{ok} = 1$), there is no vibrations and good oil pressure ($\lambda_{\sim v}$ and $\lambda_{op} = 1$) is much higher (85%) as follows: $\theta_{\sim v|ok} \theta_{ok} \theta_{op|ok} = 0.9 * 0.99 * 0.95 = 0.84645$.

Posterior marginals can be computed from the joint distribution:

$$p(BH, V, OP) = \prod_{\theta_{s|x}} \theta_{s|x} \prod_{\lambda_s} \lambda_s$$

where $\theta_{s|x}$ indicates a state's conditional probability and λ_s indicates whether or not state s is observed. Here, θ variables are known as variables, λ variables as indicators.

Summing all individual joint distribution entries yields a multi-linear function—at the core of arithmetic circuit evaluation—referred to as the *network polynomial* f

(Darwiche, 2009):

$$f = \lambda_{ok} \lambda_v \lambda_{op} \theta_{v|ok} \theta_{ok} \theta_{op|ok} + \lambda_{ok} \lambda_v \lambda_{\sim op} \theta_{v|ok} \theta_{ok} \theta_{\sim op|ok} + \lambda_{ok} \lambda_{\sim v} \lambda_{\sim op} \theta_{\sim v|ok} \theta_{ok} \theta_{\sim op|ok} + \lambda_{ok} \lambda_{\sim v} \lambda_{op} \theta_{\sim v|ok} \theta_{ok} \theta_{op|ok} + \lambda_{\sim ok} \lambda_v \lambda_{op} \theta_{v|\sim ok} \theta_{\sim ok} \theta_{op|\sim ok} + \lambda_{\sim ok} \lambda_{\sim v} \lambda_{op} \theta_{\sim v|\sim ok} \theta_{\sim ok} \theta_{op|\sim ok} + \lambda_{\sim ok} \lambda_v \lambda_{\sim op} \theta_{v|\sim ok} \theta_{\sim ok} \theta_{\sim op|\sim ok} + \lambda_{\sim ok} \lambda_{\sim v} \lambda_{\sim op} \theta_{\sim v|\sim ok} \theta_{\sim ok} \theta_{\sim op|\sim ok},$$

or in other words

$$f = \sum_E \prod_{\theta_{s|x}} \theta_{s|x} \prod_{\lambda_s} \lambda_s$$

where E indicates evidence of a network instantiation.

An arithmetic is a compact representation of a network polynomial. An arithmetic circuit (AC) is a directed acyclic graph (DAG) in which leaf nodes represent variables (parameters and indicators) while other nodes represent addition and multiplication operators. Size, in terms of number of AC edges, is a measure of complexity of inference. Unlike treewidth, another complexity measure, AC size can take network parameters (such as determinism and local structure) into account.

Answers to probabilistic queries, including marginals and MPE, are computed using algorithms that operate directly on the arithmetic circuit. A bottom-up pass over the circuit, from input to output, evaluates the probability of a particular evidence setting (or clamping of λ parameters) on the state of the network. And a top-down pass over the circuit, from output to input, computes partial derivatives. From these partial derivatives one can compute many marginal probabilities, provide information about how change in a specific node affects the whole network (sensitivity analysis), and perform MPE computation (Darwiche, 2009).

3. BAYESIAN NETWORKS FOR SOFTWARE HEALTH MANAGEMENT

At a first glance, the SWHM does look very similar to a traditional integrated vehicle health management system (IVHM): sensor signals are interpreted to detect and identify any faults, which are then reported. Such FDIR systems are nowadays commonplace in the aircraft and for other complex machinery. It seems like it would be straight-forward to attach a software to be monitored (host software) to such an FDIR. However, there are several critical differences between FDIR for hardware and software health management. Most prominently, many software faults do not develop gradually over time (e.g., like an oil leak); rather they occur instantaneously. Whereas some of the software faults directly impact the current software module (e.g., when a division-by-zero is detected), there are situations where the effects of a software fault manifest themselves in an entirely different subsystem, as discussed in the F-22 example above. For this reason, and

the fact that many software problems occur due to problematic SW/HW interactions, both software and hardware must be monitored in an integrated fashion.

Based upon requirements as laid out in Section 1., we are using Bayesian networks to develop SWHM models. On a top-level, data from software and hardware sensors are presented to the nodes of the Bayesian network, which in turn performs its reasoning (i.e., updating the internal health and status nodes) and returns information about the health of the software (or specific components thereof). The information about the health of the software is extracted from the posterior distribution, specifically from health nodes. In our modeling approach, we chose to use Bayesian networks, which do not reason about temporal sequences (i.e., dynamic Bayesian networks) because of their complexity. Therefore, all sensor data, which are usually time series, must undergo a pre-processing step, where certain (scalar) *features* are extracted. These values are then discretized into symbolic states (e.g., “low”, “high”) or normalized numeric values before presented to the Bayesian health model (Section 3.3).

3.1 Bayesian SWHM

3.1.1 Nodes

Our Bayesian SWHM models are set up using several kinds of nodes. Please note that all nodes are discrete, i.e., each node has a finite number of mutually exclusive and exhaustive states.

CMD node C Signals sent to these nodes are handled as ground truth and are used to indicate commands, actions, modes or other (known) states. For example, a node `Write_File_System` represents an action, which eventually will write some data into the file system, has been commanded. For our reasoning it is assumed that this action is in fact happening.³ The CMD nodes are root nodes (no incoming edges). During the execution of the SWHM, these nodes are always directly connected (clamped) to the appropriate command signals.

SENSOR node S A sensor node S is an input node similar to the CMD node. The data fed into this node are sensor data, i.e., measurements that have been obtained from monitoring the software or the hardware. Thus, this signal is not necessarily correct. It can be noisy or wrong altogether. Therefore, a sensor node is typically connected with a health node, that describes the health status of the sensor node.

HEALTH node H The health nodes are nodes that reflect the health status of a sensor or component. Their posterior probabilities comprise the output of an SWHM

³If there is a reason that this command signal is not reliable, the command node C is used in combination with a H node to impact state U as further discussed below. Alternatively, one might consider using a sensor node instead.

model. A health node can be binary (with states, say, `ok` or `bad`), or can have more states that reflect health status at a more fine-grained level. Health nodes are usually connected to sensor and status nodes.

STATUS node U A status node reflects the (unobservable) status of the software component or subsystem.

BEHAVIOR node B Behavior nodes connect sensor, command, and status nodes and are used to recognize certain behavioral patterns. The status of these nodes is also unobservable, similar to the status nodes. However, usually no health node is attached to the behavioral nodes.

3.1.2 Edges

The following informal way to think about edges in Bayesian networks are useful for knowledge engineering purposes: An edge (arrow) from node C to node E indicates that the state of C has a (causal) influence on the state of E .

Suppose that S is a software signal (e.g., within the aircraft controller) that leads into an input port I of the controller. Let us assume that we want S being 1 to cause C to be 1 as well. Failure mechanisms are represented by introduced a health node H . In our example, we would introduce a node H and let it be a (second) parent of I . More generally, the types of influences typically seen in the SWHM BNs are as follows:

$\{H, C\} \rightarrow U$ represents how state U may be commanded through command C , which may not always work as indicated. This is reflected by the health H of the command mechanism’s influence on the state.

$\{C\} \rightarrow U$ represents how state U may be changed through command C ; the health of the command mechanism is not explicitly represented. Instead, imperfections in the command mechanism can be represented in the CPT of U .

$\{H, U\} \rightarrow S$ represents the influence of system status U on a sensor S , which may also fail as reflected in H . We use a sensor to better understand what is happening in a system. However, the sensor might give noisy readings; the level of noise is reflected in the CPT of S .

$\{H\} \rightarrow S$ represents a direct influence of system health H on a sensor S , without modeling of state (as is done in the $\{H, U\} \rightarrow S$ pattern). An example of this approach is given in Figure 1.

$\{U\} \rightarrow S$ represents how system status U influences a sensor S . Sensor noise and failure can both be rolled into the CPT of S .

Table 2 shows the CPT for the last case. Here, we consider the status of a file system (FS). The file system can be `empty`, `full`, or filled to more than 95% (`full95`). If more space is available, its state is labeled `ok`. This (unobservable) state is observed by a software sensor, which measures the current

capacity of the file system (FC). Because this sensor might fail, a health node (FH) indicates the health of FC sensor as ok or bad.

Because the sensor node FC has two parents (status node FS and health node FH), the CPT table is 3-dimensional. Table 2 flattens out this information: the rows correspond to the states of the sensor node (1st group for healthy sensor, 2nd group for bad sensor). The rightmost four columns refer to the states of the FS node. In this particular example, a file system sensor, which is not working properly will not report if the file system is almost full or full. Such a bad sensor will only report empty or ok. This is reflected by the zero-entries in the lower right corner of the CPT.

FS	FH	$p(FC FH, FS)$			
		empty	ok	full195	full
empty	ok	0.88	0.05	0.01	0.01
ok	ok	0.1	0.6	0.2	0.1
full195	ok	0	0.2	0.7	0.1
full	ok	0	0	0	1
empty	bad	0.9	0.1	0	0
ok	bad	0.1	0.9	0	0
full195	bad	0.5	0.5	0	0
full	bad	0.5	0.5	0	0

Table 2. CPT table for $p(FC|FH, FS)$.

3.1.3 Developing Conditional Probability Tables (CPTs)

The CPT entries are set based on a priori and empirical knowledge of a system’s components and their interactions (Ricks & Mengshoel, 2009; Mengshoel et al., 2010). This knowledge may come from different sources, including (but not restricted to) system schematics, source code, analysis of prior software failures, and system testing. As far as a system’s individual components, mean-time-to-failure statistics are known for certain hardware components, however similar statistics are well-established for software. Consequently, further research is needed to determine the prior distribution for health states, including bugs, for a broad range of software components. As far as an interaction between a system’s components, CPT entries can also be obtained from understanding component interactions, a priori, or testing how different components impact each other. As an example, consider a testbed like NASA’s advanced diagnostic and prognostic testbed (ADAPT) (Poll et al., 2007), which provides both schematics and testing opportunities. Using a testing approach, one may inject specific states into the navigation system and record the impact on states of the guidance system, and perform statistical analysis, in order to guide the development of CPT entries for the guidance system. Setting of software component CPTs to reflect their interactions with hardware can be conducted in a similar way. Clearly, the well-known limitation of brute-force testing apply, and when this

occurs one needs to utilize design artifacts, system schematics, source code, and other sources of knowledge about component interactions.

3.2 Software Sensors

Information that is needed to reason about software health must be extracted from the software itself and all components that interact with the software, i.e., hardware sensors, actuators, the operating system, middleware, and the computer hardware. Different software sensors provide information about the software on different levels of granularity and abstraction. Table 3 gives an impression of the various layers of information extraction.

Only if information is available on different levels, the SWHM gets a reasonably complete picture of the current situation, which is an enabling factor for fault detection and identification. Information directly extracted from the software (Table 3) provide very detailed and timely information. However, this information might not be sufficient to identify a failure. For example, the aircraft control task might be working properly (i.e., no faults show up from the software sensors). However, some other task might consume too many resources (e.g., CPU time, memory, etc.), which in turn can lead to failures related to the control task. We therefore extract a multitude of different, usually readily available information about the software.

Software	
errors	flagged errors and exceptions
memsize	used memory
quality	signal quality
reset	filter reset (for navigation)
Software Intent	
fs_write	intent to write to FS
fork	intent to create new process(es)
malloc	intent to allocate memory
use_msg	intent to use message queues
use_sem	using semaphores
use_recursion	using recursion
Operating system	
cpu	CPU load
n_proc	number of processes
m_free	available memory
d_free	percentage of free disk space
shm	size of available shared memory
sema	information about semaphores
realtime	missed deadlines
n_intr	number of interrupts
l_msgqueue	length of message queues

Table 3. SWHM informations sources

3.3 Preprocessing of Software and Hardware Sensor Data

The main goals of preprocessing are to extract important information from the (large amounts of) temporal sensor data and to discretize continuous sensor data to be used with our discrete SWHM models. For example, the sensor for the file system (*FC*) has the states *empty*, *ok*, *full195*, *full*. Preprocessing steps, which extract temporal features from raw sensor data, enable us to perform temporal reasoning without having to use a dynamic Bayesian network (DBN). This is a very prominent conceptual decision. By giving up the ability to do full temporal reasoning by means of DBNs, which may be complex in design and execution, we are able to use much simpler static health models and handle the temporal aspects during preprocessing.

In particular, we use the following preprocessing techniques (which can also be combined):

discretization A continuous value is discretized using a number of monotonically increasing thresholds. For example, Table 4 shows the discretization for file system sensor *FC*.

min/max/average The minimal/maximal value or the mean of the entire available time series is taken.

moving min/max/average A moving min/max/mean value (with a selectable window size) is taken. In contrast to the features above, we only consider the last few seconds of the signal.

sum (integral) The sum (integral) of the sensor value is taken. For example, the sum of “bytes-written-to-file-system” (per time unit) approximates the amount of data in the file system (assuming nothing is being deleted).

temporal Temporal states of sensor signals can be extracted, e.g., time difference between event *A* and event *B*.

time-series analysis Kalman filters can be used to correlate signals against a model. Residual errors then can be used as sensor states (e.g., close-to-model, small-deviation, large-deviation). Fast Fourier transformation (FFT) can be used to detect cyclic events, e.g., vibrations or oscillations.

Percentage (df)	State
$0 \leq df < 5\%$	<i>empty</i>
$5 \leq df < 80\%$	<i>ok</i>
$80 \leq df < 95\%$	<i>full195</i>
$95 \leq df$	<i>full</i>

Table 4. Discretization into states (right) by means of thresholds (left).

4. DEMONSTRATION EXAMPLE

4.1 System Architecture

For demonstration purposes, we have implemented a simple system architecture on a platform that reflects real-time embedded execution typical of aircraft and satellite systems. Trampoline,⁴ an emulator for the OSEK⁵ real-time operating system (RTOS), is used as a platform rather than other RTOSes more established in the aerospace industry (such as Wind River’s VxWorks or GreenHills’ INTEGRITY). OSEK is easily available, widely used for embedded control systems in the automotive industry, and its capabilities were sufficient for the purpose of our experiments.

The basic system architecture (Figure 2) for running SWHM experiments consists of the OSEK RTOS, which runs a number of tasks or processes at a fixed schedule. For this simple SWHM demonstration system, (1) the simulation model of the plant is integrated as one of the OSEK tasks, and (2) hardware actuators and sensors are not modelled in detail, which would have required drivers and interrupts routines. Despite its simplicity, this architecture is sufficient to run a simple simulation of the aircraft and the GN&C software in a real-time environment (fixed time slots, fixed memory, inter-process communication, shared resources).

The software health management executive, including preprocessing, is executed as a separate OSEK task. It reads software and sensor data, performs preprocessing and provides the data as evidence to the sensor nodes of the (compiled) Bayesian network. The reasoning process then yields the posterior probabilities of the health nodes.

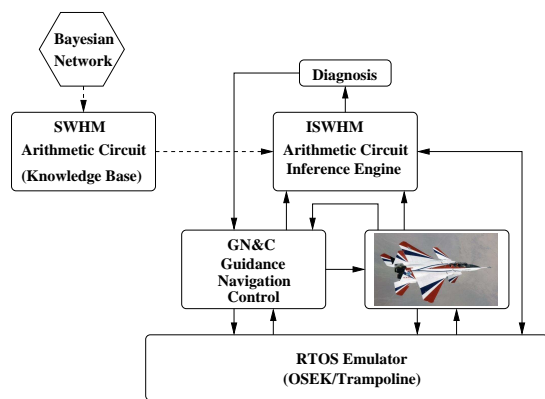


Figure 2. Demonstration system architecture. The Bayesian network model is compiled (before deployment) into an arithmetic circuit representing the knowledge base. The real-time operating system schedules three tasks: the controller, the plant, and the SWHM inference engine.

⁴<http://trampoline.rts-software.org/>

⁵<http://www.osek-vdx.org/>

4.2 Example Scenario

An experimental scenario aimed at the study of faults related to file systems, inspired by the Mars rover SPIRIT reboot cycle incident (Adler, 2006), has been implemented using the system architecture. A short time after landing, the Mars rover SPIRIT encountered repeated reboots, because a fault during the booting process caused a subsequent reboot. According to reports (Adler, 2006), an on-board file system for intermediate data storage caused the problem. After this storage was filled up, the boot process failed while trying to access that file system. The problem could be detected on the ground and was resolved successfully.

In a more general setting, this scenario is dealing with bad interaction due to scarce resources, and delays during access. Even if no errors show up, a blocking write access to a file system that is almost full, or the delivery of a message through a lengthy message queue, can in the worst case cause severe problems and emerging behavior.

For the purpose of demonstration, we designed a flawed software architecture with a global message queue that buffers all control signals and logs them in the file system (blocking) before forwarding them (Figure 3). This message queue is also used to transport image data from an on-board camera (e.g., for UAV) to the radio transmitter. The relevant software components of this simple architecture are: GN&C, message queue, logging to file system, camera, transmitter, and plant. On-board camera and transmitter are shown in Figure 3 but not used in the experiments described in this paper.

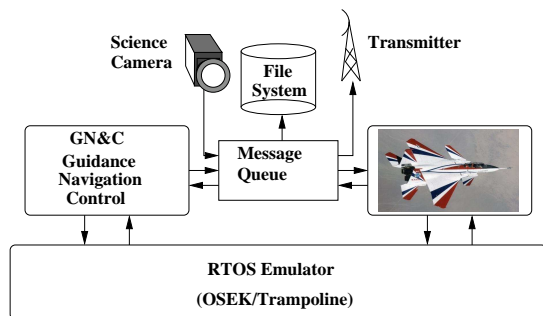


Figure 3. Software architecture for file system related fault scenarios, diagnosed using SWHM system.

Here, we are running the following scenario: the file system is initially set to almost full. Subsequent control messages, which are being logged, might stay longer in the message queue, because the blocking write into an almost full file system takes substantial time. This situation potentially causes overflow of the message queue or leads to loss of messages. However, even a small delay (i.e., a control message is not processed within its allotted time frame, but one or more timeframes later) can cause *oscillation* of the entire aircraft. This oscillation, similar to PIO (pilot induced oscillation) can lead

to dangerous situations or even loss of the aircraft.

In this scenario, the software problem does not manifest itself within the software system (e.g., in form of errors or exceptions). Rather, the overall behavior of the aircraft is impacted in a non-obvious way.

Other possible scenarios with this setup, to be diagnosed by the SWHM task, are:

- The pilot's or autopilot's stick commands are delayed, which again results in oscillations of the aircraft.
- Non-matching I/O signal transmit/read/processing rates between control stick and actuators result in plant oscillations whose root causes are to be disambiguated.
- An unexpectedly large feed from the on-board camera (potentially combined with a temporary low transmission bandwidth) can cause the message queue to overflow, which results in delays or dropped messages with similar effects as discussed above.
- The controller and the science camera compete for the message queue, which could (when not implemented correctly) cause message drops or even deadlocks.

With our SWHM, the observed problem (oscillation) should be detected properly and traced back to the root cause(s).

4.3 The SWHM Model

A Bayesian SWHM model for this architecture was designed using the SamIam tool.⁶ A modular BN design approach was attempted by first designing the SWHM model for the basic system including relevant nodes such as—in the aircraft case—the pitch-up and pitch-down command nodes. The pitch status nodes, the fuel status node, and the software, pitch, and acceleration health nodes were introduced. Other subnetworks were then added to this core Bayesian network to obtain the complete SHWM model for the specific architecture used for SWHM experiments. The relevant nodes of the subnetwork module added for SWHM experiment with file system related faults are shown in Figure 4.

The `Write_File_System` command node indicates whether a write to the file system is being executed. The health nodes for the file system and the message queue reflect the probabilities that they might malfunction. The status nodes for the file system and the message queue represent their unobservable states, while their sensor nodes reflect sensor readings after preprocessing.

The only non-standard software sensor node in this SWHM model is a sensor to detect oscillations or vibrations. A fast Fourier transform (FFT) performs a simple time-series analysis on major aircraft signals (e.g., accelerations or pqr rates). With such a sensor, low-frequency oscillations (e.g., pilot-induced oscillations (PIO)) or vibrations (with a higher frequency) can be detected and fed into the SWHM model. The

⁶<http://reasoning.cs.ucla.edu/samiam>

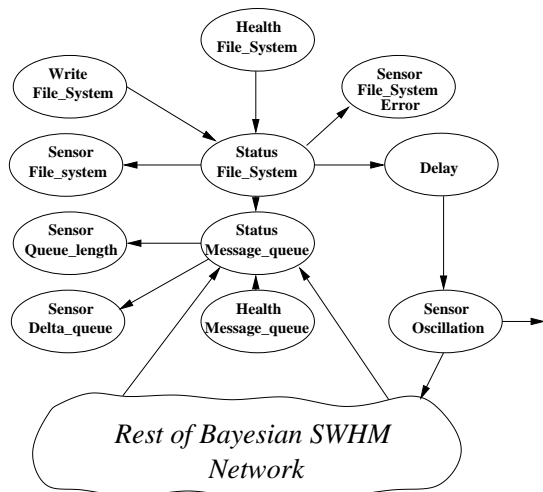


Figure 4. Partial Bayesian network for diagnosing faults potentially taking place in the software architecture shown in Figure 3.

SWHM reasoning then tries to perform a disambiguation on the cause of the oscillation or vibration.

This Bayesian network is compiled into an arithmetic circuit, which is integrated with the rest of the system as shown in Figure 2.

4.4 Results

Analysis of experimental runs with this architecture indicated that the system undergoing SHWM runs fine in the nominal case (Figure 5). However, the SWHM inference engine was instrumental in pointing toward the root cause of oscillations when pitch-up and pitch-down commands to the aircraft plant are affected by faults originating in the file system, causing the aircraft to oscillate up and down rather than maintain the desired altitude. For the purpose of our experiments, the file system was set to almost full at the start of the run, and as the system runs and controls are issued and logged, delays in executions start taking place at time $t = 30s$ (Figure 6) but no software errors are flagged. Eventually, altitude oscillations are detected by a fast Fourier transform performed on the altitude sensor readings shown in the middle panel of Figure 6. The bottom panel indicates that when the fast Fourier transform eventually detects oscillations around $t = 100s$, the SWHM infers that the posterior probability of good software health drops substantially, while the posteriors of good health of pitch and accelerometer systems are mostly high despite some transient lows. This indicates a low degree of belief in the good health of the software and that the most likely cause for a state with oscillations would be a software fault. For the purpose of this experiment, no additional pilot inputs were assumed.

SHWM can also be instrumental in disambiguating the root

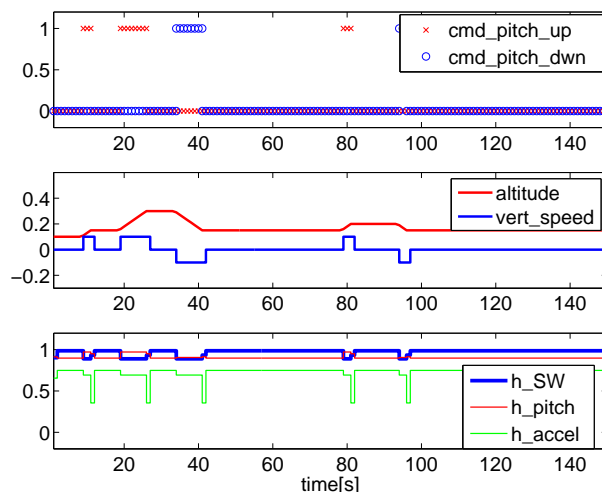


Figure 5. Temporal trace for the nominal case of file system based scenarios. The top panel shows pitch up and down commands to the aircraft. The middle panel shows the readings of altitude and vertical speed. The bottom panel shows the degree of belief in the good health of the accelerometer sensor (h_accel , green), of the pitch signal (h_pitch , red), and of the software (h_SW , thick blue line).

cause of oscillations when we add a pilot input node connected to the oscillation detection fast Fourier transform sensor node. The SWHM reasoner can then disambiguate the diagnosis by evaluating whether the fault is due to PIO or a software problem.

The SWHM models, which we have presented here are able to recognize and disambiguate known failure classes. In general, the handling of emergent behavior, i.e., the occurrence of events or failures that have not been considered or modeled, is an important task for a system-wide health management system. Such failures can occur if the system is operated in a new environment, or due to unexpected interaction between components.

Our SWHM approach can—albeit with some restrictions—detect and diagnose emergent behavior. If we model the software behavior using safety and performance requirements (in addition to specific pre-analyzed) failure modes, emergent behavior, which manifests itself adversely by violating safety requirements or lowers performance, can be detected and diagnosed.

In our experimental setting, relevant performance or safety requirements could be: no vibrations or oscillations should occur, and a smooth flight path without specific pilot input should not require substantial actuator activation. With the existing sensors and the reasoning capabilities of the Bayesian network, the failure scenario discussed above would raise an alarm due to the violation of these requirements.

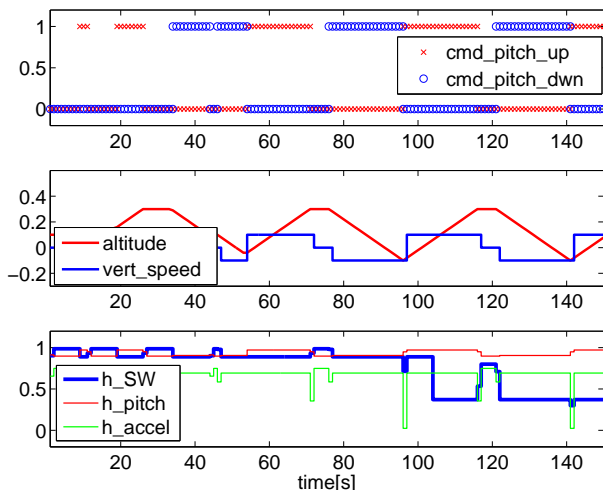


Figure 6. Temporal trace for a file system related fault scenario resulting in oscillations. The SWHM inference engine’s evaluation outputs show that the degree of belief in the good health of the system’s software (blue in bottom panel) substantially drops when oscillations are eventually detected by a fast Fourier transform at about $t = 100s$, after overflow of the file system resulted in delayed pitch up and pitch down command signals from the controller. Readings from the altitude sensor (blue in middle panel) show oscillating altitude starting at about $t = 30s$.

5. CONCLUSIONS

Software plays an important and increasing role in aircraft. Unfortunately, software (like hardware) can fail in spite of extensive verification and validation efforts. This obviously raises safety concerns.

In this paper, we discussed a software health management (SWHM) approach to tackle problems associated with software bugs and failures. The key idea is that an SWHM system can help to perform on-board fault detection and diagnosis on aircraft.

We have illustrated the SWHM concept using Bayesian networks, which can be used to model software as well as interfacing hardware sensors, and fuse information from different layers of the hardware-software stack. Bayesian network system health models, compiled to arithmetic circuits, are suitable for on-board execution in an embedded software environment.

Our Bayesian network-based SWHM approach is illustrated for a simplified aircraft guidance, navigation, and control (GN&C) system implemented using the OSEK embedded operating system. While OSEK is rather simple, it is We show, using scenarios with injected faults, that our approach is able to detect and diagnose non-trivial software faults.

In future work, we plan to investigate how the SWHM con-

cept can be extended to robustly handle unexpected and unmodeled failures, as well as how to more automatically generate SWHM Bayesian models based on information in artifacts including software engineering models, source code, as well as configuration and log files.

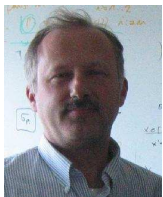
ACKNOWLEDGMENT

This work is supported by a NASA NRA grant NNX08AY50A “ISWHM: Tools and Techniques for Software and System Health Management”.

REFERENCES

- Adler, M. (2006). *The Planetary Society Blog: Spirit Sol 18 Anomaly*. Retrieved 02/2010, from <http://www.planetary.org/blog/article/00000702/>
- Chavira, M., & Darwiche, A. (2007). Compiling Bayesian Networks Using Variable Elimination. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07)* (p. 2443-2449). Hyderabad, India.
- Darwiche, A. (2001). Recursive conditioning. *Artificial Intelligence*, 126(1-2), 5-41.
- Darwiche, A. (2003). A Differential Approach to Inference in Bayesian Networks. *Journal of the ACM*, 50(3), 280–305.
- Darwiche, A. (2009). *Modeling and Reasoning with Bayesian Networks*. Cambridge, UK: Cambridge University Press.
- Jensen, F. V., Lauritzen, S. L., & Olesen, K. G. (1990). Bayesian Updating in Causal Probabilistic Networks by Local Computations. *SIAM Journal on Computing*, 4, 269–282.
- Lauritzen, S., & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society series B*, 50(2), 157–224.
- Li, Z., & D’Ambrosio, B. (1994). Efficient Inference in Bayes Nets as a Combinatorial Optimization Problem. *International Journal of Approximate Reasoning*, 11(1), 55–81.
- Mengshoel, O. J. (2007). Designing Resource-Bounded Reasoners using Bayesian Networks: System Health Monitoring and Diagnosis. In *Proceedings of the 18th International Workshop on Principles of Diagnosis (DX-07)* (pp. 330–337). Nashville, TN.
- Mengshoel, O. J., Chavira, M., Cascio, K., Poll, S., Darwiche, A., & Uckun, S. (2010). Probabilistic Model-Based Diagnosis: An Electrical Power System Case Study. *IEEE Trans. on Systems, Man, and Cybernetics*, 40(5), 874–885.

- Musliner, D., Hendler, J., Agrawala, A. K., Durfee, E., Strosnider, J. K., & Paul, C. J. (1995, January). The Challenges of Real-Time AI. *IEEE Computer*, 28, 58–66. Available from citeseer.comp.nus.edu.sg/article/musliner95challenges.html
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.
- Poll, S., Patterson-Hine, A., Camisa, J., Garcia, D., Hall, D., Lee, C., et al. (2007). Advanced Diagnostics and Prognostics Testbed. In *Proceedings of the 18th International Workshop on Principles of Diagnosis (DX-07)* (pp. 178–185). Nashville, TN.
- Ricks, B. W., & Mengshoel, O. J. (2009). Methods for Probabilistic Fault Diagnosis: An Electrical Power System Case Study. In *Proc. of Annual Conference of the PHM Society, 2009 (PHM-09)*. San Diego, CA.
- Ricks, B. W., & Mengshoel, O. J. (2010). Diagnosing Intermittent and Persistent Faults using Static Bayesian Networks. In *Proc. of the 21st International Workshop on Principles of Diagnosis (DX-10)*. Portland, OR.
- Schumann, J., Mengshoel, O., & Mbaya, T. (2011). Integrated Software and Sensor Health Management for Small Spacecraft. In *Proc. SMC-IT*. IEEE.
- Shenoy, P. P. (1989). A valuation-based language for expert systems. *International Journal of Approximate Reasoning*, 5(3), 383–411.
- Zhang, N. L., & Poole, D. (1996). Exploiting Causal Independence in Bayesian Network Inference. *Journal of Artificial Intelligence Research*, 5, 301–328. Available from citeseer.nj.nec.com/article/zhang96exploiting.html



Johann Schumann (PhD 1991, Dr. habil 2000, Munich, Germany) is Chief Scientist for Computational sciences with SGT, Inc. and working in the Robust Software Engineering Group at NASA Ames. He is engaged in research on software health management, verification and validation of health management systems, and data analysis for air traffic control. Dr. Schumann's general research interests focus on the application of formal and statistical methods to improve design and reliability of advanced safety- and security-critical software. Dr. Schumann is author of a book on theorem proving in software engineering and has published more than 90 articles on automated deduction and its applications, automatic program synthesis, software health management, and neural network oriented topics.



Ole J. Mengshoel received the B.S. degree from the Norwegian Institute of Technology, Trondheim, Norway, in 1989, and the Ph.D. degree from the University of Illinois at Urbana-Champaign, Illinois, in 1999, both in computer science. He is currently a Senior Systems Scientist with Carnegie Mellon University (CMU), Silicon Valley, CA, and affiliated with the Intelligent Systems Division, National Aeronautics and Space Administration (NASA) Ames Research Center, Moffett Field, CA. Prior to joining CMU, he was a Senior Scientist and Research Area Lead with US-RA/RIACS and a Research Scientist with the Decision Sciences Group, Rockwell Scientific, and Knowledge-Based Systems, SINTEF, Norway. His current research focuses on reasoning, machine learning, diagnosis, prognosis, and decision support under uncertainty - often using Bayesian networks and with aerospace applications of interest to NASA. He has published more than 50 papers and papers in journals, conference proceedings, and workshops. He is the holder of four U.S. patents. Dr. Mengshoel is a member of the Association for the Advancement of Artificial Intelligence, the Association for Computer Machinery, and IEEE.



Timmy Mbaya is a computer science student at the University of Massachusetts in Boston. His interests are Artificial Intelligence, probabilistic reasoning, and real-time embedded systems.

Commercialization of Prognostics Systems Leveraging Commercial Off-The-Shelf Instrumentation, Analysis, and Data Base Technologies

Preston Johnson

National Instruments, Austin, Texas, 78759, United States of America

preston.johnson@ni.com

ABSTRACT

There are many facets of a prognostics and health management system. Facets include data collection systems that monitor machine parameters; signal processing facilities that sort, analyze and extract features from collected data; pattern matching algorithms that work to identify machine degradation modes; database systems that organize, trend, compare and report information; communications that synchronize prognostic system information with business functions including plant operations; and finally visualization features that allow interested personnel the ability to view data, reports, and information from within the intranet or across the internet.

A prognostic system includes all of these facets, with details of each varying to match specific needs of specific machinery. To profitably commercialize a prognostic system, a generic yet flexible framework is adopted which allows customization of individual facets. Customization of one facet does not materially impact another.

This paper describes the framework, and provides case studies of successful implementation.

1. INTRODUCTION

The objective of a prognostic system is to predict the need for maintenance before serious equipment breakdown occurs and to predict the remaining useful life of the equipment components. A prognostics system should operate where possible on its own, to lessen the need for human involvement. This is a tall order for the prognostics systems developer. To ease the required efforts, commercial off the shelf (COTS) components can be used to allow more focus on prognostics algorithms and recommendation reporting.

This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Prognostics Systems have several components, commonly grouped into data acquisition, signal processing and analysis, and decision making, Figure 1. Figure 1 can be expanded to include communications, visualization, and database components.

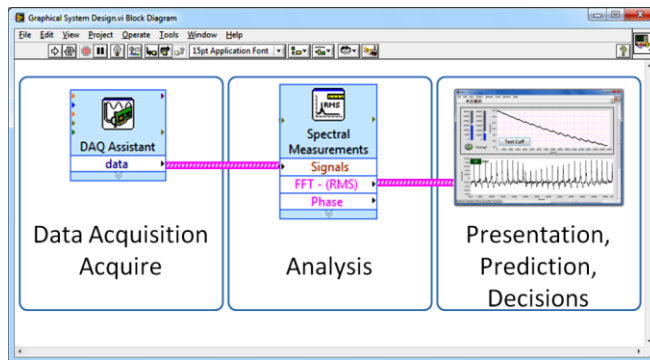


Figure 1: Basic components of prognostic system

To allow flexibility for analysis types, and machinery types, each component needs to be modular to the extent components can be easily interchanged. This interchangeability extends to hardware as well as software. The system needs to scale from small machines up to large machines, and from test cell applications down to portable systems and into on-line embedded systems. Finally, the on-line embedded system components need to be priced competitive to existing data collecting systems. In other words, constraints exist in hardware, software, and development tools in order to maximize modularity, cost and ease of customization. A framework of hardware and software components makes this commercialization possible, Figure 2.

From a cost perspective, it quickly becomes apparent that commercial off-the shelf (COTS) components provide the best cost model for the framework. With COTS, the prognostics systems designer minimizes electronic design as well as software design efforts. Instead, the designer is able to leverage development work already in play within the

COTS suppliers' organization. Further, COTS systems typically provide a faster validation cycle as much of the hardware and software components have validation certificates.

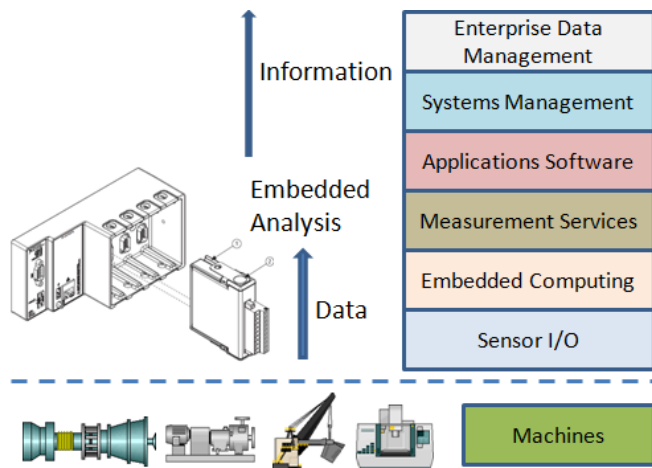


Figure 2: Modular prognostics architecture

This paper examines the prognostics architecture framework and its core components including COTS component options.

2. MODULAR ARCHITECTURES

There are several publications promoting a modular architecture for condition monitoring and prognostics systems. One such standard is the ISO 13374 standard, Figure 3, ISO (2003).

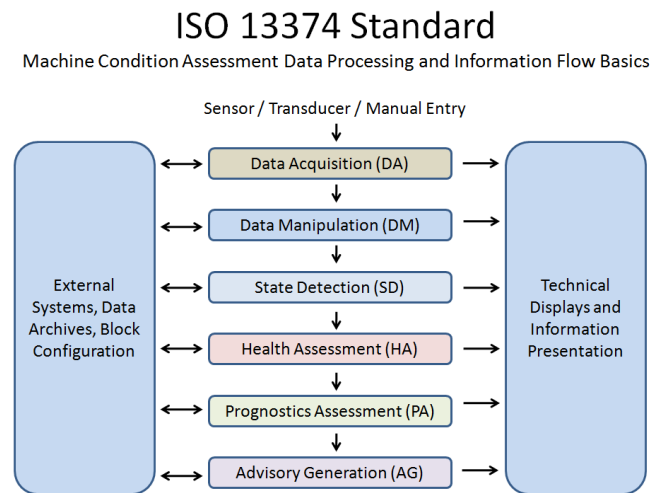


Figure 3: ISO 13374 condition monitoring standard

The ISO 13374 standard divides the condition monitoring and prognostics system into eight subsystems. These include data acquisition (DA) and data manipulation (DM)

or data acquisition and signal processing. The ISO 13374 standard also calls out state detection (SD). State detection is often defined as the determination of deviation from normal or healthy operation. It can also be defined as determining the operational state of the machine, for example high speed operation and low speed operation.

Three prognostic functions in the ISO 13374 standard include Health Assessment (HA), Prognostic Assessment (PA) and Advisory Generation (AG). These three blocks perform the hard work of diagnostics, prediction, and information delivery. The outer two blocks on the left and right of the six middle blocks further define data base storage and retrieval, as well as visualization including technical displays and reporting. When following this model, the prognostics developer can save costs by foregoing the need to design these architectures. Further, when following a defined standard, it is possible to mix and match components from multiple commercial suppliers, each of which may specialize in a specific component area.

The University of Cincinnati Intelligent Maintenance Center (IMS Center), for example, takes a unique approach in adapting the ISO 13374 model by adding data filtering and sorting during the data acquisition (DA) step in the process, Figure 4, Lee (2009). The University recommends sorting data into operating regimes.

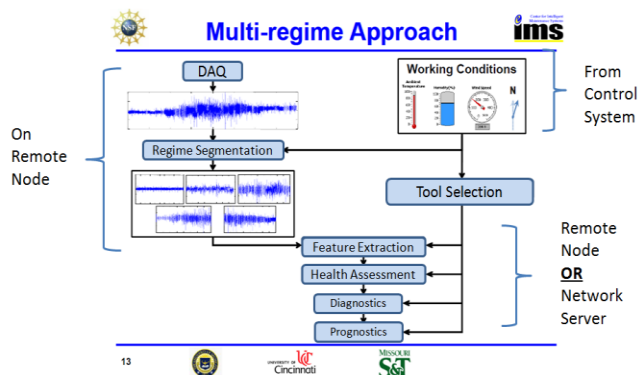


Figure 4: IMS Center multi-regime approach

Operating regimes are distinguished by speeds, loads, and even mechanical failure modes. These regimes are identified during the data collection process. Data is labeled with the regime name, for example speed and load. Downstream, categorization of information is made easy with regime tags made at the data collection point.

In either case, adaption of the ISO 13374 model to specific implementation provides modularity, flexibility, and promises to lower costs.

3. DATA ACQUISITION COMPONENT OF PROGNOSTICS

The data acquisition component (Figure 3) of the prognostic system has the important role of correctly recording sensory information from the machine for downstream analysis and decision making. Data acquisition typically consists of sensors, signal conditioning, and an analog to digital converter. Sensors may include digital sensors as well as analog sensors. Analog sensors commonly used in mechanical system prognostics include temperature, electrical power, strain, speed, and vibration. Often, electrical power, strain, speed, and vibration sensors need fast and high resolution analog to digital converters to transform the analog output of the sensor into a digital format the embedded data acquisition can process, store, and report.

To obtain the best sensory information, many dynamic signals including vibration, need a wide dynamic range data acquisition device. Dynamic range is a measure of the data acquisition's systems ability to detect both strong and weak signals at the same time. In the case of vibration, it is important as many vibration sensors measure vibration from multiple machine components at the same time. In other words, high-amplitude low frequency vibration from unbalance is measured along with low-amplitude high frequency vibration from roller bearing and gears. A wide dynamic range data acquisition system, such as a 24 bit delta sigma analog to digital converter is beneficial in prognostic applications. The difference in amplitudes at various frequencies can be seen in the Figure 5.

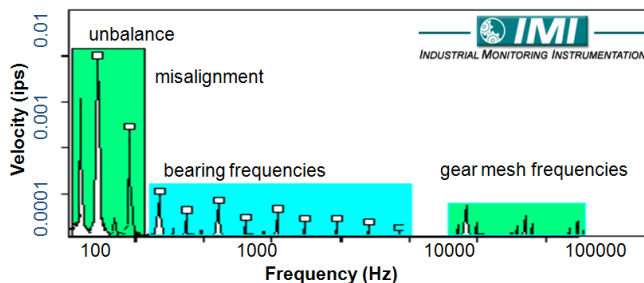


Figure 5: Vibration frequency and amplitude spectrum

A high dynamic range then allows the single sensor to correctly digitize unbalance vibration, mechanical looseness vibration, bearing fault vibration, and even gear mesh vibration.

In addition to dynamic range, there are several other factors for consideration in the data acquisition component. These include anti-aliasing filters, amplification, and sensor power. Typical 24 bit data acquisition hardware includes anti-aliasing filters that remove high frequency noise from the

measured signal. While pre-amplification of signals is not often needed with 24 bit hardware, attenuation may be desirable to provide for dynamic displacement or proximity probe sensors. The latest 24 bit data acquisition hardware offers a +/- 30V input range at bandwidths of 40kHz, creating a universal input for accelerometers, proximity probes, voltage inputs, and tachometer signals. Finally, most 24 bit vibration acquisition hardware devices provide configurable IEPE 24V power to power accelerometers, laser tachometers, dynamic pressure, acoustic emission and microphone sensors.

To provide a data acquisition component of the prognostic system, there exist three core choices. First, it is possible to design the data acquisition system from the ground up, selecting analog to digital and signal conditioning semiconductor components, board manufacturers, embedded processors, programming languages, and so on. While this approach can lead to the lowest manufacturing cost for higher volumes, the electronic design domain expertise and time to market costs become prohibitive.

A second choice is to purchase a series of board level products following any number of standards such as PC-104. This choice offers the prognostics developer a range of suppliers to choose from and a range of generic processor and analog to digital boards that can work together. In most cases however, the processor and analog boards have limited software facilities for analysis, data storage, and downstream diagnostics or prognostics. In other words, they provide a fundamental capability, primarily designed for control applications with limited dynamic range, and have limited software support. These products typically are designed to compete on price, with limited advanced features often needed for embedded or downstream analysis. The prognostics developer then must create AND validate signal processing, filtering and other related algorithms in addition to data storage and communications. This effort can become a significant software development effort.

A third choice is to build the prognostics system on a modular system designed for high fidelity sensor measurements with wide dynamic range, with a wide range of hardware certifications, and with a full featured signal processing library for downstream or embedded prognostic analysis.

The second and third options are differentiated by software development tools including mathematics as well as system certification activities. Figure 6 shows a comparison of complete custom (option 1) and modular system (option 3). There is considerable reduction in development effort required when using a instrumentation class modular system as the basis of a prognostics system. A COTS system providing modular instrumentation class data acquisition then allows the prognostic systems developer to focus attention on health assessment and prognostics.

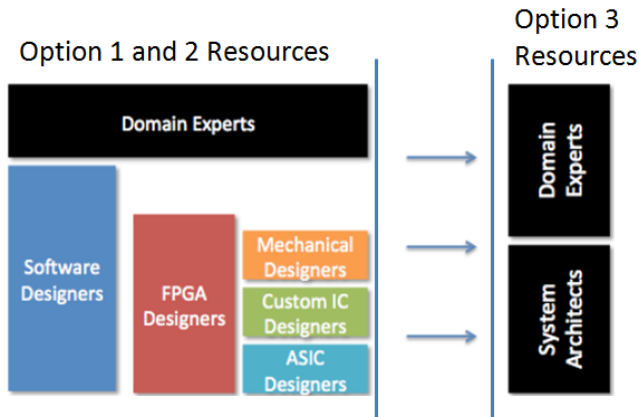


Figure 6: Modular system development effort reduction

Given, there is a data acquisition modular hardware platform in place, it is possible to adjust the sensory input capabilities of the system using modular hardware I/O to best match the machine for which the prognostic system will be used. A modular system allows analog input modules for a variety of sensor types to be selected based on the needs of the system. For example, vibration, temperature, speed, and process variable modules can be added (or removed) based on the measurements that best serve the prognostic system.

Finally, hardware should be industrial grade, meeting high standards for rugged environments. Key measures of rugged reliability include temperature, shock, and vibration. The prognostic systems designer should also consider supplier quality systems and hardware warranty. Table 1 provides common rugged specifications.

Table 1. Common rugged specifications for hardware

Item	Standard	Measure
Operating temperature	IEC 60068-2-1, IEC 60068-2-2	-40 to 70 °C
Storage temperature	IEC 60068-2-1, IEC 60068-2-2	-40 to 85 °C
Operating vibration random	IEC 60068-2-64	5 grms, 10 to 500 Hz
Operating vibration sinusoidal	IEC 60068-2-6	5 g, 10 to 500 Hz
Operating shock	IEC 60068-2-27	30 g, 11 ms half sine, 50 g, 3 ms half sine, 18 shocks at 6 orientations

A number of COTS suppliers provide modular industrial grade data acquisition hardware. Example suppliers include National Instruments, Spectris, Advantech, and Measurement Computing Corporation.

With a solid hardware framework, the prognostics developer is able to focus on the systems architecture, prognostics algorithms, and information delivery aspects of the system.

3.1 Data recording filters

Once the data acquisition system is chosen, it is prudent to determine what signal processing and data storage strategies should be embedded into the data acquisition system component of the prognostic system. On one end of the scale, all time series data from all sensors is continuously recorded. While this strategy insures no loss of data from the machine, it burdens communications and downstream signal processing. For example, simply recording four channels of vibration data at 51,400 samples per second continuously for a week yields 605 Giga-Bytes of data. That is a lot of work for communications, off-line processing, and human interpretation. Much of the data is repetitive.

An alternative is to filter data to limit on board recording to just data that contains new information. This filtering is typically done by onboard analysis.

By analyzing data, it is possible to determine whether the data has changed. On board analysis is performed on monitored sensory data such as speed, temperature, vibration, strain, and electrical power. Examples of onboard analysis include statistical analysis and spectral analysis. The prognostics system should be configurable to allow for deviation limits of sensory information to be used as data storage triggers. With this implementation in place, sensory data is recorded only when it has changed, on a periodic basis, or when an operator request has occurred. Further, the recording includes the condition which caused the data to be recorded. By recording a range of sensory metrics or features along with the recorded data, it is possible to sort the data downstream in the prognostic process. These sensory metrics, then allow the downstream prognostics functions to categorize operational and failure patterns of the same machine and similar machines.

When reviewing the capabilities of COTS technologies for a prognostic system, it is prudent to consider software templates, routines, facilities, etc. that allow for data filtering and data sorting, Figure 7. With the ability of the data acquisition system to filter and sort data, downstream prognostic consumers of the data are more focused and productive.

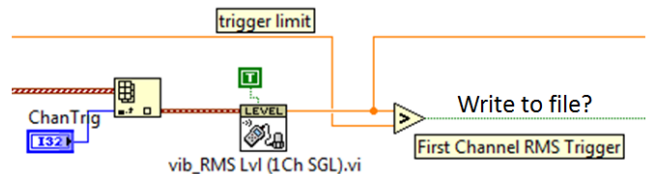


Figure 7: In line analysis drives data recording

Figure 7 shows a simple data recording trigger of vibration RMS level. This block can be replaced and enhanced with a wide range of embedded signal processing including order

analysis, envelope analysis, statistics, etc. The ability to customize storage triggering with embedded analysis is an important modularity feature of several COTS hardware data acquisition platforms.

3.2 Data storage format considerations

When considering data storage formats, it is best to leverage a technology or format that works well for the embedded data acquisition system, and provides rich descriptive capabilities for downstream prognostics analysis. One common “schema” for data recording is the Common Relational Information Schema, or CRIS, as defined by the Machinery Information Management Open Systems Alliance (MIMOSA) organization, MIMOSA (2002). This schema defines data types including time waveform data, spectral data, alarm status, process data, and a range of sensory source information including machine asset and sensor asset information. An illustration of the MIMOSA schema is given in Figure 8.

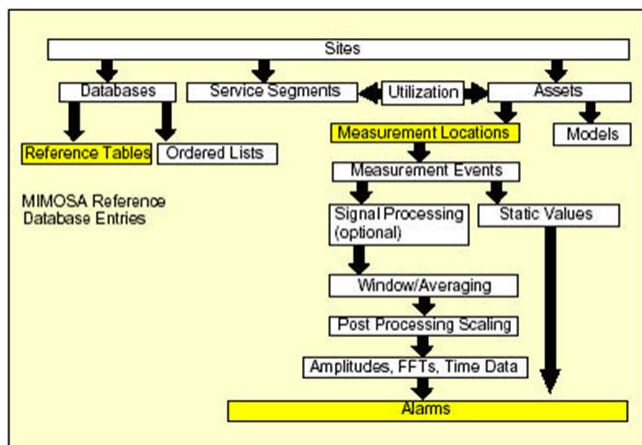


Figure 8: MIMOSA CRIS schema

The MIMOSA CRIS data schema describes a rich data architecture allowing for a combination of time waveforms, spectrum, scalar values, images, and related data types to be stored in a unified data base. When the data sets are organized by sensor, mechanical component, etc, a view of related data sets is easily obtained. For example, opening a sectional view under a roller bearing, one would see time series vibration data, temperature trends, vibration spectra, oil particulate count trends, etc. All of the information is organized as sensory information related to the bearing.

There are several ways to implement an embedded data storage capability which supports this rich data structure. These include common relational database structures and data structures specifically designed for embedded monitoring applications. An example of a embedded data structure format is shown in Figure 9.

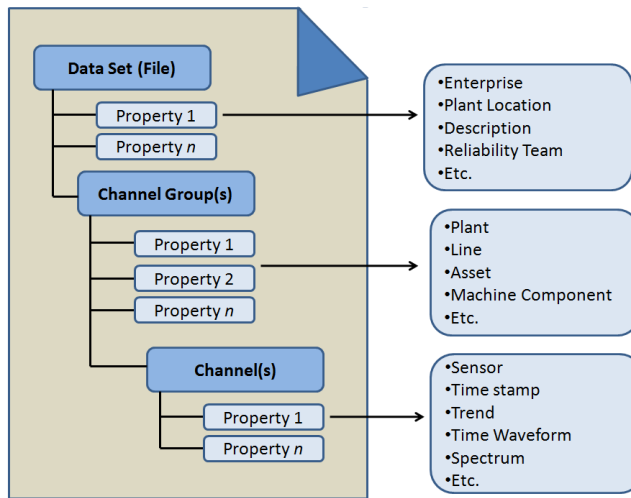


Figure 9: Example embedded data recording structure

Figure 9 illustrates a data structure that is efficient in recording with high speed streaming capabilities. It is rich in data descriptors with the use of data property strings for each data element or channel stored in the data file. In other words, information about the sensor, location, scaling factors, filtering, and mechanical component beneath the sensor, can be stored as labels that describe the time waveform recording. In addition, properties in the data file describe the conditions that caused the data file to be recorded, whether it be an analysis result threshold limit, a time limit, speed change, or operator request.

A second feature of this data structure is the ability to add information along the prognostic system analysis chain. In other words, as the data file record is moved from the data acquisition device downstream to an engineering workstation, additional analysis can be performed on both time series data and extracted features which are stored alongside the original sensory data record, Figure 10.

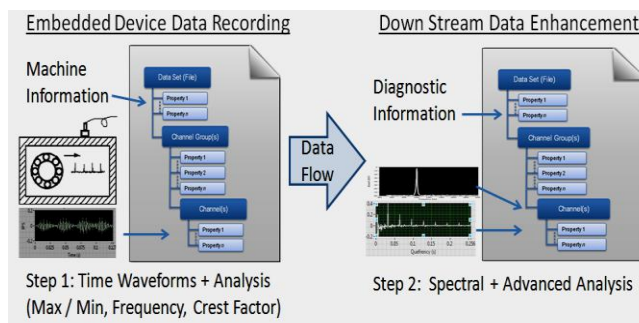


Figure 10: Progression of data structure

The task of analyzing and categorizing data is rarely complete. With a flexible data store, additional analysis results, methods, comparisons, labeling, etc can be added to the data set at any time in the progression of the prognostics process. Going further, if specific empirical patterns emerge, the data files become new models or fault mode references. A flexible and modular COTS data acquisition system provides a core framework for the important task of digitizing and storing necessary sensory data.

3.3 Data acquisition system communications






A third important element of the data acquisition system is communications capabilities. While TCP/IP communications is a common monitoring systems requirement, there are situations where alternative methods are beneficial. These communications protocols can include Controller Area Network (CAN) based protocols including DeviceNet, CanOpen, Modbus, etc. (National Instruments 2010). Further, TCP/IP communications may vary in physical form including copper, fiber optic, cellular, and 900 MHz communications. The data acquisition platform framework should be able to easily accommodate many of these communications variants, allowing adaptation of the prognostics systems to oil and gas machinery, to mining equipment, to wind turbines, to remote equipment and many others. With a flexible architecture, the data acquisition system abstracts the communications to a higher level, where specific communications protocols can plug in easily.

By leveraging flexible communications architectures, the embedded component of the prognostics system is able to easily adapt to the needs of the industry and its machines. In addition, with the embedded data recording structure described above, data can be stored locally, and forwarded to the engineering team at the pace of the communications system and when the communications network is available. This “store and forward” capability is valuable for remote machinery locations with sporadic and slower communications.

4. SIGNAL PROCESSING AND VISUALIZATION

Signal processing functions operate on sensory data to extract features or measurements from data acquired from sensors placed strategically on the machine. Signal processing can occur in the data acquisition system, downstream on an engineering or database computer, or even across the internet leveraging emerging cloud computing technologies. Signal processing plays a part in state detection, health assessment, and prognostic assessment steps in the complete prognostic system. Table 2 and Figure 11 illustrate several signal and data processing functions that can play a part in the commercial prognostic system, Zhang (2008).

Table2: Signal processing options for feature extraction

Graphic	Signal Characteristic	Analysis Methods	Machine Example
	Narrow frequency band lasting for a long time	Frequency Analysis Fourier Transform Power Spectrum	Unbalance in a single speed machine
	Narrow frequency band with harmonics lasting for a long time	Quefreny Cepstrum	Damaged bearing in a machine with roller element bearings
	Time varying frequency band	Time-frequency analysis Order analysis	Unbalance in a variable speed pump
	Wide frequency band signal lasting for a short time	Wavelet analysis AR Modeling	Low speed machine with compressor valve impacts
	Narrow frequency band signal lasting for a short time	WaveletAnalysis	Electrical motor driven machine with rub and knock noise.

As Table 2 indicates, there are a wide range of signal processing options for condition monitoring and prognostics applications. The choice of signal processing function is made on feature extraction needs, mechanical phenomenon indication desired, and domain expertise and preference of the prognostic system designer. It is important that the software development tools used to implement the prognostic system, offer a wide range of signal processing capabilities.

The IMS Center at the University of Cincinnati has added performance prediction, assessment, and diagnostic pattern matching as a supplement to advanced signal processing, Intelligent Maintenance Systems (2007). These capabilities operate downstream from embedded data acquisition by categorizing extracted features into operating modes and failure modes.

Underlying signal processing and prognostics algorithms is a wide range of mathematics. It is important then that the underlying math meets applicable standards and quality metrics. One such reference is the Numerical Mathematics Consortium, (NMC 2009). In the case of sound and vibration numerical functions, there exist several standards including ANSI, ISO, and IEC. When using signal processing algorithms that meet existing standards, the prognostics system developer is able leverage the certification and validation work of the algorithm supplier.

Health or performance assessment and prediction or prognostics assessment build on signal processing used in the data acquisition, data filtering and sorting, and feature extraction steps of the upstream prognostic components. These additional steps, including logic regression, self organizing maps (SOM), and even the field of statistical pattern recognition; provide tools for matching current measurements with data driven models of system health and failure modes. In other words, the discovery of impacting and out of balance features in vibration data can match

patterns of induced stress on roller bearings and help predict a specific bearing failure.

The leverage of signal processing for feature extraction and health indication measurements, leads to visualization of data and signal processing results that the human uses to understand a problem or degradation in the machine. Figure 11 offers one example of visualization graphics.

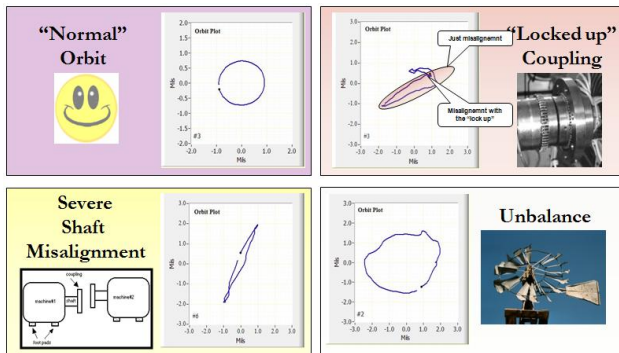


Figure 11: Orbit plot visualization of shaft vibrations

Orbit plots are a common diagnostic and health indicator graphic used in turbine driven machinery applications. These plots indicate the severity of out-of-balance, alignment, and coupling machined degradation issues. The shape and size of the orbit plot indicates the progression of specific shaft vibration problems in the machine. The shape and size of the orbit plot can be analyzed by human domain experts as well as analytically with mathematical algorithms.

Additional visualization tools exist in prognostics software development libraries to summarize multiple machine or system components. These summary plots provide a high level of machine health and allow for selection of suspect machines for further study. The University of Cincinnati's Intelligent Maintenance Systems Center offers several visualization tools for information delivery, Lee (2009), Figure 12.

These graphics provide visual display of health information. The Confidence Value trend chart shows the mechanical health of a specific machine component using a measure of 1 (very healthy) to 0 (badly damaged). The confidence value is commonly calculated using statistical pattern matching described earlier. The Health Radar Chart shows the confidence value of multiple components on a single chart. The Health Map combines machine operational states with machine failure modes. The Risk Radar Chart combines machine state and health indicators along with safety and financial parameters to indicate an element of risk.

Results of Smart Prognostics Tools for Asset Health Information

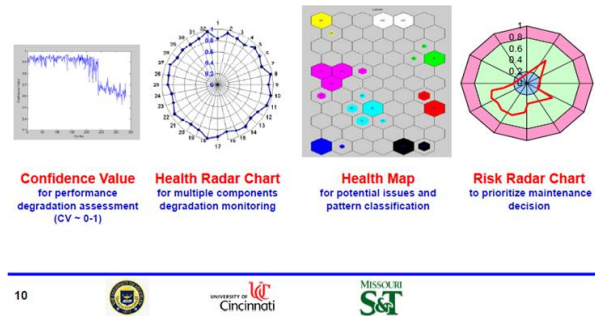


Figure 12: Visualization of machine health and prognostics

Armed with these reports, operations and maintenance teams are best prepared to make operational and maintenance decisions. Of course these end reports build on solid data collection and signal processing techniques described earlier.

The signal processing and visualization components of the prognostic system can be utilized in the embedded data acquisition portion of the system, at the local engineering workstation computer, and over the network and remote engineering centers or data centers. The flexibility of location of mathematical analysis offers the prognostic systems designer options to choose the best place for advanced prognostics in the data acquisition, filtering, storage, and post processing components of the prognostic system.

5. CASE STUDIES

There are several case studies worth review where a modular system framework is in use for condition monitoring and prognostics applications. While several are relatively new to the market, each leverages a common modular hardware data acquisition platform, with modular software architecture allowing for the placement of signal processing and prognostic functions to be placed anywhere along the data acquisition, off-line data manipulation, and visualization sequence of prognostic system activities.

In power generation applications, wind energy continues to lead renewable energies as next generation sources of power. However, these machines are complex and operate in a variety of speed, load, and environmental conditions. These energy generating machines historically have shown to have reliability problems in the drive train. Much interest in research and industry is focused on improved monitoring, diagnostics, and prognostics systems to support

wind energy applications. One such example is illustrated in Figure 13.

Wind Farm Prognostics

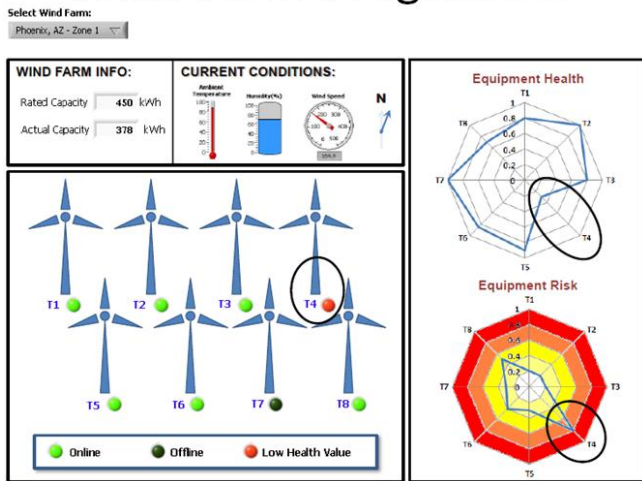


Figure 13: IMS Center view of wind farm prognostics

The IMS Center continues work to adapt state of the art prognostics systems technologies to wind energy applications. In partnership with National Instruments, the IMS center leverages rugged COTS embedded data acquisition technologies, signal processing algorithms, and local and web based visualization tools to implement a wind farm prognostics framework. This framework is used to further research in wind turbine prognostics and to develop an advanced commercial condition monitoring and prognostics system for the wind energy industry.

Another example in wind energy applications is the use of the modular COTS hardware and software prognostics development platform by bearing supplier, FAG, Figure 14.

Wind Farm Condition Monitoring

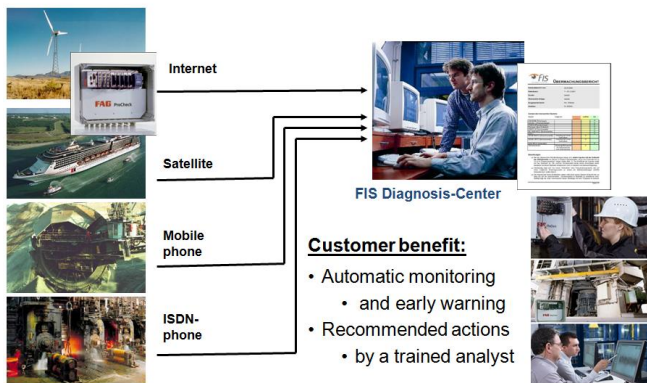


Figure 14: FIS condition monitoring system

FAG Industrial Services, the service division of FAG industrial bearings, has developed an advanced condition monitoring system based on modular COTS embedded data acquisition and signal processing platforms. The monitoring systems are used both by wind farm operators and maintenance service teams, as well as FAG’s bearing service and support center. Embedded intelligence in the monitoring system, specifically envelope analysis, reduces sensory data at the data acquisition source to information that is more actionable when it reaches operations and maintenance personnel. The information is transmitted over existing controls networks or wirelessly leveraging cellular and RF technologies, Langer (2006).

Another case study, explores distributed condition monitoring and prognostics in nuclear power, Shumaker (2010), Figure 15.

- High Flux Isotope Reactor
- Routine surveillance
- Difficult locations
- Cabling is very expensive
- “Data Dashboard”
- MCSA, Vibration, Temp

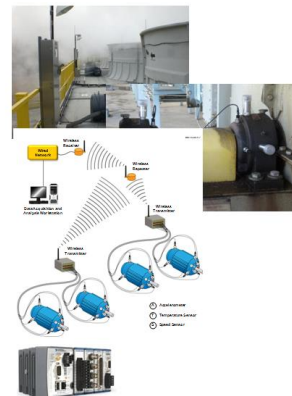
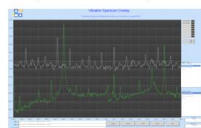


Figure 15: AMS-Corp nuclear pump monitoring systems

Analysis and Measurement Services Corporation (AMS) specializes in testing of process instrumentation and development of specialized test equipment and software products for power and process industries. This project proposes a comprehensive effort to expand and commercialize previous research projects to provide passive, in-containment use of wireless technology at nuclear power plants. Specifically, the effort of the subsequent phases of the project will focus on assembling a complete, commercial, wireless on-line data monitoring and analysis system that can be adapted for use in any pressurized water reactor containment. The system would be used for condition monitoring during plant operation and/or outage time to provide additional measurements that may be needed by the maintenance crews, operations or plant management. Because of the nature and purpose of nuclear plant containment, the introduction of a wireless network/communication system inside the confined area is challenging and, yet, very advantageous. The immediate benefit to the nuclear plant is the reduced cost for monitoring equipment and/or processes within containment

and to provide additional data as needed for maintenance work during refueling outages and normal operation.

This particular example, leverages COTS technology including rugged embedded data acquisition and signal processing to gather and digitize sensory information, to store and forward the sensory data over wireless TCP/IP and to format the data in a flexible data schema for off-line analysis and reporting.

Finally, in the mining and materials industry, a wide range of conveying and grinding machinery is used. Crushers are important assets in material processing plants. O'mos developed a condition monitoring solution using a modular architecture to monitor the health of cone crushing equipment, Epie (2011).

The modular conical mill condition monitoring system uses accelerometers, temperature sensors, and pressure switches. O'mos is a service company, providing maintenance services for its customers. With remote monitoring and in-line signal processing, O'mos is able to improve its service offerings to its material processing customers. The ability to leverage COTS embedded data acquisition and analysis components frees O'mos to focus their expertise on off-line analysis, prediction, and reporting. O'mos lowers their cost of service thru data acquisition automation while working to improve reporting and recommendation results leveraging specific conical mill domain expertise.

Several other prognostics suppliers are working to adapt COTS technologies as the foundation for their prognostic offerings. Example prognostics offerings are IMS Center Watchdog™ Agent, Global Technologies Corporation's PEDs hms™, and Impact Technologies ReasonPro™.

6. CONCLUSION

By leveraging commercial off the shelf (COTS) technologies and a flexible modular architecture or framework, it is possible to develop and bring to market a prognostic system that adapts to a wide range of machines, industries, and applications. The prognostics system developer is able to get to market rapidly and at less cost, than the alternative of developing components that are otherwise commercially available. This benefit is specifically realized, when the COTS components are flexible in data storage, and signal processing capabilities

making it possible to adapt the COTS components for specific prognostic algorithms and methods.

REFERENCES

- Epie, C. (2011), Developing a monitoring system for predictive maintenance of a conical mill, National Instruments case study, no. 13377, <http://sine.ni.com/cs/app/doc/p/id/cs-13377>
- Intelligent Maintenance Systems (2007). Watchdog Agent™ documentation, Center for Intelligent Maintenance Systems, www.imscenter.net
- ISO Standards (2003), Condition monitoring systems standards ISO-13374, www.iso.org
- Langer, G. (2006). FAG Industrial Services; how to detect bearing faults; *National Instruments NIWeek 2006 Conference*; Presentation E227; Austin, Texas USA, August 2006; <http://zone.ni.com/wv/app/doc/p/id/wv-85>
- Lee, Dr. J. (2009). Advanced prognostics for smart Systems. intelligent maintenance systems (IMS) introduction, pp. 6-8. www.imscenter.net
- Machinery Information Management Operational Systems Alliance (2002), Overview of MIMOSA www.mimosa.org.
- National Instruments (2010). Connectivity to industrial communications. On-line tutorial. <http://zone.ni.com/devzone/cda/tut/p/id/10473#toc5>
- Numerical Mathematics Consortium (2009). Numerical mathematics consortium releases new draft standard for numerical algorithm development. <http://www.nmconsortium.org/news/release.aspx?id=39>
- Shumaker, B. (2010). Online monitoring of nuclear reactors using CompactRIO. National Instruments website case study. <http://sine.ni.com/cs/app/doc/p/id/cs-13207>
- Zhang, N. (2008). Advanced signal processing algorithms and architectures for sound and vibration; *National Instruments NIWeek 2008 Conference*; Presentation TS 1577; Austin, Texas, USA; August.

Comparison of Fault Detection Techniques for an Ocean Turbine

Mustapha Mjit, Pierre-Philippe J. Beaujean, and David J. Vendittis

Florida Atlantic University, SeaTech, 101 North Beach Road, Dania Beach, FL 33004 USA
mmjit@fau.edu
pbeaujea@fau.edu
dvendittis@aol.com

ABSTRACT

The Southeast National Marine Renewable Energy Center at Florida Atlantic University, which supersedes the Center for Ocean Energy Technology (Driscoll et al., 2008), is conducting research and development to support the implementation of ocean current and ocean thermal energy technologies for a low environmental-impact extraction of energy.

Fault detection capability is needed for these offshore ocean turbines (and other systems) because access to these machines for maintenance is difficult and costly. Techniques that offer reliable and early (incipient) detection allow for preventive maintenance to prevent the development of secondary faults that may be generated by the primary faults. Several methods for processing and displaying vibration data are compared and evaluated relative to synergistic detection utilizing data from a prototype (dynamometer) of an ocean current turbine. The results may generically apply to other machines, such as wind turbines.¹

1. INTRODUCTION

An ocean turbine (OT) is subject to high and varying loads, locations that are difficult to access and extreme environment conditions; therefore, it requires special predictive monitoring strategies (Sloan et al., 2009; Beaujean et al., 2009). For many machines, a vibration condition monitoring program is considered as one of the most important tools to detect the presence of anomalous behavior, thus allowing for early remedial actions to reduce both maintenance costs and premature breakdown. Since access is difficult and costly, monitoring techniques that detect these faults reliably (and early) for machines like offshore ocean turbines offer an advantage over the more

standard techniques (e.g. vibration level trending), allowing for preventive maintenance and to prevent the development of secondary faults that may be initiated by the primary faults.

This paper discusses several approaches, procedures and techniques considered to detect and diagnose the faults of an ocean turbine, utilizing vibration data. Specifically, modulation detection techniques utilizing the Cepstrum or the Hilbert transform and transient detection techniques (Short Time Fourier Transform (STFT) and kurtosis) are considered. Such methods have shown to be efficient, (Fernandez et al., 2005; Kim et al., 2006), for detecting faults that affect the component health of machines (e.g. motors, gearboxes, fans and generators) generically similar to those that may be considered subsystems of an OT (Figure 1).

A LabVIEW model for on-line vibration condition monitoring was developed (Mjit, 2009; Mjit et al., 2010). It contains the advanced fault detection techniques mentioned above as well as diagnostic techniques that provide information about the type, severity and identification of the fault. The principal monitoring method utilizes the Power Spectral Density (PSD) for in-depth analysis of the vibration signal and for vibration level trending, assuming acceptable stationary of the vibration signal. The model was exercised using data acquired from a rotor end of a dynamometer (Figure 2), which is representative of the electrical and mechanical equipment of the actual OT (Figure 1). The data were processed in several different ways to evaluate the relative ability of the detection techniques to detect the types of incipient faults expected of the OT. Actual turbine data may differ because of the presence of the dynamometer's motor drive and additional gearbox. Varying loads and structural fluid loading of the OT may affect the frequency of structural resonances; however the types of mechanical faults should be generically the same. The purpose of this effort was to determine if the conclusions and recommendations made in (Fernandez et al., 2005; Kim et al., 2006) apply to the dynamometer and, possibly, the OT.

¹ This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Most machines display some non-stationary behavior, mainly because of varying loads or structural variations. This affects the vibration data (Fernandez et al., 2005), mainly by causing frequency shifts. Because of the additional variations caused by changes in the water current velocity, it is expected that the vibration data collected from the ocean turbine will be even less stationary in nature than those from the dynamometer. Therefore, the use of wavelets, possibly combined with other algorithms, such as the Hilbert transform, may be necessary to assess changes in the vibration levels, (Fernandez et al., 2005; Fan et al., 2006; Tyagi et al., 2003; Wald et al., 2010). Such a combination is under development and will be evaluated with the in-water turbine data.

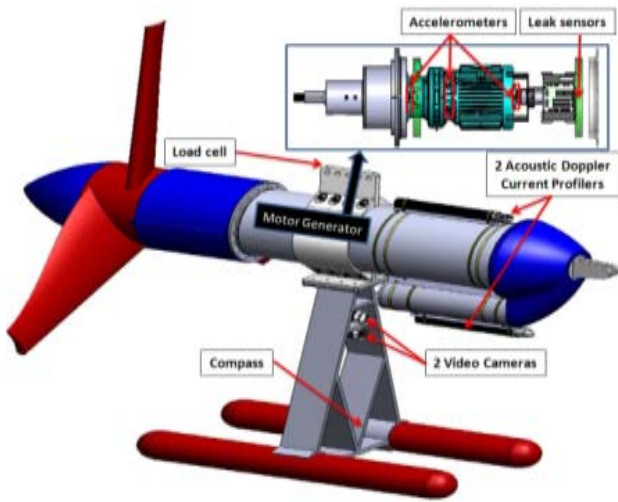


Figure 1. Ocean Turbine, Conceptual.

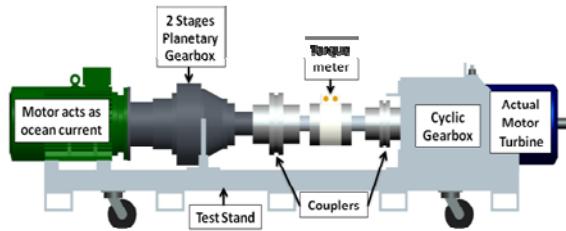


Figure 2. Dynamometer.

2. LISTING AND DESCRIPTION OF THE FAULT DETECTION TECHNIQUES

2.1 Power Spectral Density and Fractional Octave Analysis

The PSD is derived from the vibration time waveform by performing a Fast Fourier Transform (FFT). The PSD is well-suited to analysis and diagnosis as it shows more

clearly the forcing frequencies of the rotating components. This technique is very accurate for stationary machines. The PSD is averaged over fractional octave bands, and is used for trending and detection as it covers a large frequency range. The trending of fractional octave spectra is very accurate especially if there is small speed variation. The PSD is also very accurate for stationary machine where the forcing frequencies of the components do not vary with time. The PSD can also be used for (slightly) non-stationary machines if one is only interested in the spectral components that exist in the signal, but not interested in what time each spectral component occurs. Most of the peaks in the PSD are directly proportional to the running speed of the machine. The PSD may be normalized during each iteration before the averaging process to avoid smearing in the case of non stationary machine speeds.

2.2 Cepstrum Analysis

The power cepstrum is the inverse FFT of the logarithm of the power spectrum of a signal; it is used to highlight periodicities in the vibrations spectrum, in the same way that the spectrum is used to highlight periodicities in the time waveform. Thus, harmonics and sidebands in the spectrum are summed into one peak in the cepstrum (called rahmonic), allowing identification and trending of modulation frequencies associated with a specific fault.

$$C(f) = \int_{-\infty}^{+\infty} \log \left(|F\{x(t)\}|^2 \right) \times e^{-j2\pi s/f} ds. \quad (1)$$

F is the Fourier transform operator; $x(t)$ is the time signal and f is the frequency in hertz.

2.3 Kurtosis

Kurtosis is a statistical parameter, derived from the fourth statistical moment about the mean of the probability distribution function of the vibration signal and is an indicator of the non-normality of that function. The kurtosis technique has the major advantage that the calculated value is independent of load or speed variations. The kurtosis analysis is good for faults and transient effect detection, but it does not give an indication of the specific source of the problem (Reimche et al., 2003); however, the kurtosis will diminish with increased distance from the source of the transients. The kurtosis will be equal to 3 for a healthy machine and greater than 3 if the machine's vibrations contain transients. The general definition of the kurtosis is,

$$Kurtosis = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{(n-1)\sigma^2} \quad (2)$$

The variables $x_1, x_2 \dots x_n$ represent the population data of the signal, \bar{x} is the mean of x , σ is the variance of x and n is the number of samples.

2.4 Hilbert Transform Analysis

The Hilbert transform of a real signal is defined as the convolution of the signal with the function $1/\pi t$,

$$\hat{x}(t) = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{x(\tau)}{t - \tau} d\tau. \quad (3)$$

The complex analytic signal is:

$$\tilde{x}(t) = x(t) + j\hat{x}(t). \quad (4)$$

The envelope of the original signal is defined as follow:

$$e(t) = \tilde{x}(t) e^{-j2\pi ft}. \quad (5)$$

f is the frequency of the modulated signal.

The Hilbert transform is used to demodulate the signal so as to obtain the low frequency variations (faulty signal) in a higher frequency signal (forcing or resonance frequency). When a fault starts developing, the vibrations caused by a bearing or gear fault is obscured (especially at low frequency) by the noise or the vibrations from other rotating parts like shafts, gears, etc. In this case, the bearing or gear frequencies cannot be seen in either the time waveform or the spectrum of the vibration. The Hilbert transform can be used to highlight and extract the modulating signal (faulty signal) from the modulated signal (characteristic frequency of the machine). The Hilbert transform technique removes the carrier signals which are of no interest for fault detection. Amplitude modulation occurs for example when a gear rides on a bent or misaligned shaft, while frequency modulation occurs for example when the shaft speed varies with time. In the case of a narrow-band detection process, a band-pass filter (whose pass band includes the fault frequencies) filters out the selected part of the spectrum. The output is shifted (heterodyned) to low frequency and subjected to envelope detection.

$$BSP(f, f') = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \tilde{x}(\tau) w_1(t, \tau) e^{-j2\pi(f'\tau + ft)} d\tau dt \quad (6)$$

BSP is the bispectrum of the analytical signal \tilde{x} , f and f' are the modulated and modulating frequencies respectively. w_1 is a time window.

2.5 Short Time Fourier Transform

For non-stationary machines, the Short Time Fourier Transform (STFT) of the signal should be used to clearly identify non-stationary vibration data related to speed variation, from vibrations caused by the inception of anomalies. Indeed, the PSD may not provide sufficient information about the presence of transient effect, since abrupt change in the time signal is spread out over the entire frequency range. Time-frequency Analysis results are displayed in a spectrogram, which shows how the power of a signal is distributed in the time-frequency domain. Narrow-band, periodic signals, transients and noise appear very distinctly on a spectrogram. The STFT is based on the following mathematical operations,

$$PS(t, f) = \left| \int_{-\infty}^{+\infty} x(t') w_2(t' - t) e^{-j2\pi ft'} dt' \right|^2 \quad (7)$$

PS is the power spectrogram of the signal $s(t)$ and $w_2(t)$ is a real and symmetric window translated by t . t and f are the instantaneous time and frequency.

3. DATA ACQUISITION

Vibration data were acquired from the dynamometer running at various RPM with simulated faults to evaluate the ability of the detection algorithms to detect the presence of incipient faults. Processing features such as stationary assumptions and smoothing windows were also evaluated to insure a high quality for the data. The faults were simulated for selected levels of severity to determine whether the conclusions depended on existing signal-to-noise levels. Note that the motor and second gearbox (simulated rotor) section of the dynamometer were detached and instrumented (Figure 3). The motor was operated at selected speeds with and without several weights that increased the rotor shaft imbalance. These tests were performed to evaluate the envelope analysis using Hilbert, the PSD and cepstrum techniques. Additionally, a hammer was used to introduce impact transients. The response of the monitoring system to such impacts was evaluated using the kurtosis and the STFT. The anomaly detection techniques were implemented and assessed relative to their detection capabilities.

In this paper, the motor was operated at 1,144 RPM under normal and augmented imbalance condition. The shaft rotated at 52.47 RPM due to the reduction ratio (1:21.8) of the two stages reduction planetary gearbox. The expected mechanical forcing frequencies (Singleton, 2006) relative to the motor speed of 1,144 RPM are summarized in Tables 1 and 2. These forcing frequencies are calculated

automatically in the LabVIEW program. Note that Tables 1 and 2 show only the fundamental forcing frequencies and not their harmonics.

The experiment was performed on the rotor section of the dynamometer in three different situations: without any weight added to the shaft, with a light weight (two magnets, 0.5 lbs total) and with a large weight (two magnets and two blocks, 2.875 lbs total) attached to the very end of the output shaft. The distance from the shaft axis to the location of the weights was 3.7 inches. Changes in third octave bands, power spectral density, envelope and kurtosis were measured. The acceleration data were collected using a Low Frequency (0.2 to 3000 Hz) piezoelectric accelerometer with a sensitivity of 500 mv/g mounted on the torque meter. The sampling frequency F_s was 5,000 Hz or 20,000 Hz. The number of data points in each sample was 20,000. The corresponding frequency resolution was 0.25 Hz (at $F_s = 5,000$ Hz) or 1 Hz (at $F_s = 20,000$ Hz). A total of 500,000 points were acquired (25 samples). A Hanning window was used to smooth the data.

4. PSD, HILBERT AND CEPSTRUM ANALYSES

Figures 4 to 6 show the baseline (without imbalance condition introduced) PSD, in three different frequency regions, of the data acquired at 1,144 RPM of the motor. The major frequency components (derived from the known forcing frequencies) are identified in these figures. These forcing frequencies are tabulated in Table 3. In figure 4, the average of the PSD (25 samples) in the low frequency region (0.5 to 50 Hz) was calculated using 500,000 data points (20,000 for each sample) for a sampling rate of 5,000 Hz to achieve a frequency resolution of 0.25 Hz. High frequency resolution was needed in the low frequency region (below 50 Hz) as the forcing frequencies are closer

to each other. In Figures 5 and 6, the average of the PSD (25 samples) in medium and high frequency region - calculated using 500,000 data point for a sampling rate of 20,000 Hz (20,000 points per sample) and a frequency resolution of 1 Hz - are shown. Figure 7 shows the baseline PSD of the motor running at two different speeds (not harmonically related), 1,144 RPM and 1,593 RPM; this allows for the identification of the resonant frequencies of the system as both PSD should display the same peaks at these frequencies.

	1st stage planetary gear	2nd stage planetary gear
Carrier speed	3.57 RPS	0.87 RPS
Planet speed	9.62 RPS	2.70 RPS
Planet absolute speed	6.04 RPS	1.82 RPS
Planet gear mesh frequency	279.02 Hz	59.47 Hz
Sun gear mesh frequency	343.2 Hz	78.69 Hz
Planet passing frequency	10.73 Hz	2.62 Hz
Sun gear side band defect frequency	46.46 Hz	8.10 Hz
Planet gear side band defect frequency	19.24 Hz	5.40 Hz

Table 1. Expected forcing frequencies from planetary gears of the rotor side of the dynamometer, motor speed 1,144 RPM (19.06 Hz).

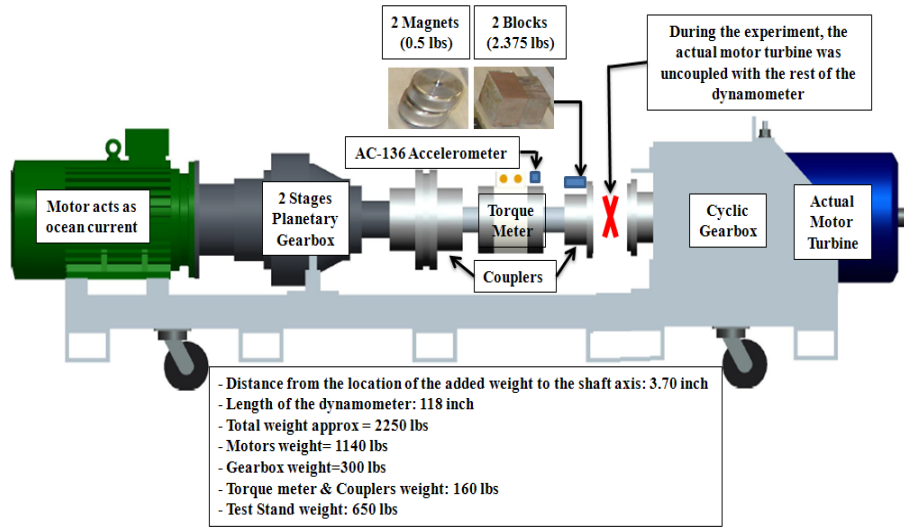


Figure 3. Rotor end of the Dynamometer.

	Bearing 1	Bearing 2	Bearing 3
Outer ring frequency	58.58	58.26	7.69
Inner ring frequency	93.95	94.26	13.29
Roller bearing frequency	77.80	76.29	3.03
Cage frequency	7.32	7.28	0.32

Table 2. Expected forcing frequencies (in Hz) from bearings of the rotor side of the dynamometer, motor speed 1,144 RPM (19.06 Hz).

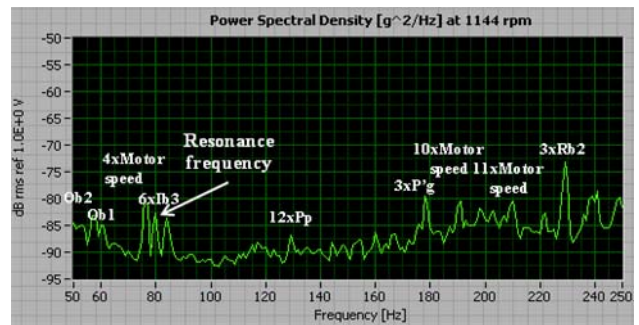


Figure 5. Forcing frequencies identification on the PSD plot (50 Hz to 250 Hz).

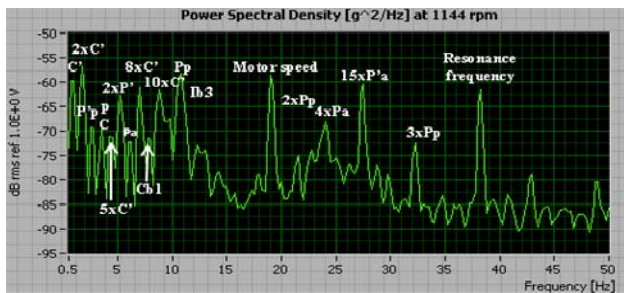


Figure 4. Forcing frequencies identification on the PSD plot (0.5 Hz to 50 Hz).

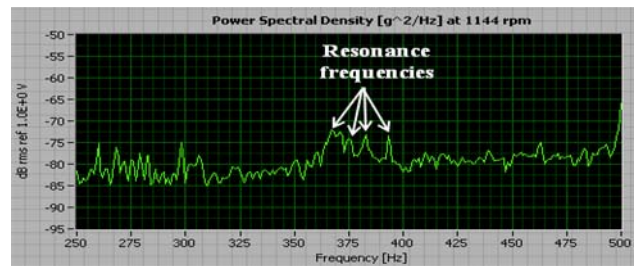


Figure 6. Forcing frequencies identification on the PSD plot (250 Hz to 500 Hz).

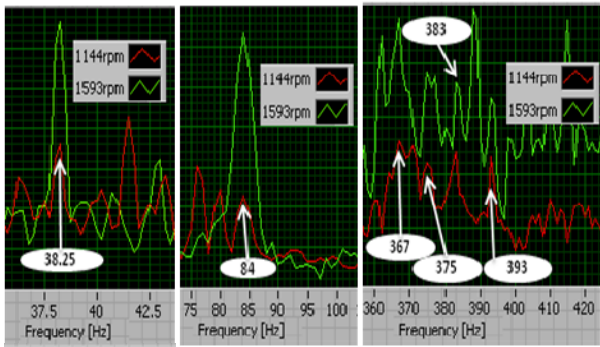


Figure 7. The PSD at 1,144 RPM and 1,593 RPM show the same peaks at resonance frequencies.

In Figure 8, the difference in levels between the imbalance and baseline in the third octave bands (63 Hz, 125 Hz, 160 Hz, 200 Hz, 250 Hz and 315 Hz) exceed the threshold (level increase allowance) of 5dB, causing the alarm for each of those third octave bands to switch on. In figure 9, the PSD and the spectrograms clearly show peaks at about 60 Hz and 120 Hz relative to the modulation in the case of extreme imbalance. The increase in level is due to the imbalance of the shaft that causes the planet gear meshing frequency (59.5 Hz) and its harmonics (119 Hz, 178.5 Hz) to be modulated by the rotational frequency (0.88 Hz) and its harmonics (1.76 Hz, 2.64 Hz, 3.52 Hz) (Figure 11); for brevity, only the modulation of the fundamental planet gear meshing frequency is shown. Similar conclusions were made from data acquired with the motor running at 1,593 RPM. Figure 10 shows comparisons between imbalance effect on the PSD and Hilbert envelope analyses.

The Hilbert envelope analysis shows the major modulation of the gear meshing frequency much more clearly than the PSD do. Table 4 summarizes the amplitude change relative to the baseline using the third octave, PSD and Hilbert envelope analysis, in the case where the unbalance is caused either by two magnets attached to the shaft or two magnets and two blocks attached to the shaft. The table shows that the PSD levels and the levels of the demodulating

frequencies (envelope analysis) increased with imbalance condition. Figures 11 and 12 show the PSD and the spectrograms with two magnets attached to the shaft, and two magnets and two blocks, in the frequency ranges 130-250 Hz and 215-500 Hz, respectively. The spectrograms show clearly the peaks (at the modulating frequencies and its harmonics) that are causing the third octave bands shown in figure 8 to exceed their baselines. The kurtosis was not affected by the imbalance at either speed, but would have changed significantly if the imbalance was causing damage to the gears or bearings – an example of data fusion and a potential tool for prognosis.

Freq.	Forcing frequencies	Symbol
19.06	Motor speed	-
0.87	Shaft speed 2 nd stage gear	C'
3.57	Shaft speed 1 st stage gear	C
2.62	Planet passing frequency 2 nd stage	P'p
10.73	Planet passing frequency 1 st stage	Pp
1.82	Planet absolute frequency 2 nd stage	P'a
6.04	Planet absolute frequency 1 st stage	Pa
178	Planet gear mesh frequency 2 nd stage	P'g
7.32	Cage defect frequency 1 st bearing	Cb1
58.58	Outer race defect frequency bearing 1	Ob1
58.26	Outer race defect frequency bearings 2	Ob2
7.69	Inter race defect frequency bearings 3	Ib3
76.29	Roller bearing frequency bearing 2	Rb2
38.25, 84, 367, 375, 383, 393	Resonance frequencies	-

Table 3. Observed forcing frequencies (in Hz) in the PSD for motor running at 1,144 RPM (19.06 Hz).

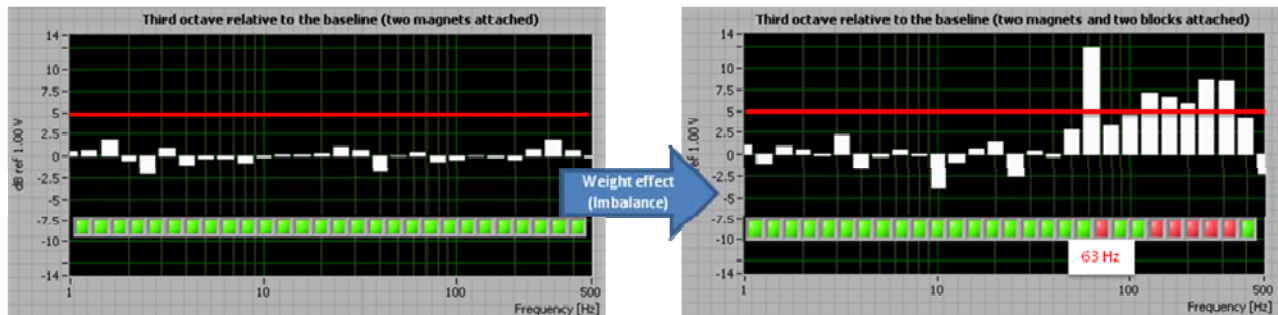


Figure 8. Relative amplitude with respect to the baseline (normal condition) using third-octave analysis; two magnets attached to the shaft (left) and, two magnets and two blocks (right).

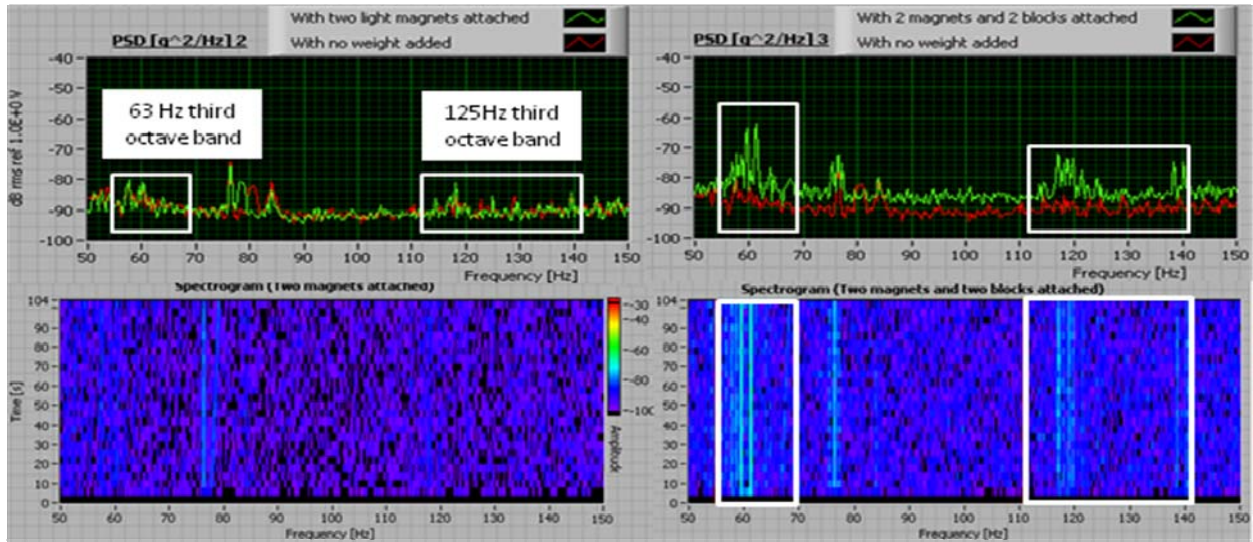


Figure 9. Power spectral density and spectrogram in the frequency range 50-150 Hz, with two magnets attached to the shaft (left) and two magnets and two blocks (right).

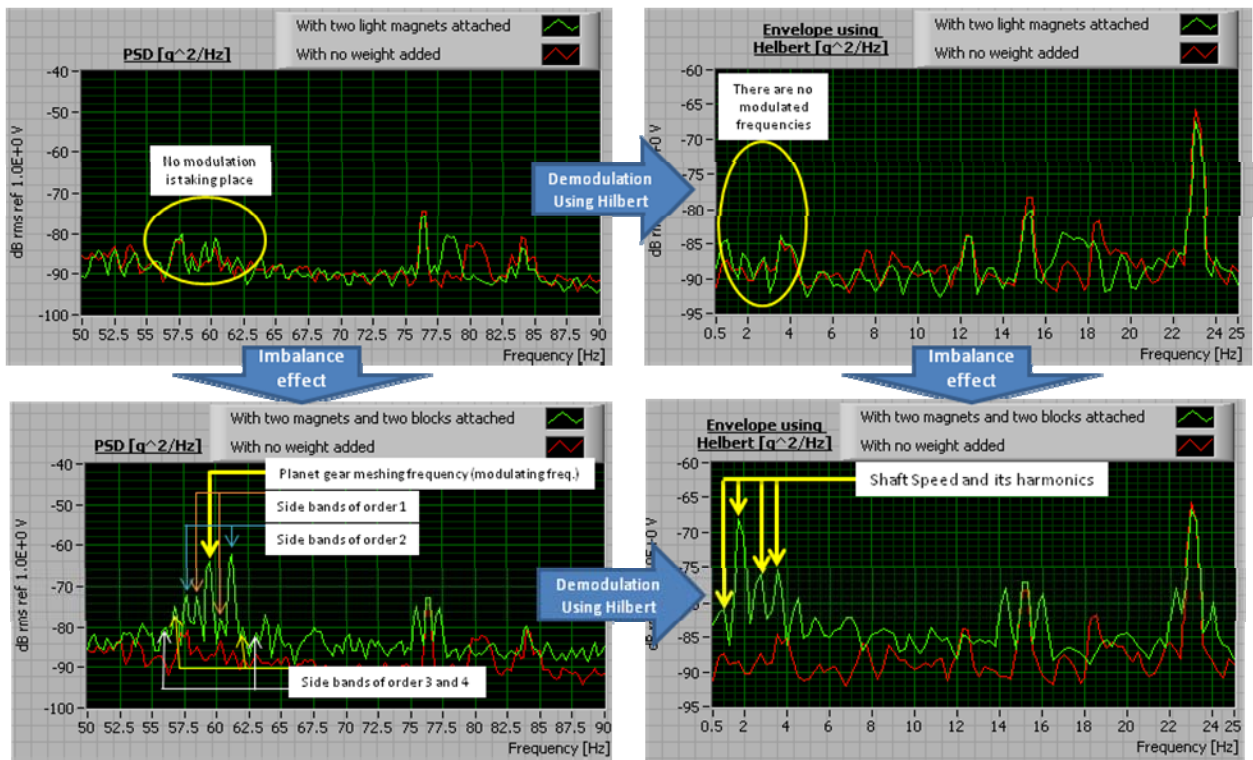


Figure 10. Power spectral density and its Hilbert envelope analysis; two magnets attached to the shaft (top) and two magnets and two blocks (bottom).

	Two magnets					Two magnets & Two blocks				
	59.5Hz tone	1 st order sideband	2 nd order sideband	3 rd order sideband	4 th order sideband	59.5Hz tone	1 st order sideband	2 nd order sideband	3 rd order sideband	4 th order sideband
PSD relative level (dB)	4.63	2.37	-1.34	2.35	3.31	23.04	5.82	23.59	8.26	5.21
Hilbert envelope relative level (dB)		2.87	0.19	-0.4	0.04		11.11	18.17	6.5	6.37
1/3octave Analysis (dB)	0.92 (relative level in the 63Hz third octave band)					12.47 (relative level in the 63Hz third octave band)				

Table 4. Comparison of the features (PSD, envelope and third octave level) relative to the baseline (no weight added), for two levels of unbalance severity, one with two light magnets, and another with two magnets and two blocks.

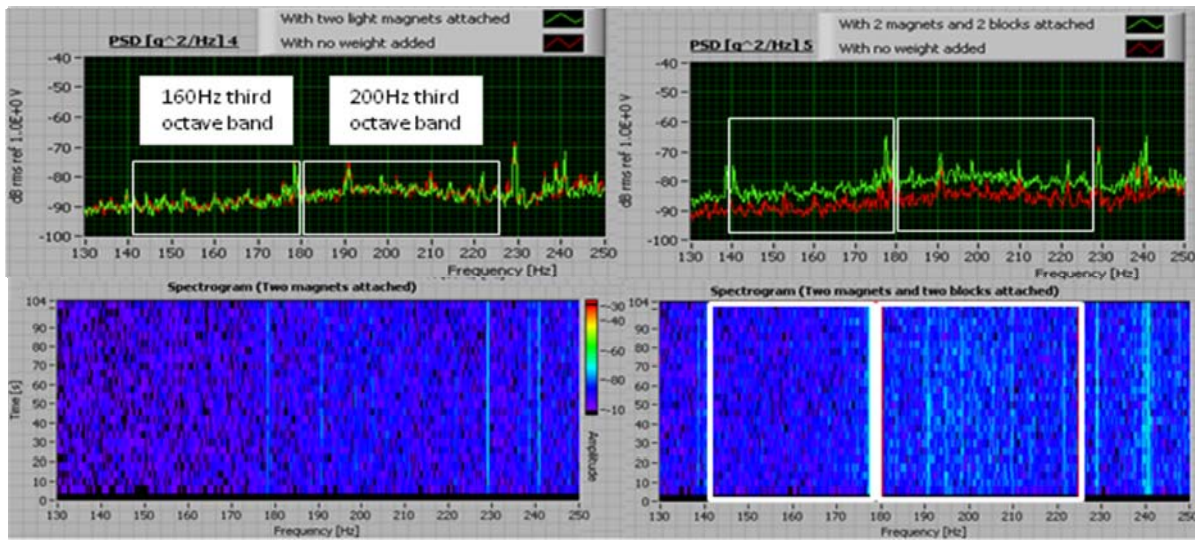


Figure 11. Power spectral density and the spectrogram in the frequency range 130-250 Hz, with two magnets attached to the shaft (left) and two magnets and two blocks (right).

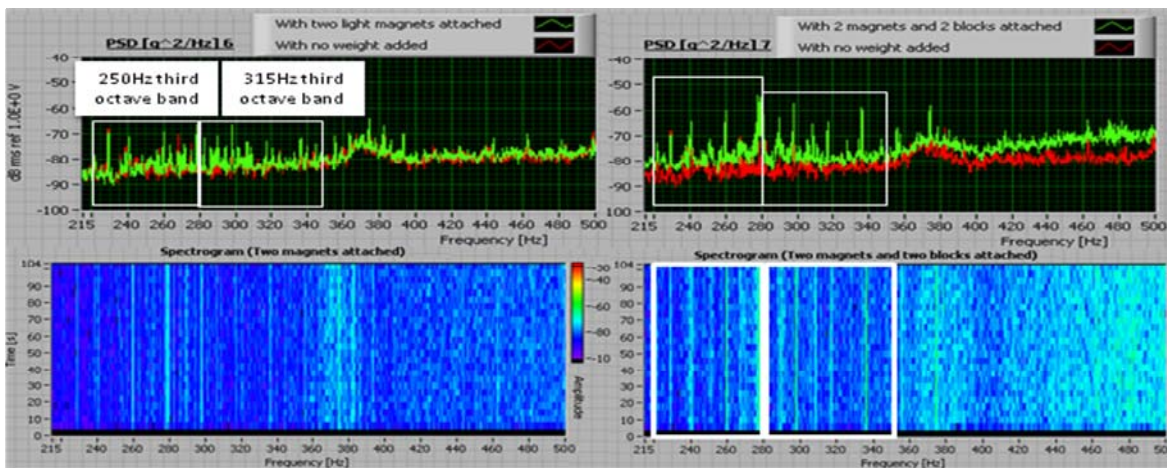


Figure 12. Power spectral density and the spectrogram in the frequency range 215-500 Hz, with two magnets attached to the shaft (left) and two magnets and two blocks (right).

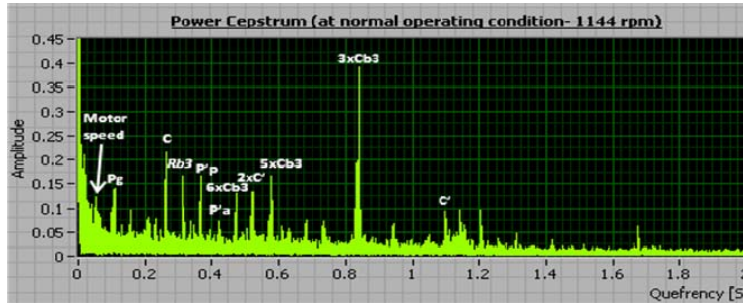


Figure 13. Forcing frequencies identification on the Cepstrum plot.

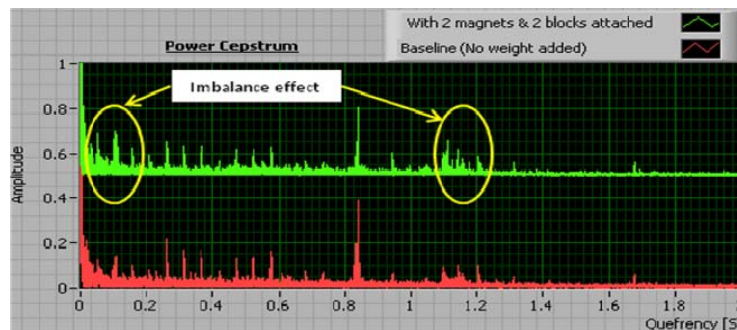


Figure 14. Cepstrum averaged over 25 samples (motor running at 1,144 RPM)-baseline (blue) shifted up by 0.5 compared to extreme imbalance case (green)

Cepstrum analysis was evaluated for early detection or fault diagnosis. In Figures 13 and 14, the increased imbalance of the shaft caused by the weight resulted in amplitude increases of the harmonically related frequencies. These changes are difficult to assess because the imbalance changes the dynamics of the system; e.g., the stress localized on a tooth due to the imbalance of the shaft produces modulation of the tooth-meshing frequencies with the shaft speed. Also, a large number of sidebands around the tooth-meshing frequency and its harmonics in the spectrum are generated, which are spaced by the rotation frequency of the shaft. As discussed earlier, the use of the Hilbert transform based techniques allows for easier interpretations to the monitoring of the envelope at specific frequencies, such as bearing or gear related frequencies. The easier interpretation increases the probability of early detection.

5. TRANSIENT ANALYSIS

A transient analysis utilizing kurtosis and STFT was performed using a calibrated hammer. The hammer hit the structure every second with increasing intensity over 96 seconds; a significant increase in the kurtosis

and spikes on the spectrograms were observed. Figure 15 shows the time waveforms (green curve in normal operating condition and yellow curve with hammer tests) and the hammer forces (red curve) recorded during 96 seconds, the kurtosis and the short time Fourier transforms for several different conditions. Spectrograms on the top of the figure show the STFT resulting from normal operating condition; the kurtosis was 3.32 for each stage. Spectrograms on the bottom show the STFT for the extra light hammer hits (4 to 36 s), for light hammer hits (36 to 68 s) and strong hammer test (68 to 100 s), respectively. The time step and frequency resolution of the STFT were set to 0.125 s and 8 Hz, respectively.

The hammer hits experiment was performed on the coupled dynamometer (Figure 2). As the vibration level increases with the RPM, the hammer hits should be larger in high speed than in lower speed (to avoid being masked by the vibration level). The speed of the drive motor was selected to be 300 RPM (low speed) to avoid damage to the gearbox. Table 5 shows the amplitude change relative to the baseline (no hammer hits) of the kurtosis in the case of extra light, light and strong hammer hits.

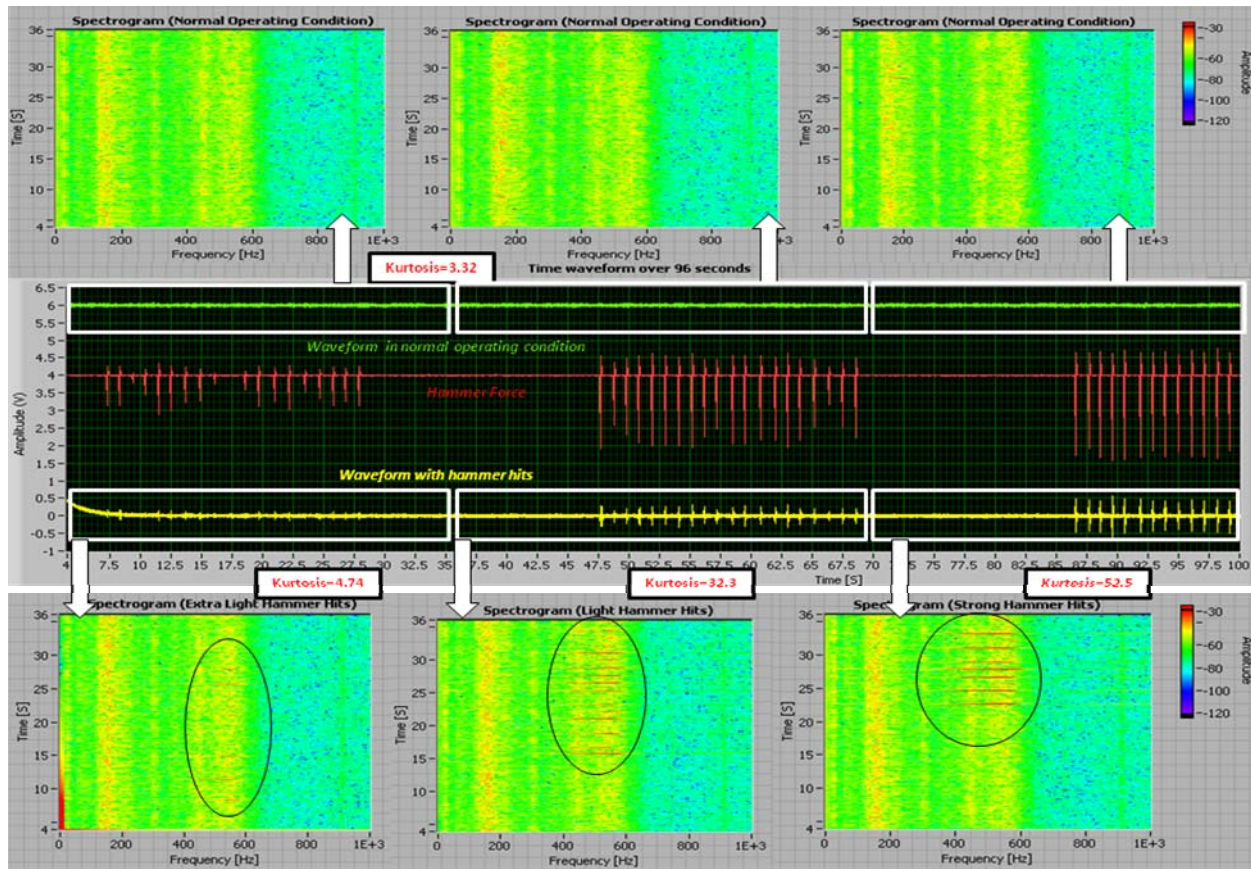


Figure 15. Spectrogram and kurtosis of the time waveform, for extra light (4-32 s), light (36-68 s) and strong hammer hit (68-100 s) at 300 RPM. Curve in red represents the hammer force. Yellow and green curves are the time waveform, with hammer test and in normal operating condition respectively.

	Relative Kurtosis
Extra light hammer hits	1.42
Light hammer hits	28.98
Strong hammer hits	49.18

Table 5. Relative kurtosis to the baseline for increasingly stronger hammer hits.

6. CONCLUSION

During the process of data acquisition and processing, several findings were made that are believed noteworthy:

- (1) Augmenting the imbalance caused the planet gear meshing frequency to be modulated by the output shaft speed of the second stage reduction gearbox (0.88 Hz for 1,144 RPM and 1.25Hz for 1595). The modulation level increased with increased imbalance.
- (2) The PSD was a better indicator of level change than the cepstrum, although the cepstrum is a better tool to identify harmonic relationships.
- (3) Envelope analysis using Hilbert transform techniques is a better indicator of modulation content than the PSD

and the cepstrum; this is consistent with reference (Fernandez et al., 2005). It may outperform the kurtosis analysis in the presence of transients.

(4) The kurtosis seems to be a good indicator for transient effects; the kurtosis had similar values with and without imbalance. However, had the imbalance been introduced while the shaft was rotating (transient), the value of the kurtosis would have changed significantly. Also, the kurtosis would have been increased if the imbalance had caused gear or bearing damages.

(5) The envelop analysis was performed on the planet gear meshing frequency. The results were similar to those found in (Fernandez et al., 2005; Yong-Han Kim et al., 2006) using bearing frequencies.

(6) The data comparisons indicate that the use of more than one technique for fault detection and identification increases the reliability of the conclusions. This might decrease the false alarms rate and the use of lower alarms levels, allowing for earlier fault detection.

In the light of these findings, the use of envelop and kurtosis analyses for detection of bearing and gear related

faults should be considered in addition to PSD levels. This allows for more reliable of fault identification and for evaluation of the severity of the problem.

REFERENCES

- Beaujean, P.-P. J., Khoshgoftaar, T. M., Sloan, J. C. Xiros, N. & Vendittis, D. (2009). Monitoring Ocean Turbines: a Reliability Assessment. In *Proceedings of the 15th ISSAT International Conference on Reliability and Quality in Design*.
- Driscoll, F. R., Skemp, S. H., Alsenas, G. M., Coley, C. J. & Leland A. (2008). Florida Center for Ocean Energy Technology. In *Proceedings of the MTS/IEEE Oceans'2008*.
- Fernandez, A., Bilbao J., Bediaga I., Gaston, A. & Hernandez, J., (2005). *Feasibility study on diagnostic methods for detection of bearing faults at an early stage*. In *Proceedings of the WSEAS Int. Conf. on dynamical systems and control*.
- Fan, X. & Zuo, M.J. (2006). *Gearbox fault detection using Hilbert and wavelet packet transform*. Mechanical Systems and Signal Processing, 20, 966–982.
- Singleton, K. (2006). *Case Study, Two Stages Planetary Gearbox*. www.vibration.org.
- Kim, Y.-H., Tan, A. C. C., Mathew, J. & Yang, B.-S. (2006). *Condition monitoring of low speed bearings: a comparative study of the ultrasound technique versus vibration measurements*. In *Proceedings of the World Congress on Engineering Asset Management*.
- Mjit, M. (2009). *Methodology for fault detection and diagnostics in an ocean turbine using vibration analysis and modeling*. Master's thesis. Florida Atlantic University.
- Mjit, M., Beaujean, P.-P.J. & Vendittis, D. J. (2010). *Fault Detection and Diagnostics in an Ocean Turbine using Vibration Analysis*. In *Proceedings of ASME IMECE10*.
- Reimche, W., Südmersen, U., Pietsch, O., Scheer, C., Bach, F.-W. (2003). *Basics of Vibration Monitoring For Fault Detection and Process Control*. In *Proceedings of the 111 Pan-American Conference for Non-Destructive Testing*.
- Sloan, J. C., Khoshgoftaar, T. M., Beaujean, P.-P. J. & Driscoll, F. (2009). Ocean Turbines – a Reliability Assessment. *International Journal of Reliability, Quality and Safety Engineering*, 16(5), 1-21.
- Tyagi, S. (2003). *Wavelet Analysis And Envelope Detection For Rolling Element Bearing Fault Diagnosis, A Comparative Study*. In *Proceedings*

of the 11th National Conference on Machines and Mechanisms.

- Wald, R., Khoshgoftaar, T. M., Beaujean, P.-P. J., & Sloan, J. C. (2010). Combining wavelet and Fourier transforms in reliability analysis of ocean systems. In *Proceedings of the 16th ISSAT International Reliability and Quality in Design Conference*.



Mustapha Mjit is an engineer in structural vibrations for the Southeast National Marine Renewable Energy Center at Florida Atlantic University. He earned his M.S. degree in Ocean Engineering at Florida Atlantic University (2009) and his Master's in Mechanical Engineering at the Université de Technologie de Compiègne, France (2006). In 2010, he received a fellowship award from Siemens Energy under the Gas Turbine Industrial fellowship Program. His professional experience includes vibration and health monitoring of rotating machines. Mustapha Mjit is a member of SNAME (2009), Tau Beta Pi Engineering Honor Society (2010) and Ordre des Ingénieurs du Québec (2010).



Dr. Pierre-Philippe J. Beaujean is an associate professor at Florida Atlantic University in the Department of Ocean and Mechanical Engineering. He received his Ph.D. in Ocean Engineering at Florida Atlantic University in 2001. He specializes in the field of underwater acoustics, signal processing, sonar design, data analysis, machine-health-monitoring and vibrations control. Dr. Beaujean is an active IEEE, ASA and MTS member.



Dr. David J. Vendittis [Ph.D. (Physics) - American University, 1973] is a Research Professor (part time) at the Center for Ocean Energy and Technology, Florida Atlantic University. Additionally, he is the Technical Advisory Group (ASA/S2) chairman for an International Standards Organization subcommittee, ISO/TC108/SC5 - Machinery Monitoring for Diagnostics. This committee writes international standards that support Machinery Condition Monitoring for Diagnostics. He was appointed to this position by the Acoustical Society of America.

Comparison of Parallel and Single Neural Networks in Heart Arrhythmia Detection by Using ECG Signal Analysis

Ensieh Sadat Hosseini Rooteh¹, Youmin Zhang^{2,3}, and Zhigang Tian³

^{1,2}*Department of Mechanical and Industrial Engineering, Concordia University, Montreal H3G 2W1, Canada*

e_hoss@encs.concordia.ca
ymzhang@encs.concordia.ca

³*Concordia Institute for Information Systems Engineering, Concordia University, Montreal H3G 2W1, Canada*

tian@ciise.concordia.ca

ABSTRACT

In this study, we have presented a method for detecting four common arrhythmias by using wavelet analysis along with the neural network algorithms. The method firstly includes the extraction of feature vectors with wavelet analysis. Then, the vectors will be categorized by means of the neural network into four classes. Input signals are recorded from two different leads. In addition, we have used both continuous and discrete wavelet analyses simultaneously for feature extraction. This results into increasing the accuracy of feature vectors extraction. Also, using the continuous wavelet in a specific scale can lead to better extraction of coefficients as well as more accurate data. In order to decrease the computational efforts and increase the training speed, the dimensions of the feature vectors have been reduced by substituting the wavelet coefficients with their statistical parameters. Furthermore, two approaches are introduced in classification of feature vectors. The first approach comprises four neural networks in the parallel form for detection of four classes, while the second approach makes use of one network for four classes. Numerical simulation results show that in comparison with the previous studies, the proposed methods are more accurate and faster. In addition, it is observed that the second approach has better capabilities in classification of data than the first one. On the other hand, the first approach is believed to have a good function for complicated data spaces.

1. INTRODUCTION

The most common way for studying and diagnosing cardiac dysfunctions is the Electrocardiogram (ECG) signal analysis. ECG is a record of the origin and the propagation of the electrical potential through cardiac muscles. The

Hosseini et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

normal ventricular complexes (N) are provoked by the sinus node and are related with regular conduction path through the ventricles, which assures their normal narrow waveform. The existence of ectopic centers, as well as some blocked regions in the ventricles changes the path propagation of the activation front and leads to generation of QRS complexes with wide and bizarre waveforms related to premature ventricular contractions (PVC) and left and right bundle branch blocks (LBBB, RBBB). Detection of these diseases by means of the convenient medical approaches is usually not easy and not accurate. On the other hand, signal analyses based on ECG signals has a big potential in the diagnosis.

Various methods are used for heart beat disease detection. Accuracy of detection depends on three basic factors – the used heartbeat feature set, the applied classification method and the organization of the training strategy.

The literature contains information about various feature extraction rules, including wavelet transform (Al-Fahoum and Howitt, 1999), (Shahidi Zandi and Moradi, 2006), (Ghaffari and Golbayani, 2008), Fourier transform (Minami, Nakajima, and Toyoshima, 1999) Lyapanov exponents (Ubeyli and Gular, 2004), (Casaleggio and Braiotta, 1997), independent component analysis (Sung-Nien and Kuan-To, 2007), (Wang, He, and Chen, 1997) principle component analysis (Ceylan and Ozbay, 2007) and also contains a lot of methods for classification such as neural network (Al-Nashash, 2000), (Foo, Stuart, Harvey, and Meyer-Baese, 2002) and neuro-fuzzy method (Engin and Demirag, 2003), (Engin, 2004), (Acharya, Bhat, Iyengar, Roo, and Dua, 2002) and K-th nearest neighbor (Christov, Jekova and Bortolan, 2005), (Jekova, Bortolan, and Chridstov, 2007), and mixture of experts (Hu, Palreddy and Tompkins, 1997) etc. In previous studies, selecting a powerful classifier was discussed and feature extraction stage was only a stage for reducing signal information. However, regarding to the neural network input data influence on the network performance,

the feature extraction stage is very important. If the feature vector determines the signal characteristics better and effectively shows the discrimination between patient signals. Then the classifier can serve better and subsequently the diagnosis processes will be done more accurate. Jekova, *et al.* (2007) used the geometrical parameters and discriminating features while their method was performed manually. Here, in the present study, the features are extracted using both continuous and discrete wavelet transforms and in order to have all of observable characteristics of signals they are recorded with two leads. It should be pointed out that in most relevant works which use the advantage of discrete wavelet transform for feature extraction while for reducing the dimension of the feature vectors they ignore the coefficients of some stages which leads to missing part of information through the signal. The statistical parameters are used to replace the coefficients of wavelet transform and finally the neural networks were used by two different methods for classifying signals to four classes. Lastly, the results of these two different methods in signal classification are compared with together and with some previous studies. The presented approach, in comparison to the existing methods, is demonstrated to detect heart arrhythmia accurate and efficient under the study conditions in this paper.

2. MATERIAL AND METHODS

2.1 ECG Signals

This study involves 8 ECG recording from the **MIT-BIH** (the MIT university arrhythmia signal database) arrhythmia database. Each recording has 30 min duration and includes two leads, the modified limb lead II as well as one of the modified leads V1, V2, V3, V4 or V5. The sampling frequency is 360 Hz and the resolution is 200 samples per mV. The study focuses on the classification of the four largest heartbeat classes in the **MIT-BIH** arrhythmia database: (1) normal beats (N); (2) premature ventricular contraction (PVC); (3) right bundle branch block (RBBB); (4) left bundle branch block (LBBB). All the recorded data from this website are labeled and it is clear that each signal is belonged to which four above classes. In the present study, each data is made of 200 alternative samples which make a heartbeat to involve P, QRS and T waves (they are three waves which make a complete heart beat.) which will be used through the neural network.

2.2 Wavelet Transform (WT)

The ECG signals are considered as representative signals of cardiac physiology which are useful in diagnosing cardiac disorders. The most complete way for displaying this information can perform spectral analysis. WT provides very general techniques which can be applied to many tasks in signal processing. One of the most important applications of WT is its ability for computing and manipulating of data

in compressed parameters which are often called features. Thus, the ECG signal, consisting of many data points, can be compressed into few parameters. These parameters characterize the behavior of the ECG signals. This feature uses a smaller number of parameters to represent the ECG signal which, particularly, is important for recognition and diagnostic purposes (Guler and Ubeyli, 2005). The continuous wavelet transform (CWT) of a continuous signal $x(t)$ is defined as:

$$CWT_x(\tau, a) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t-\tau}{a} \right) dt \quad (1)$$

where $\psi(t)$ is the mother wavelet, and a is the scale factor which can be thought as the inverse of frequency. As shown in Eq. (1), the mother wavelet $\psi(t)$ is scaled by a and shifted by τ to provide the basis of time-frequency representation of $x(t)$. Using the CWT, a time-scale (time-frequency) description of a signal, which is very useful to investigate the signal behavior in time and frequency domains simultaneously, is obtained (Shahidi Zandi and Moradi, 2006).

In discrete wavelet analysis, a multi-resolution formulation is used in wavelet analysis to decompose a signal event into finer and finer details. The procedure of multi-resolution decomposition of a signal $x[n]$ is schematically shown in Fig. 1.

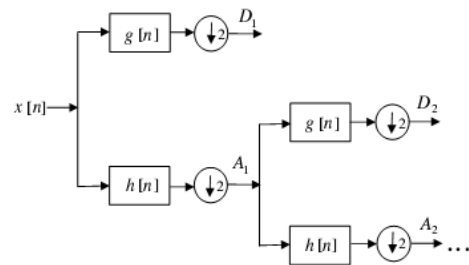


Fig. 1. Sub band decomposition of DWT implementation;

$g[n]$ is the high-pass filter and $h[n]$ is the low-pass filter.

Each stage of this scheme consists of two digital filters. The first filter $g[n]$ is the discrete mother wavelet, high-pass in nature, and the second, $h[n]$ is its mirror version, with low-pass in nature. The outputs of first decomposition stage are D_1 and A_1 , in which A_1 is further decomposed and this process is continued as shown in Fig. 1 (Guler and Ubeyli, 2005).

2.3 Neural Network Classifier

Artificial neural networks (ANNs) may be defined as structures comprised of densely interconnected adaptive simple processing elements (neurons) that are capable of performing massively parallel computations for data

processing and knowledge representation. ANNs can be trained to recognize patterns and the nonlinear models developed during training and allow neural networks to generalize what they have previously encountered. The multilayer perceptron neural networks (MLPNNs) are the most commonly used neural network architectures since their nice features such as ability to learn and to generalize, with smaller training set requirements, faster operation, and ease of implementation. A MLPNN consists of (1) an input layer with neurons representing input variables to the problem, (2) an output layer with neurons representing the dependent variables (what are being modeled), and (3) one or more hidden layers containing neurons to help to capture the nonlinearity in the data (Guler and Ubeyli, 2005). Fig. 2 shows a general structure of the MLPNNs.

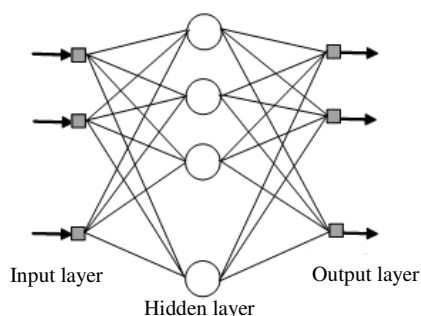


Fig. 2 The general structure of MLPNNs

3. EXPERIMENTAL RESULTS

3.1 Computation of Feature Vectors

In the present study, four various classes of ECG beats have been considered which are shown in Figs. 3(a)-(d). According to the fact that with inappropriate inputs even the best classifiers will give unacceptable results, then the selection of inputs for the neural network seems to be most important factor in designing a neural network for the patterns classification. In order to select appropriate data it should be noted that which elements of the pattern or which kind of the input data are the best description of the given data. Also it is possible that all information of a signal is not observable through a unit lead. Then, for having more information and reducing the possibility of data loss, in this study, for each heart signal two available leads from the MIT-BIH have been used. In addition, since we are eager to compare the results with each other, it is necessary to use a similar leads for all data. This matter has been considered within the records selection and all of the records have been described with two MLII and V1 leads.

Also for extraction of feature vectors, both continuous and discrete wavelet transform have been used. Continuous wavelet transform is used with Haar function and discrete wavelet is used with Daubechies function. Continuous wavelet transform with Haar function based on the Ghaffari

and Golbayani (2008) can extract some information about the shape of signal and if all the wave of signals occurred or not? Also discrete wavelet with Daubechies function based on the Ceylan and Ozbay (2007) can extract some information about the sudden changes in the signal rhythm. Using these two transform simultaneously helps to extract more information from the signals.

After the wavelet transform the statistical parameters like: max, mean and standard deviation are used to compact the information of continuous and discrete wavelet coefficient more. Then feature vectors with dimension 36×1 are made.

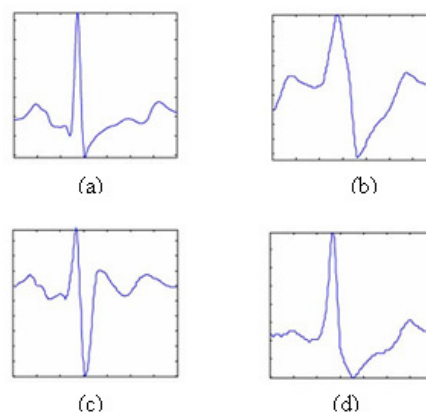


Fig. 3 (a) Normal beat (b) Premature ventricular contraction (c) Right bundle branch block (d) Left bundle branch block

3.2 Applying Neural Network on ECG Signals

In this study 110 signals are used as the test signals as shown in Table 1:

Signal Type	Number of Test Signals
N	25
PVC	25
RBBB	30
LBBB	30

Table 1- Number of test signals

These signals will classified with neural networks by two different methods. Each of these methods is explained as follows:

Method 1- Four neural networks are considered for data classification and each of these neural networks diagnoses one class of signals. For instance a neural network diagnoses normal signals and this network divides all data into two classes: 1- normal signals and 2- abnormal signals. The first network is called normal network. The second

neural network is called PVC network and is dividing signals in two classes: 1-PVC signals and 2-other signals. The third network is called RBBB network and is dividing signals into two classes: 1-RBBB signals and 2-other signals. The fourth network is called LBBB network and is dividing signals into two classes as 1-LBBB signals and 2-other signals. All neural networks have three layers: input layer, hidden layer and output layer. Normal, PVC and RBBB networks have 36 neurons in input layer and 8 neurons in hidden layer and 2 neurons in output layer. LBBB network has 36 neurons in input layer and 12 neurons in hidden layer and 2 neurons in output layer.

The test results of these four neural networks are given as follows:

		Output of Normal Network	
Signal Type	Number of test signals	Normal signals	Other signals
Normal signals	25	25	0
Other signals	85	1	84

Table 2- Confusion* Matrix for Normal Network

This table shows that 25+85 signals are tested with the normal network and all of the 25 normal signals are detected as normal signal correctly and also 1 signal which is not normal is detected as a normal signal wrongly. This shows that the normal network has high separation ability in separating normal signals from the other signals.

		Output of PVC Network	
Signal Type	Number of Test signals	PVC Signals	Other Signals
PVC Signals	25	23	2
Other Signals	85	4	81

Table 3- Confusion Matrix for PVC Network

This table shows that 25+85 signals are tested with the PVC network and 23 of 25 PVC signals are detected as PVC signal correctly and also 4 signals which are not PVC are detected as PVC signal wrongly.

		Output of RBBB Network	
Signal Type	Number of Test Signals	RBBB Signals	Other Signals
RBBB Signals	30	30	0
Other Signals	80	2	78

Table 4- Confusion Matrix for RBBB Network

This table shows that 30+80 signals are tested with the RBBB network and all of the 30 RBBB signals are detected as RBBB signal correctly and also 2 signals which are not RBBB are detected wrongly. This shows that the RBBB network has high separation ability in separating RBBB signals from the other signals.

		Output of LBBB Network	
Signal Type	Number of Test Signals	LBBB Signal	Other Signals
LBBB Signal	30	29	1
Other Signals	80	1	79

Table 5- Confusion Matrix for LBBB Network

This table shows that 30+80 signals are tested with the LBBB network and 29 of 30 LBBB signals are detected as LBBB signal correctly and also 1 signal which is not LBBB is detected as LBBB signal wrongly.

For each network two different accuracies are determined as:

- 1- Specific accuracy: This shows the network accuracy in detecting the signals of its class. It is obtained for example for normal network by dividing number of signals which they detected normal to the number of tested signals which they are normal. Then for normal network this accuracy will be 100% (25/25).
- 2- Total accuracy: This shows the network accuracy in detecting the signals for both two classes. It is obtained by dividing the number of signals which they are detected correct to the number of total signals. For example for normal network it will be 99%.

* Confusion matrix is a visualization tool typically used in supervised learning . Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class.

Table 6 shows these two accuracies for each neural network:

Network	Specific Accuracy (%)	Total Accuracy (%)
Normal	100	99
PVC	92	94.5
RBBB	100	98.2
LBBB	96.7	98.2

Table 6- Neural Network Accuracies

The other important issue is the network ability in data separation. Network separation ability for example about normal network is to determine how many signals of PVC, RBBB, LBBB signals are correctly detected and assigned to abnormal signal class. For calculating this item, the number of signals which detected will bring by more details in the table below:

Network	Number of Signals			
	Normal	PVC	RBBB	LBBB
Normal	-----	24	30	30
PVC	23	-----	29	29
RBBB	25	23	-----	30
LBBB	25	24	30	-----

Table 7- Number of Signals which correctly detected

The separation abilities for the networks are in the table below:

Network	Signal Class			
	Normal	PVC	RBBB	LBBB
Normal	-----	96%	100%	100%
PVC	92%	-----	96.7%	96.7%
RBBB	100%	92%	-----	100%
LBBB	100%	96%	100%	-----

Table 8 - Result of Network Separation Ability

From the recorded results in the above table it can be seen that the accuracy of the normal network in separating PVC signals from normal signals is 96%. It means that 24 signals of 25 signals of PVC class are correctly assigned to abnormal signal class. Also the accuracy for separating RBBB and LBBB signals from normal signals is 100%.

By using neural network in parallel form, after training of 4 networks the test vectors are fed to all four networks and the class of each test signal is determined by these four outputs. It can often happen that a signal will be detected by two networks. For final classification a logical decision must be helpful to detect a correct class for signal. In this study three methods for this logical decision are explained as below:

- A- If a signal is only detected by a network, this signal is belonging to the class of this network. If a signal is detected by two or three networks simultaneously, this signal is considered as an unclassified signal. Therefore, in this method the signals are either detected correctly or wrongly, or remained unclassified.
- B- In this method, the class of a signal is determined according to the more accurate network's detection. For example if a signal is detected by two normal network and PVC network, by considering that the specific accuracy of normal network is 100% and this accuracy for PVC network is 92% then it will be concluded that the signal is normal.
- C- This method is based on separation ability of the neural networks. On the other hand, if a signal is detected by two networks simultaneously, the signal is assigned to the class of network with higher separating ability. For example if a signal is detected by two normal and PVC networks, this signal is assigned to normal signal class. This detection is because of difference between the separation ability of normal neural network in separating normal signals from PVC signals (96%) and this ability for PVC network (92%).

As the above explanations it is clearly seen that this method of classification leads to reduction in classification error. We are using the neural networks in the parallel form. It means that each signal is fed to all networks for class detection. Then, the networks can cover their weaknesses and therefore the final result will be more accurate. The results of implementation of these three logical decisions on four network outputs are as follows:

Signal Classes	Detected Class				
	Normal	PVC	RBBB	LBBB	Unclassified
Normal	24	-----	-----	-----	1
PVC	1	21	1	-----	2
RBBB	-----	-----	29	-----	1
LBBB	-----	1	-----	29	-----

Table 9- Confusion Matrix for result of Method A

Signal Classes	Detected Class			
	Normal	PVC	RBBB	LBBB
Normal	25	-----	-----	-----
PVC	1	21	2	1
RBBB	-----	-----	30	-----
LBBB	-----	1	-----	29

Table 10- Confusion Matrix for result of Method B

Signal Classes	Detected Class			
	Normal	PVC	RBBB	LBBB
Normal	25	-----	-----	-----
PVC	1	23	1	-----
RBBB	-----	1	29	-----
LBBB	-----	1	-----	29

Table 11- Confusion Matrix for result of Method C

For evaluating the classification methods some statistical parameters are defined as follows:

$$1- \text{Specificity } (Sp_i) = \frac{TN_i}{TN_i + FP_i}$$

$$2- \text{Sensitivity } (Se_i) = \frac{TP_i}{TP_i + FN_i}$$

where TP_i (true positive) is the number of heartbeats of the i th class, which are correctly classified, TN_i (true negative) is the number of heartbeats which is not belonging to and classified in the i th class, FP_i (false positive) is the number of heartbeats classified erroneously in the i th class and finally FN_i (false negative) is the number of heartbeats of i th class which is classified in a different class. These statistical parameters for three methods are showed in the Tables 12-14 below:

Network	Sp_i %	Se_i %
Normal	98.8	96
PVC	98.8	84
RBBB	98.8	96.7
LBBB	100	96.7

Table 12- The Result of Method A

Network	Sp_i %	Se_i %
Normal	98.8	100
PVC	98.8	84
RBBB	97.5	100
LBBB	98.8	96.7

Table 13- The Result of Method B

Network	Sp_i %	Se_i %
Normal	98.8	100
PVC	97.6	92
RBBB	98.8	96.7
LBBB	100	96.7

Table 14- The Result of Method C

In Tables 15-16, comparison results of three methods in terms of specificity and sensitivity are presented. As shown in Table 15, all of these three methods have same specificity

ability for normal signal. Methods A and B have better results for PVC signals and also methods A and C have better results for RBBB and LBBB signals. Overall, method A shows the best results.

Signal Type	Sp_i (%)		
	Method A	Method B	Method C
Normal	98.8	98.8	98.8
PVC	98.8	98.8	97.6
RBBB	98.8	97.5	98.8
LBBB	100	98.8	100

Table 15- Comparison Three Methods in Specificity Factor

The sensitivity of three methods is compared in Table 16 below:

Signal Type	Se_i (%)		
	Method A	Method B	Method C
Normal	96	100	100
PVC	84	84	92
RBBB	96.7	100	96.7
LBBB	96.7	96.7	96.7

Table 16- Comparison Three Methods in Sensitivity Factor

In summary with considering these given parameters and accuracy parameter, it can be concluded that method C provides the best performance based on the separating ability.

Method 2- For classifying data in four classes by neural network, a MLPNNs with three layers is considered, having 36 neurons in input layer, 12 neurons in hidden layer and 4 neurons in output layer. The outputs of neural network for four classes are assigned to four target vectors as follows: normal signal (1,0, 0, 0) , premature ventricular contraction (0,1,0,0) , right bundle branch block (0,0,1,0) and left bundle branch block (0,0,0,1). The training method of neural network is chosen to be back propagation error. For increasing the learning speed the Levenberg-Marquardt method has been used. The results of neural network training are described in Confusion matrix as below.

Signal type	Number of signal	Neural network output			
		N	PVC	RBBB	LBBB
N	25	25	0	0	0
PVC	25	1	24	0	0
RBBB	30	0	0	30	0
LBBB	30	0	1	0	29

Table 17- Confusion Matrix

According to the Confusion matrix it is observed that all of the normal signals and the right bundle branch block signals are diagnosed correctly but one of the signals between premature ventricular contractions is diagnosed incorrectly and is assigned to normal signals class. In addition, one of right bundle branch signals is also diagnosed incorrectly and assigned to be in premature ventricular contraction class.

The statistical parameters are computed for 4 classes and are listed in Table 18.

Signal type	Sp_i (%)	Se_i (%)
N	98.8	100
PVC	98.8	96
RBBB	100	100
LBBB	100	96.7

Table 18- Statistical Parameter Value of Neural Network Performance

4. DISCUSSION AND CONCLUSION

In this paper, wavelet transform and neural network are used for heart arrhythmia signal classification. Selected signals are belonging to four different classes and signals are recorded from two leads (MLII & V1). Wavelet transform is used for feature extraction and then feature vectors are classified by two different methods by using neural networks. The key results of these two methods are compared in Tables 19 and 20.

Signal Type	Se_i (%)			
	Method 1			Method 2
	Method A	Method B	Method C	
Normal	96	100	100	100
PVC	84	84	92	96
RBBB	96.7	100	96.7	100
LBBB	96.7	96.7	96.7	96.7

Table 19- Methods Comparison in Sensitivity

Signal Type	Sp_i (%)			
	Method 1			Method 2
	Method A	Method B	Method C	
Normal	98.8	98.8	98.8	98.8
PVC	98.8	98.8	97.6	98.8
RBBB	98.8	97.5	98.8	100
LBBB	100	98.8	100	100

Table 20- Methods Comparison in Specificity

According to Tables 19 and 20 the results of second method is better than first method which has three shapes. Therefore the second method of signal classification, which uses a neural network for signal classification, is more accurate than first method.

The key results of the second method are compared with previous study in Table 21.

	This study(Method2)	Sung-Nien [9]
	<i>WT-BPNN method</i>	<i>ICA-PNN method</i>
Signal type	Sp_i (%)	Sp_i (%)
N	98.8	99.9
PVC	98.8	98
RBBB	100	99.97
LBBB	100	99.65

Table 21- comparison of the second method results with the pervious study

According to Table 21 the results of the second method are much better than the previous study Sung-Nien and Kuan-To (2007) in three types of signals: premature ventricular contraction, right bundle branch block and left bundle branch block. According to the same method for classification in two studies the difference between their results is because of different feature extraction methods. Therefore the feature extraction method that is used in this study is a better method to determine signal characteristics.

The results of the second method are compared with the Ceylen and Ozbay (2007) study in Table 22. In their study such as second method of this study, wavelet is used feature extraction and neural network is used for classification. It is clearly to see that the second method of this study serves more effectively than the previous one. The use of statistical indices of wavelet coefficients in second method of this

study provides considerable increase in training speed and the accuracy of diagnosis.

	This Study(Method2)	Ceylen-Ozbay's Study
Test error	0.158	0.4
CPU time (s)	10	85.44

Table 22- Comparison of second method with pervious study

REFERENCES

- Acharya, R. , Bhat, P. S., Iyengar, S. S. , Roo, A. & Dua, S. (2002), Classification of heart rate data using artificial neural network and fuzzy equivalence relation. *The Journal of the Pattern Recognition Society*, 36 , 61 – 68.
- Al-Fahoum, A.S. & Howitt, I. (1999), Combined wavelet transformation and radial basis neural networks for classifying life-threatening cardiac arrhythmias. *Medical and Biological Engineering and Computing*, 37: 566–573.
- Al-Nashash, H. (2000), Cardiac arrhythmia classification using neural networks. *Techno Health Care*, 8:363–72.
- Casaleggio, A., Braiotta, S. (1997). Estimation of Lyapunov exponents of ECG time series-the influence of parameters. *Chaos, Solitons & Fractals*, 8 (10):1591-1599.
- Ceylan, R., Ozbay, Y. (2007), Comparison of FCM, PCA and WT techniques for classification ECG arrhythmias using artificial neural network. *Expert Systems with Applications*, 33: 286–295.
- Christov, I. , Jekova, I., Bortolan, G. (2005). Premature ventricular contraction classification by the Kth nearest neighbors rule. *Physiol Meas*, 26:123–30.
- Engin, M. & Demirag, S. (2003). Fuzzy-hybrid neural network based ECG beat recognition using three different types of feature set. *Cardiovascular Engineering: An International Journal*, 3(2): 1–80.
- Engin, M. (2004). ECG beat classification using neuro-fuzzy network. *Pattern Recognition Letters*, 25: 1715–1722.
- Foo, S. Y., Stuart, G., Harvey, B. & Meyer-Baese, A. (2002). Neural network-based ECG pattern recognition. *Engineering Applications of Artificial Intelligence*, 15: 253–260.
- Ghaffari , A., Golbayani, H. (2008), A new mathematical based QRS detector using continuous wavelet transform, *Computer & Electrical Engineering*, 34(2): 81-91.
- Guler, I., Ubeyli, E. (2005), A modified mixture of experts network structure for ECG beats classification with diverse features. *Engineering Application of Artificial Intelligence*, 18: 845-856.
- Hu, Y.H. , Palreddy, S. & Tompkins, W.J. (1997). A Patient-adaptable ECG beat classifier using a mixture of experts approach. *IEEE Transactions on Biomedical Engineering*,44(9): 891–900.
- Jekova, I., Bortolan, G., Chridstov, I. (2007). Assessment and comparison of different methods For heartbeat classification. *Med Eng Phys* (2007).
- Minami, K., Nakajima, H. & Toyoshima, T. (1999). Real-time discrimination of ventricular tachyarrhythmia with Fourier transform neural network. *IEEE Transactions on Biomedical Engineering*, 46(2): 179–185.
- Shahidi Zandi, A., Moradi, M. H. (2006), Quantitative evaluation of a wavelet-based method in ventricular late potential detection. *Pattern Recognition*, 39: 1369-1379.
- Sung-Nien, Y., Kuan-To, Ch. (2007), Integration of independent component analysis and neural networks for ECG beat classification. *Expert Systems with Applications*, 34, 2841–2846.
- Ubeyli, E.D., Gular, I. (2004). Detection of electrocardiographic changes in partial epileptic patients using Lyapunov exponents with multilayer perceptron neural networks. *Engineering Applications of Artificial Intelligence*, 17(6): 567-576.
- Wang, Z. , He, Z. & Chen, J. Z. (1997). Blind EGG separation using ICA neural networks. In *Proceedings-19th annual international conference of the IEEE-EMBS*, Vol. 3 (pp. 1351–1354). Chicago, IL, USA.



Ensieh Sadat Hosseini Rooteh is a PhD student in the Department of Mechanical and Industrial Engineering at Concordia University, Montreal, Canada. She received her Master of Science degree in 2009 from the Department of Mechanical Engineering, Khaje Nasir Toosi University of Technology, Tehran, Iran. Earlier in 2006 she graduated in Bachelor of Science in Mechanical Engineering from Tehran University, Tehran, Iran. She joined Cardiovascular Research Group (CRG) at KNT University of Technology

in 2006. Her research interests include control and its application in biomechanics, fault diagnosis, pattern recognition and signal processing.



Youmin Zhang received his Ph.D. degree in 1995 from the Department of Automatic Control, Northwestern Polytechnical University, Xian, China. He is currently an Associate Professor with the Department of Mechanical and Industrial Engineering at Concordia University, Montreal, Canada. He held several teaching and research positions in Northwestern Polytechnical University, University of New Orleans, Louisiana State University, State University of New York at Binghamton, The University of Western Ontario, and Aalborg University, respectively. His main research interests and experience are in the areas of condition monitoring, fault diagnosis and fault-tolerant (flight) control systems; cooperative guidance, navigation and control of unmanned aerial/ground vehicles; dynamic systems modeling, estimation, identification and control; and advanced signal processing techniques for diagnosis, prognosis and health management of safety-critical systems and manufacturing processes. He has published 4 books, over 200 journal and conference papers.

He is a senior member of AIAA, a senior member of IEEE, a member of IFAC Technical Committee on Fault Detection, Supervision and Safety for Technical Processes (SAFEPROCESS) and AIAA Infotech@Aerospace Program Committee on Unmanned Systems, and a member of Association for Unmanned Vehicle Systems International (AUVSI), Unmanned Systems Canada (USC), Canadian Aeronautics and Space Institute (CASI), Canadian Society for Mechanical Engineering (CSME). He is an editorial board member of several international journals and an IPC member and session chair/co-chair of many international conferences.



Zhigang Tian is currently an Assistant Professor at Concordia Institute for Information Systems Engineering at Concordia University, Montreal, Canada. He received his Ph.D. degree in 2007 in Mechanical Engineering at the University of Alberta, Canada; and his M.S. degree in 2003, and B.S. degree in 2000 both in Mechanical Engineering at Dalian University of Technology, China. His research interests focus on reliability analysis and optimization, prognostics, condition monitoring, and maintenance optimization. He is a member of IIE and INFORMS.

Condition Based Maintenance Optimization for Multi-component Systems for Cost Minimization

Zhigang Tian¹, Youmin Zhang², and Jialin Cheng³

^{1,3}*Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Quebec, H3G2W1, Canada*

tian@ciise.concordia.ca

²*Department of Mechanical and Industrial Engineering, Concordia University, Montreal, Quebec, H3G2W1, Canada*

ymzhang@encs.concordia.ca

ABSTRACT

Most existing condition based maintenance (CBM) work reported in the literature only focuses on determining the optimal CBM policy for single units. Replacement and other maintenance decisions are made independently for each component, based on the component's age, condition monitoring data and the CBM policy. In this paper, a CBM optimization method is proposed for multi-component systems, where economic dependency exists among the components subject to condition monitoring. In a multi-component system, due to the existence of economic dependency, it might be more cost-effective to replace multiple components at the same time rather than making maintenance decisions on components separately. Deterioration of a multi-component system is represented by a conditional failure probability value, which is calculated based on the predicted failure time distributions of components. The proposed CBM policy is defined by two failure probability thresholds. A method is developed to obtain the optimal threshold values in order to minimize the long-term maintenance cost. An example is used to demonstrate the proposed multi-component CBM method.

1. INTRODUCTION

Condition based maintenance (CBM) generally aims to determine the optimal maintenance policy to minimize the overall maintenance cost based on condition monitoring information. The health condition of a piece of equipment is monitored and predicted via collecting and analyzing the inspection data, such as vibration data, acoustic emission data, oil analysis data and temperature data. Various CBM policies and optimization methods have been proposed (Banjevic et al, 2001, Jardine et al, 2006). However, most existing condition based maintenance (CBM) work reported in the literature only focuses on determining the optimal CBM policy for single units. Replacement and other

maintenance decisions are made independently for each component, based on the component's age, condition monitoring data and the CBM policy.

For multi-component systems which involve multiple components, economic dependency exists among the components subject to condition monitoring. For example, in the replacement of bearings on a set of pumps at a remote location, the fixed maintenance cost, such as sending a maintenance team to the site, is incurred whenever a preventive replacement is performed. Thus, for multi-component systems, it might be more cost-effective to replace multiple components at the same time rather than making maintenance decisions on components separately. Tian and Liao (2011b) developed a proportional hazards model based approach for CBM of multi-component systems. In this paper, we propose an approach which can utilize prediction information from more general prediction tools. More specifically, the proposed CBM can be used as long as the prediction tool can produce predicted failure time values and their associated uncertainty information. The simulation-based cost evaluation method is presented. An example is used to illustrate the proposed approach.

2. COMPONENT HEALTH CONDITION PREDICTION

The output of component health condition prediction is the predicted failure time values and the associated uncertainty information. That is, at a certain inspection point, health condition prediction tools can generate the predicted failure time distribution. In this section, we present a method that can be used for generating the predicted failure time distribution.

Suppose, at a certain inspection point, the age of the component is t , the predicted failure time is $T_{n,t}$, and the actual failure time of the component is T_m , where subscript n indicates a predicted failure time while the subscript m indicates the actual failure time. The prediction error is

Zhigang Tian et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>).

defined in this paper as $e_{n,t} = (T_{n,t} - T_m)/T_m$. We also define the life percentage as $p_t = t/T_m$. The prediction error is a measure of the prediction accuracy. To obtain the predicted failure time distribution, Tian et al (2010, 2011a) developed a method to calculate the standard deviation of the predicted failure time, while using the artificial neural network (ANN) prediction model. The basic idea is that the ANN life percentage prediction errors can be obtained during the ANN training and testing processes, based on which the mean, μ_p , and standard deviation, σ_p , of the ANN life percentage prediction error can be estimated. These values can be used to build the predicted failure time distribution at a certain inspection point. Suppose the component age is t and the ANN life percentage output is P_t . The predicted failure time will be $t/(P_t - \mu_p)$, and the standard deviation of the predicted failure time will be $\sigma_p \cdot t/(P_t - \mu_p)$. That is, the predicted failure time T_p at the current inspection point follows the normal distribution as:

$$T_p \sim N\left(t/(P_t - \mu_p), \sigma_p \cdot t/(P_t - \mu_p)\right). \quad (1)$$

It is assumed that the ANN life percentage prediction error follows normal distribution, and the predicted failure time at a certain inspection point also follows normal distribution. It is also assumed that the standard deviation of the ANN life percentage prediction errors is constant and does not change over time.

3. THE MULTI-COMPONENT CBM APPROACH

In this section, we present the CBM policy for multi-component systems, and the cost evaluation method for the CBM policy.

3.1 The CBM policy

In multi-component systems, the conditional probability Pr^* is used to determine not only when and also which components should be preventively replaced at each inspection time. The CBM policy for multi-component systems is proposed as below:

- 1) Identify the number of components in multi-component systems.
- 2) Regularly inspect these components which are subjected to condition monitoring. Calculate the predictive failure probability of each component at each inspection time based on the prediction method.
- 3) When a component's predicted failure probability Pr exceeds the level-1 threshold value Pr_1^* , preventively replace the component.
- 4) When a component fails, replace it by a new one.
- 5) When there is a preventive replacement or a failure replacement performed on any component in the

system, simultaneously replace other components if their Pr values exceed the level-2 threshold value Pr_2^* .

At each inspection time, one of the following events takes place exclusively for each component i :

1. Component i reaches $Pr_1^* \rightarrow$ a preventive replacement is performed on i .
2. Component i reaches Pr_2^* if there is a failure replacement or a preventive replacement that needs to be performed on one of the components in the multi-component systems \rightarrow preventively replace component i simultaneously.
3. Component i fails \rightarrow a failure replacement is performed, the component is replaced by a new one.
4. None of the above \rightarrow component i continues its normal operation.

3.2 A simulation method for cost evaluation

In this work, a simulation method is used to find the optimal condition failure probability threshold value which corresponds to the minimum expected replacement cost. We assume that there are N components in the multi-component systems. The procedure of the simulation method for CBM policy cost evaluation is shown in Figure 1, and is discussed in details as follows.

Step 1: Define the maximum simulation iteration.

Set the maximum simulation iteration NT , for example, 100,000 inspection points. It means we start from inspection point 0 and end with inspection points 100,000. Between each inspection point, there is a fixed inspection interval L , say, 20 days.

Step 2: Generate a random failure time as the actual failure time of each component.

At the starting point of a new life cycle of component i , generate a random failure time, FT_i , which follows Weibull distribution with the parameters α, β .

Step 3: Generate a random predicted failure time of a component.

At inspection point k ($k = 0, \dots, NT$), generate a random predicted failure time for component i based on ANN remaining useful life (RUL) prediction error. In a simulation process, this random predicted failure time simulate the predicted result based on ANN model using condition monitoring data at each inspection time. The predicted lifetime is denoted by PT_{ki} and follows normal distribution:

$$PT_{ki} \sim N(\mu, \sigma^2) \quad (k = 0, \dots, NT; i = 1, \dots, N) \quad (2)$$

where $\mu = FT_i$, $\sigma = \sigma_p \times FT_i$, σ_p is standard deviation of the remaining useful life prediction error.

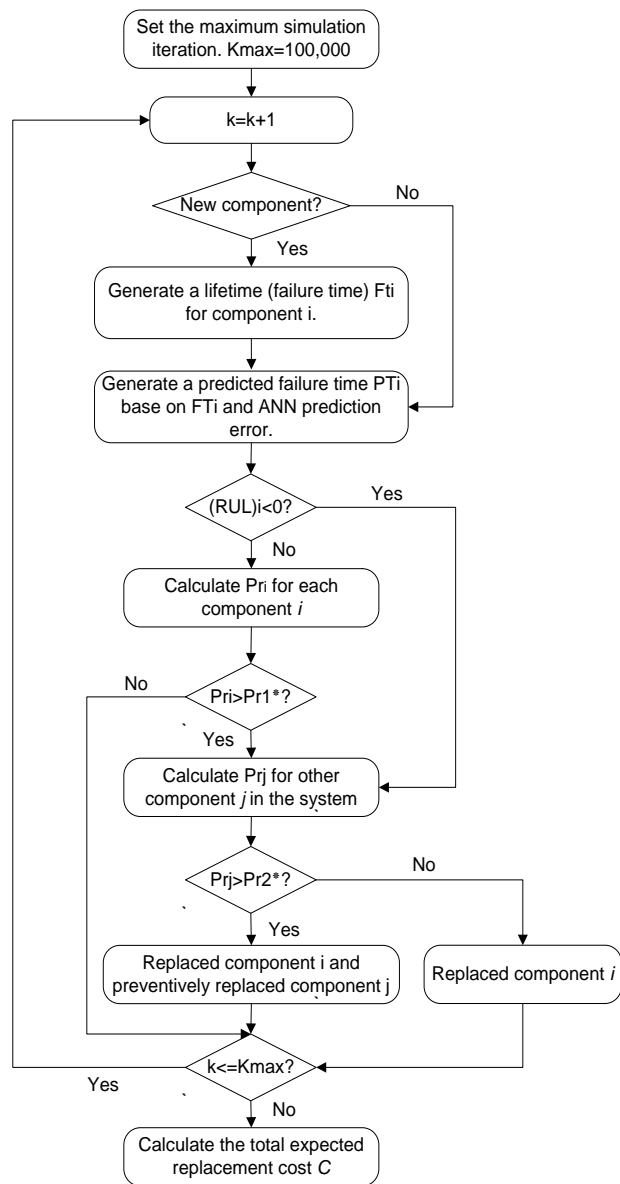


Figure 1. The procedure of the simulation method for cost evaluation in multi-component

Step 4: calculate the failure probability.

During a lifetime of component i , calculate conditional failure probability $P_{r_{ki}}$ in each inspection point by using equation below (Tian et al, 2011a):

$$P_{r_{ki}} = \frac{\int_{t_i}^{t_i+L} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t_i-\mu)^2}{2\sigma^2}} dt}{\int_{t_i}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t_i-\mu)^2}{2\sigma^2}} dt}$$

$(k = 0, \dots, NT; i = 1, \dots, N)$

where t_i is cumulated inspection time of component i in one life cycle, L is the constant inspection interval, μ is the predicted failure time of different component at different inspection point of time PT_{ki} , and $\sigma = \sigma_p \times FT_i$, where σ_p is standard deviation of ANN RUL prediction error. The failure probability in the formula above basically refers to the conditional failure probability during the next inspection interval given that it is still working now.

If $P_{r_{ki}}$ is greater than the level-1 condition failure probability threshold Pr_1^* , preventively replace the component at inspection point k . If there is no preventive replacement performed during the lifetime of the component, perform failure replacement at the inspection point just past the generated failure time FT_i . When there is a preventive replacement or a failure replace taking place at inspection time k , check other components in the system, if $P_{r_{kj}}$ ($j = 1, \dots, N$) is greater than the level-2 failure probability threshold Pr_2^* , perform preventive replacement on component j simultaneously.

We also introduce two variables to represent the stature of the component i in the multi-component systems:

$$\Delta_{p_{ki}} = \begin{cases} 1 & \text{Component } i \text{ is preventively replaced;} \\ 0 & \text{No preventive replacement} \end{cases}$$

$$\Delta_{f_{ki}} = \begin{cases} 1 & \text{Failure replacement on Component } i \\ 0 & \text{No failure replacement on component } i \end{cases}$$

If $\Delta_{p_{ki}} = 0$ & $\Delta_{f_{ki}} = 0$, component i continues its normal operation.

Step 5: New life cycle starts.

Start a new life cycle of component i after a preventive or a failure replacement taking place, go back to Step 2 and set the cumulated inspection time, t_i , equal to 0. The iteration would not stop until the maximum simulation iteration is reached.

Step 6: Estimation of the total expected replacement cost.

The expected replacement cost for multi-component system can be obtained by the following equation:

$$C_r = \frac{\text{Cost}_{total}}{\text{Time}_{total}} = \frac{\sum_{k=0}^{NT} C_k}{NT \times L} \text{ (\$/day)} \quad (3)$$

where C_k is the total cost occurs at inspection point k , NT is the total inspection point of the simulation process, and L is the inspection interval.

$$C_k = C_f \cdot \sum_{i=1}^N \Delta_{fki} + C_p \cdot \sum_{i=1}^N \Delta_{pki} + I(\Delta_{pki}) \cdot C_{p0} \quad (4)$$

where $I(\Delta_{pki}) = 1$, when $\sum_{i=1}^N \Delta_{pki} \geq 1$ & $\sum_{i=1}^N \Delta_{fki} = 0$; otherwise $I(\Delta_{pki}) = 0$. N is the number of components under condition monitoring, C_{p0} is fixed preventive replacement cost, which does not change with respect to the number of components being maintained. C_p is variable preventive replacement cost, and C_f is failure replacement cost.

At inspection point k , C_k can be in one of three possible circumstances as follows:

$C_k = C_{p0} + nC_p$, ($1 \leq n \leq N$), if there is at least one preventive replacement needed but no failure replacement;

$C_k = mC_f + nC_p$, ($1 \leq m \leq N, 0 \leq n \leq N - 1$), if there are at least one failure replacement and n preventive replacement performed;

$C_k = 0$, if neither preventive replacement nor failure replacement is needed.

Step 7: Determining the optimal CBM policy for multi-component systems.

The two-level predicted failure probability threshold values, which are defined in Section 3.1, are decision variables in the CBM policy for multi-component systems. The minimum calculated replacement cost corresponding to the predicted failure probability threshold value Pr_1^* and Pr_2^* . So once Pr_1^* and Pr_2^* are determined, the CBM policy is determined.

3.3 The CBM optimization model

The objective of the CBM optimization is to determine the optimal failure probability threshold values to minimize the long-run expected replacement cost. The optimization model can be formulated as below:

$$\min C_r(Pr_1^*, Pr_2^*) \quad (5)$$

s. t.

$$C_r \leq C_0, Pr_1^* \geq Pr_2^* \geq 0$$

where C_0 is the cost constraint value, Pr_1^* and Pr_2^* are Level-1 and Level-2 failure probability threshold values and also are the CBM optimization decision variables.

4. EXAMPLE

In this section, we present an example based on bearing vibration monitoring data collected from bearings on a group of Gould pumps at a Canadian kraft pulp mill company (Stevens 2006). We use totally 24 bearing histories which were examined at 8 pump locations, embracing 10 bearing failure histories and 14 suspension histories. For each pump, seven types of measurements were recorded: five different vibration frequency bands (8*5),

and the overall vibration reading (8*1) plus the bearing's acceleration data (8*1). So the original inspection data includes 56 (8*5+8*1+8*1) vibration measurements at each time. More information on the example can also be found in Tian et al (2010).

The software EXAKT was used to conduct the significance analysis for the 56 vibration measurements (Stevens 2006). Two of the variables were identified as having significant influence on the health of bearings. Then we use these two measurements and the age values of the components as the inputs of the ANN model. Constant usage rate is assumed here. 5 failure histories and 10 suspension histories are used as ANN training inputs and the other 5 failure histories are used as test histories. After comparing the predicted lifetime with the actual lifetimes, we found that the prediction error follows the normal distribution. The mean of prediction error is 0.1385 and the standard deviation is 0.1429.

For multi-component systems, level-1 and level-2 probability thresholds are two decision variables to determine the optimal CBM policy, and therefore, the expected replacement cost of certain CBM policy can be evaluated by giving certain probability threshold values Pr_1^* and Pr_2^* . In this case, we consider a multi-component system consisting of 5 identical bearings which are operating in parallel and subject to random failures. The lifetimes of the individual components are independent random variables and are identically distributed as Weibull distribution with parameters $\alpha = 1386.3$, $\beta = 1.8$.

The simulation procedure is as follows:

Step 1: Set the maximum simulation inspection point as 100,000. Between each inspection point, the fixed inspection interval, L , equals 20 days.

Step 2: At the starting point of each iteration for component i ($i = 1, \dots, 5$), set t_i equal 0, generate a random failure time, FT_i , of the component which follows Weibull distribution.

Step 3: At inspection point k ($k = 0, \dots, 100,000$), generate a random predicted failure time, PT_{ki} , of the component i , based on the ANN RUL prediction error. PT_{ki} follows a normal distribution. In this case: $\mu_i = FT_i$, $\sigma = \sigma_p \times FT_i$, where σ_p is standard deviation of ANN RUL prediction error. Thus, we have

$$PT_{ki} \sim N(FT_i, (0.1429 \times FT_i)^2) \quad (6)$$

Step 4: During the lifetime of component i , calculate the conditional failure probability Pr_i of each inspection point, and we have:

$$Pr_{ki} = \frac{\int_{t_i}^{t_i+20} \frac{1}{0.1429\sqrt{2\pi}} e^{-\frac{(t_i-PT_{ki})^2}{2 \times 0.1429^2}} dt_i}{\int_{t_i}^{\infty} \frac{1}{0.1429\sqrt{2\pi}} e^{-\frac{(t_i-PT_{ki})^2}{2 \times 0.1429^2}} dt_i}$$

$$(k = 0, \dots, 100,000; i = 1, \dots, 5; t_i \geq 0)$$

where t_i is cumulated inspection time in one life circle for component i .

At each inspection point k , if Pr_{ki} ($i = 1, \dots, 5$) is greater than the given level-1 condition failure probability threshold Pr_1^* ($0 < Pr_1^* < 1$), preventively replace the component at time point k . If there is no preventive replacement during the lifetime of component i , perform failure replacement at the inspection point just behind FT_i . When there is a preventive/failure replacement occurring at time k , check other components, and if Pr_{kj} ($j = 1, \dots, 5$) is greater than the given level-2 failure probability threshold Pr_2^* , perform preventive replacement on component j simultaneously.

Step 5: When there is a preventive/ failure replacement taking place on component i , start a new life circle of component i by setting $t_i = 0$, and go back to Step 2. The iteration would not stop until k equals 100,000.

Step 6: Estimate cost rate. In this case, the fixed preventive replacement cost C_{po} is 3,000 and the variable preventive replacement cost C_p is 1,800. We have:

$$C_r = \frac{\text{Cost_total}}{\text{Time_total}} = \frac{\sum_{k=0}^{100,000} C_k}{100,000 \times 20} \text{ (\$/day)} \quad (k = 0, \dots, 100,000)$$

where

$$C_k = C_f \cdot \sum_{i=1}^N \Delta_{fki} + C_p \cdot \sum_{i=1}^N \Delta_{pki} + I(\Delta_{pki}) \cdot C_{po}$$

$$\Rightarrow C_k = 16,000 \cdot \sum_{i=1}^5 \Delta_{fki} + 1,8000 \cdot \sum_{i=1}^5 \Delta_{pki} + 3,000 \cdot I(\Delta_{pki})$$

where

$$\Delta_{pki} = \begin{cases} 1 & \text{Component } i \text{ was preventively replaced;} \\ 0 & \text{No preventive replacement} \end{cases}$$

$$\Delta_{fki} = \begin{cases} 1 & \text{Failure replacement on Component } i \\ 0 & \text{No failure replacement on component } i \end{cases}$$

If $\Delta_{pki} = 0$ & $\Delta_{fki} = 0$, the component i continues its normal operation.

Step 7: find the optimal total expected replacement cost. By setting different values for Pr_1 and Pr_2 , the corresponding total expected replacement cost can be evaluated. The minimal cost value can be found and the conditional failure probability threshold value Pr_1^* and Pr_2^* can be determined.

The expected cost as a function of Pr_1^* and Pr_2^* is plotted in Figure 2. The optimal failure probability threshold

values can be observed from this figure, where the lowest expected cost exists.

The minimal expected cost for multi-component occurs when $Pr_1^* = 0.100259$ and $Pr_2^* = 4.0973 \times 10^{-4}$, and the expected maintenance cost for this multi-component system containing 5 components is \$17.5651/day.

The comparative results are showed in Table 1. Comparing to the CBM policy for single units, the optimal cost is much lower when using multi-component CBM policy, with a cost saving of 27.21%.

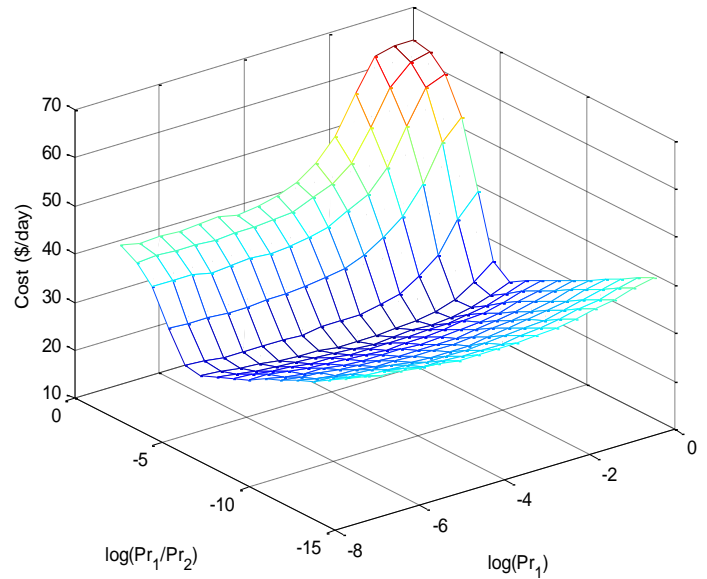


Figure 2. Cost versus two condition failure probability threshold values

Table 1. Comparison of cost between single unit and multi-component CBM policy

	Single Unit	Multi-component systems (5 components)
Cost (\$/day)	4.8264	17.5651
Cost for each component (\$/day)	4.8264	3.513
Cost savings (%)		27.21%

This comparative study demonstrates that the proposed multi-component CBM policy can achieve a lower total

expected replacement cost by taking advantage of economic dependency in multi-component systems.

5. CONCLUSION

In this paper, a CBM optimization method is proposed for multi-component systems, where economic dependency exists among the components subject to condition monitoring. Deterioration of a multi-component system is represented by a conditional failure probability value, which is calculated based on the predicted failure time distributions of components. The proposed CBM policy is defined by a two-level failure probability threshold. A method is developed to obtain the optimal threshold values in order to minimize the long-term maintenance cost. An example is used to demonstrate the proposed multi-component CBM method.

In future work, we can use discrete-event simulation tool, such as ARENA and FlexSim, to further verify the proposed method in this work and study more complex situations. We are also in the process of developing a numerical method for more accurate maintenance cost evaluation.

ACKNOWLEDGEMENT

This research is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC). We appreciate very much the help from OMDEC Inc. for providing the condition monitoring data used in the case study.

REFERENCES

- Banjevic D, Jardine AKS, Makis V., (2001) A control-limit policy and software for condition-based maintenance optimization. *INFOR*. 39: 32–50.
- Jardine AKS, Lin DM, Banjevic D., (2006) A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*. 20 (7): 1483-1510.
- Tian Z, Jin T, Wu B, Ding F., (2011a) Condition based maintenance optimization for wind power generation systems under continuous monitoring. *Renewable Energy*. 36: 1502-1509.
- Tian Z, Liao H. (2011b) Multi-component condition based maintenance optimization considering component level condition monitoring. *Reliability Engineering & System Safety*. 96(5): 581-589.
- Tian Z, Wong L, Safaei N., (2010) A neural network approach for remaining useful life prediction utilizing both failure and suspension histories. *Mechanical Systems and Signal Processing*. 24(5): 1542-1555.
- Stevens, B., (2006) *EXAKT reduces failures at Canadian Kraft Mill*, www.modec.com.
- Zhigang Tian** is currently an Assistant Professor at Concordia Institute for Information Systems Engineering at Concordia University, Montreal, Canada. He received his Ph.D. degree in 2007 in Mechanical Engineering at the University of Alberta, Canada; and his M.S. degree in 2003, and B.S. degree in 2000 both in Mechanical Engineering at Dalian University of Technology, China. His research interests focus on reliability analysis and optimization, prognostics, condition monitoring, and maintenance optimization. He is a member of IIE and INFORMS.
- Youmin Zhang** is an internationally recognized expert in the field of condition monitoring, fault diagnosis and fault-tolerant control with more than 15 years experience in the field. Dr. Zhang (with co-workers) published a first-ever research monograph (book) worldwide on “Fault Tolerant Control Systems” in 2003. He has published 4 books (two on the topic of fault tolerant control) and more than 200 referred journal and conference papers. He has been awarded an NSERC Strategic Project Grant (SPG) and an NSERC Discovery Project Grant and after he joined Concordia University in Dec. 2006.
- Jialin Cheng** obtained her master’s degree from Concordia Institute for Information Systems Engineering at Concordia University, Canada in 2010. Her research interests is in the field of condition based maintenance, multi-component systems and reliability modeling.

Cost Comparison of Maintenance Policies

Le Minh Duc¹, Tan Cher Ming² (Senior Member, IEEE)

Division of Circuits and Systems
School of Electrical & Electronics Engineering
Nanyang Technological University, Singapore

¹ lemi0006@ntu.edu.sg

² ecmtan@ntu.edu.sg

ABSTRACT

Maintenance is crucial to all repairable engineering systems as they will degrade and fail. The cost of maintenance for a manufacturing plant can occupy up to 30% of the total operating cost. If maintenance is not scheduled properly, unexpected equipment failure can induce significant cost due to reduced productivity and sub-standard products produced, both of which may result in customer penalty.

Various maintenance policies have been proposed in the past. Among the various policies, age-dependent and periodic maintenances are the common policies employed in industries. Recently, predictive maintenance or condition based maintenance policies are also proposed owing to the advancement in the sensor technology. In this work, we compare the age-dependent and periodic maintenance policies as well as the predictive maintenance policies from the perspective of cost using Markov multi-state maintenance modeling and Monte Carlo simulation. To be realistic, imperfect maintenance is included, and both the sequential and continuous inspections are considered and compared.

1. INTRODUCTION

All industrial systems suffer from deterioration due to usage and age, which may leads to system failures. To some industry, system failures cause serious consequences, especially in industries such as transportation, construction, or energy sectors. These deterioration and failure can be controlled through a proper maintenance plan.

The cost of maintenance as a fraction of the total operating budget varies across industry sectors. In the mining industry, it can be as high as 50% and in transportation industry it varies in the range of 20-30 % (Murthy, Atrens, & Eccleston, 2002), which accounts only for the actions to

keep the system in operating state. The consequential cost of failure could be much higher. Hence, it is vital to have a good maintenance policy so as to reduce the possibility of failure to the least while preserves a low maintenance cost.

Maintenance problems have been extensively investigated in the literature, and a number of maintenance policies have been proposed. These policies span from the most basic one as corrective maintenance (CM) to more advanced policy as preventive maintenance (PM). CM is carried out only when a system fails. PM is performed when the system is still operating, in attempt to preserve the system in its good condition, and the most popular PM policy is age-dependent PM policy (Barlow, Proschan, & Hunter, 1996). Under this maintenance policy, the system is preventively replaced at its age of T or at failure, whichever occurs first, where T is a constant. The extension of this maintenance policy includes considering the effect of imperfect maintenance or minimal repair at failure (Kijima, 1989; Nakagawa, 1984; SHEU, KUO, & NAGAGAWA, 1993). Another common maintenance policy is periodic PM (Barlow, et al., 1996). Under this maintenance policy, a system is preventively maintained at fixed time interval T regardless of the failure history of the system and at intervening failures. This policy is often applied to a group of units where the failure's history of one unit is often neglected. There are several modifications of this periodic PM policy. In (Nakagawa, 1986), minimal repair is performed at failure and the system is replaced at planned time kT if the number of failure exceeds n . Age-dependent PM and periodic PM can be combined as in (Berg & Epstein, 1976), in which the system is periodically replaced only if its age exceeds T_0 . Although being common and popular, age-dependent PM and periodic PM do not account for the actual condition of the system, thus these policies may result in unnecessary replacement of good units and cost expenditure.

Recently, condition-based maintenance (CBM), which is a subset of Predictive Maintenance (PdM), is proposed in order to improve the cost effectiveness of existing PM

Minh Duc Le et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

policies. CBM is to make maintenance decisions based on the actual system's health condition (Lu, Tu, & Lu, 2007; Ming Tan & Raghavan, 2008). CBM is often applied to system with degradable performance, which can be represented by different states. A CBM policy assigns a maintenance action to each system state. By its definition, CBM must be carried out based on the observation of the system's health, which is obtained using either sequential or continuous inspection. With the advancement of sensor technology, the system's health condition can be observed continuously. In (Moustafa, Maksoud, & Sadek, 2004), a CBM policy is developed for continuous inspection, and two maintenance options are considered, namely replacement and minimal repair. Although continuous inspection is commonly used in detecting system's degradation, it usually swarms with unnecessary and excessive data. Also, the inspection process can be costly, especially with complex systems which requires huge number of monitoring devices. Hence, there are several works on maintenance policies in which the system is inspected only at specific time (sequential inspection) and replaced with a new identical one only when the degradation reaches a predefined threshold (Grall, Dieulle, Bérenguer, & Roussignol, 2002; Lam & Yeh, 1994; Ohnishi, Kawai, & Mine, 1986). In the formulation of such policies, they considered the cost of operation in different states of degradation, cost of inspection, and maintenance. Tomasevicz (Tomasevicz & Asgarpoor, 2009) extended their works by considering the effect of imperfect maintenance and by introducing maintenance states, from which the system can be recovered to a better operating state. Their comprehensive cost analysis showed that an optimal choice of inspection date and replacement threshold can improve the cost effectiveness of the maintenance policy.

It is widely assumed that the imperfect maintenance restores a system to a state between as good as new (replacement) and as bad as old (minimal repair). The two extreme cases are investigated thoroughly in early works. In general, these assumptions are not true in many applications. In practice, imperfection can arise due to the maintenance engineering skills, quality of the replaced parts and complexity of the degraded systems. Several theoretical models are developed that taking into account the imperfect maintenance (Nakagawa & Yasui, 1987; Pham & Wang, 1996). They can be broadly classified into four classes, namely the *probabilistic approach* (Nakagawa & Yasui, 1987), *improvement factor* (Chan & Shaw, 1993; Malik, 1979), *virtual age* (Kijima, 1989; Kijima, Morimura, & Suzuki, 1988), and the final class which is based on the *cumulative system degradation* model (Martorell, Sanchez, & Serradell, 1999). For detailed discussion on the various maintenance models, one can refer to (Brown & Proschan, 1983; Levitin & Lisnianski, 2000; Wang & Pham, 1996).

In this work, we will compare the age-dependent and periodic maintenance policies as well as the predictive maintenance policies from the perspective of cost using Markov multi-state system modeling and Monte Carlo simulation. To be realistic, imperfect maintenance is included, and both the sequential and continuous inspections are considered and compared. The novelty of this work lies in the introduction of imperfect maintenance in the optimization of CBM policy for Markov multi-state system. A clear comparison between age-dependent PM, periodic PM and condition-based maintenance under different cost-related conditions will be shown, and the advantages and disadvantages of each maintenance policy will be discussed.

2. MAINTENANCE POLICIES

2.1. System Description

The system under study is a multi-state system, and each state represents a system's health condition. These states can be defined by either a degradation index such as vibration's intensity, temperature, etc, or simply the system's performance. The system is assumed to be in a finite number of states $1, 2, 3 \dots N$ where state 1 is the as-good-as-new state and state N is the completely failed state. The states are in ascending deteriorating order.

The degradation process is represented by the transition from one state to another state. In normal operation, the failures of a complex system have been shown (Drenick, 1960) to follow the exponential distribution despite the fact that the individual components in the system may follow different distributions. Hence, the system's deterioration process can be modeled as a continuous-time Markov process. From state i , ($1 \leq i \leq N - 1$) the system can only transit to the more degraded state j , ($i \leq j \leq N$) with a transition rate of λ_{ij} . In this work, for the sake of simplicity, we assume that the transition rates are constant for a given i and j . From the values of λ_{ij} , the probability $P_{ij}(t)$ that the system is at state j after a time t given that the system is originally at state i can be calculated. In actual cases, the transition rate can be changed after the system is maintained.

The state of the system is not known unless it is inspected. In the case of sequential inspection, the cost for each successive inspection is fixed at C^{SI} . During the time of inspection, the system state is unchanged. In the case of continuous inspection, since the system is continuously monitored, the state can be instantly detected and the cost is represented as a cost per unit time c^{CI} . The system's failure (system at state N) is detected without inspection and not recoverable by maintenance. The system upon replacement is recovered to the initial state 1 with a cost of C^R . However, the failure also results in a secondary consequential damage such as unplanned delay in production, lost of physical

assets etc, which is represented by a cost of C^f . The value of C^f depends on the nature of the failures.

Upon maintenance, the system's state is improved to a better state $j, i \geq j \geq 1$, with probability P_{ij}^M . The probability that the system is recovered to as-good-as-new state is getting smaller as the system state is approaching N, and the maintenance cost and time varies with different states. C_{ij}^M is denoted as the cost of maintenance to repair the system from state i to a less degraded state j . The maintenance cost includes the cost due to system unavailability.

The illustration of all the different quantities is as shown in Figure 1.

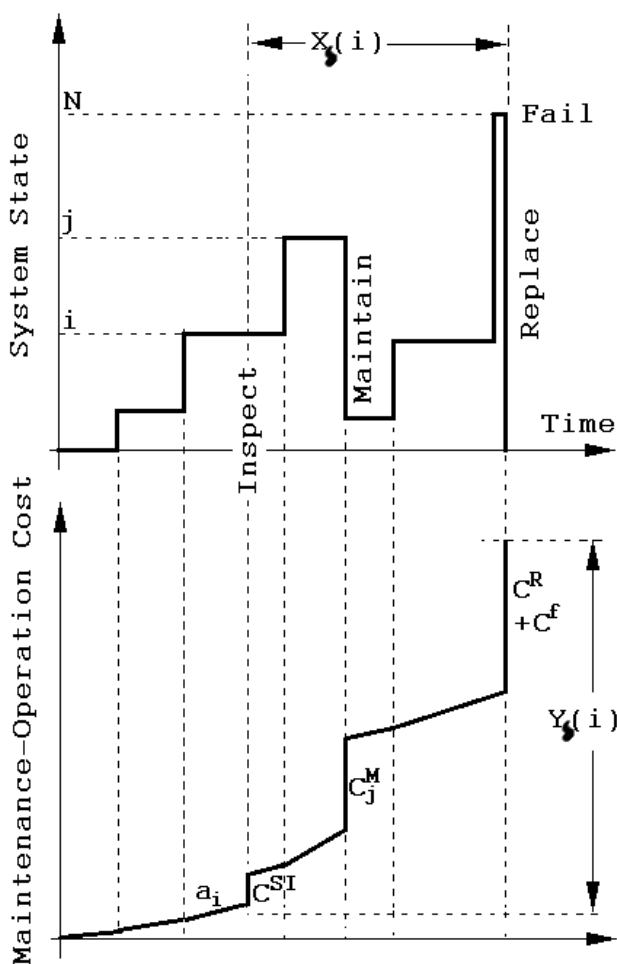


Figure 1. Schematic view of the system state degradation and its maintenance-operation cost

To proceed to the determination of the optimal maintenance policies, let us define the following terms:

δ : A policy which determines the action at each state, either replacement, maintenance or continue the inspection.

$D(i)$: Decision at state i . They can be either to inspect the system after time interval $t_i (I(t_i))$, maintain (M), replace (R) or keep monitoring in the case of employing continuous inspection (C).

$X_\delta(i)$: Mean operating time from the moment the system is detected to be at state i to the time where the system is replaced (at state N) for a given policy δ . Hence, $X_\delta(1)$ is the mean time from a new/newly replaced system till it is replaced.

$Y_\delta(i)$: Mean cost from the moment the system is detected to be at state i to the time where the system is replaced (at state N), for a given policy δ . Hence, $Y_\delta(1)$ is the mean cost from a new/newly replaced system till it is replaced.

$F_i(t)$: Probability that the system will fail in the interval $(0, t)$ given that the system is at state i .

a_i : operating cost at state i . The cost of operation is increasing with the degradation in order to accounts for the loss in profit due to the degradation in the system's performance.

The mean operating cost, given that the system initially at state i , after a time t is (Ohnishi, et al., 1986):

$$A_i(t) \triangleq \sum_{j=i}^N \int_0^t P_{ij}(u) a_j du \quad (1)$$

$P_{ij}(t)$: Probability that the system will is at state j after a time t given that the system is at state i .

2.2. Maintenance Policies

In this work, we compare four different maintenance policies for a multi-state system, namely age-dependent PM, periodic PM, sequential and continuous inspection CBM. The optimal maintenance policy refers to minimum overall operation cost rate $g^* \equiv \min_{\delta} Y_\delta(1)/X_\delta(1)$. Let us now look at the formulation of the optimization for each maintenance policy.

a. Age-dependent PM:

In this study, we only consider the most basic Age-dependent PM, which does not utilize maintenance. The system is preventively replaced at its age of T_a or at failure, whichever occurs first. T_a is chosen so that the cost rate is minimized.

The mean cost and operating time until system replacement can be expressed as:

$$Y_A = A_1(T_a) + C^R + F_1(T_a)C^f \quad (2)$$

$$X_A = \int_0^{T_a} \bar{F}_1(u)du \quad (3)$$

In (2), the terms $A_1(T_a)$, C^R and $F_1(T_a)C^f$ represent the mean operation cost in the interval $(0, T_a)$, replacement cost and mean failure-induced cost respectively. In (3), $\int_0^{T_a} \bar{F}_1(u)du$ is the expected operating time in the interval $(0, T_a)$. This can be derived by considering two possibility of the system's operation, i.e. the system can either work up to t_a with the expected operating time of $t_1 = t_a \bar{F}_1(t_a)$, or the system fails at u within the interval $(0, t_a)$ with the expected operating time of $t_2 = \int_0^{t_a} u dF(u)$. We thus have $X_A = t_1 + t_2$.

b. Periodic PM:

The system is preventively replaced at fixed time interval T_b or at intervening failures regardless of the failure history of the system. Here T_b is a constant, and it is chosen so that the cost rate is minimized.

The mean cost until system replacement can be expressed as (4).

$$Y(T_b) = C^R + (C^R + C^f)M(T_b) + C_{ope}(T_b) \quad (4)$$

In (4), $M(t)$ is the mean number of failure and is given in (5) (Barlow, et al., 1996).

$$M(t) = \int_0^t (1 + M(t-x))dF_1(x) = \sum_{n=1}^{\infty} F_1^n(t) \quad (5)$$

$$F_1^{n+1}(t) = \int_0^t F_1^n(t-x)dF_1(x), F_1^1(t) = F_1(t)$$

$C_{ope}(t)$ is the mean operation cost in the duration $(0, t)$, and they are given in (6). The term $C_k(t)$ represents the mean operation cost given that exactly k failures occur and can be calculated recursively as shown in (7). In (7), x is the time of the first failure occurs in the interval $(0, t)$. Thus, the $C_k(t)$ can be computed by integrating the summation of the operation cost before and after x for all x in $(0, t)$.

$$C_{ope}(t) = \sum_{k=1}^{\infty} C_k(t) \quad (6)$$

$$C_n(t) = \int_0^t (C_0(x) + C_{n-1}(t-x))dF_1(x) \quad (7)$$

$$C_0(t) = A_1(t)$$

The mean operating time until system replacement can be expressed as (8).

$$X(T_b) = T_b \quad (8)$$

c. Sequential Inspection CBM (SI-CBM):

The system is inspected at a planned time. The decision depends on the indicated system state i , which is either preventively replaced $D(i) = R$, maintained $D(i) = M$, or to leave the system operating until the next planned inspection time $D(i) = I(t_i)$. Maintenance is considered to be imperfect. If $i = N$, the system fails and needs to be replaced. In that case, we have $X_\delta(N) = T^R$, $Y_\delta(N) = C^R + C^f$. The decision $D(i)$ at each state is chosen so that the cost rate is minimized.

1. If $D(i) = I(t_i)$

Under this decision, the system is left to degrade until the next inspection after an interval t_i . If the system fails at $u < t_i$, it is replaced. If the system passes the time interval t_i without failure, the time to replacement will be t_i plus the mean time to replacement of the arrived state j . Once the planned inspection time t_i is reached, the system is inspected. The mean cost and operating time until renewal under the decision $D(i) = I(t_i)$ can be expressed as:

$$Y_\delta(i) = A_i(t_i) + C^{SI} \bar{F}_i(t_i) + \sum_{j=i}^N P_{ij}(t_i) Y_\delta(j) \quad (9)$$

$$X_\delta(i) = \int_0^{t_i} \bar{F}_i(u)du + \sum_{j=i}^N P_{ij}(t_i) X_\delta(j) \quad (10)$$

In (9), $A_i(t_i)$ is the mean operating time in the interval $(0, t_i)$, $C^{SI} \bar{F}_i(t_i)$ is the mean inspection cost and $\sum_{j=i}^N P_{ij}(t_i) Y_\delta(j)$ is the expected cost until replacement given that the system is in the degraded state j .

In (10), $\int_0^{t_i} \bar{F}_i(u)du$ is the expected time to replacement in the interval $(0, t_i)$ and $\sum_{j=i}^N P_{ij}(t_i) X_\delta(j)$ is the expected operating time given that the system is in the degraded state j .

2. If $D(i) = M$

The system is maintained with a maintenance cost C_{ij}^M , and the system is thus improved from the current state i to a less degraded state j with an improvement probability of P_{ij}^M . The mean cost and operating time until replacement can be expressed as

$$Y_\delta(i) = \sum_{j=1}^N P_{ij}^M (C_{ij}^M + Y_\delta(j)) \quad (11)$$

$$X_\delta(i) = \sum_{j=1}^N P_{ij}^M X_\delta(j) \quad (12)$$

3. If $D(i) = R, i \neq N$

Except the failure case $i = N$, the system can be preventively replaced. Thus, the mean operating time and cost until replacement can be expressed as

$$Y_{\delta}(i) = C^R \quad (13)$$

$$X_{\delta}(i) = 0 \quad (14)$$

Overall, we have

$$Y_{\delta}(i) = \begin{cases} A_i(t_i) + C^{SI} \bar{F}_i(t_i) + \sum_{j=i}^N P_{ij}(t_i) Y_{\delta}(j), & \text{if } D(i) = I(t) \\ \sum_{j=1}^N P_{ij}^M (C_{ij}^M + Y_{\delta}(j)), & \text{if } D(i) = M \\ C^R, & \text{if } D(i) = R \end{cases} \quad (15)$$

$$X_{\delta}(i) = \begin{cases} \int_0^{t_i} \bar{F}_i(u) du + \sum_{j=i}^N P_{ij}(t_i) X_{\delta}(j), & \text{if } D(i) = I(t_i) \\ \sum_{j=1}^N P_{ij}^M X_{\delta}(j), & \text{if } D(i) = M \\ 0, & \text{if } D(i) = R \end{cases} \quad (16)$$

d. Continuous Inspection CBM (CI-CBM):

The system is inspected continuously. The decision depends on the indicated system state, which is either preventively replaced $D(i) = R$, maintained $D(i) = M$, or to leave the system operating while keep monitoring the system's condition $D(i) = C$. Maintenance is considered to be imperfect. The decision $D(i)$ at each state is chosen so that the cost rate is minimized.

For the first two decisions, the analysis is the same with SI-CBM case. Under the decision of continuous inspection, the system is operating at state i until it changes its state to a more degraded state j . The mean cost and operating time until renewal under the decision $D(i) = C$ can be expressed as:

$$Y_{\delta}(i) = (c^{CI} + a_i) \int_0^{\infty} P_{ii}(u) du + \sum_{i=1}^N \frac{\lambda_{ij}}{\sum_{k=1}^N \lambda_{ik}} Y_{\delta}(j) \quad (17)$$

$$X_{\delta}(i) = \int_0^{\infty} P_{ii}(u) du + \sum_{i=1}^N \frac{\lambda_{ij}}{\sum_{k=1}^N \lambda_{ik}} X_{\delta}(j) \quad (18)$$

In (17) and (18), $\int_0^{\infty} P_{ii}(u) du$ is the mean time the system operate at state i , $\frac{\lambda_{ij}}{\sum_{k=1}^N \lambda_{ik}}$ is the probability that the system transit from state i to state j at any instant given that the system has to change its state. Thus, $(c^{CI} + a_i) \int_0^{\infty} P_{ii}(u) du$ is the mean operation plus inspection cost when the system is at state i ,

$\sum_{i=1}^N \frac{\lambda_{ij}}{\sum_{k=1}^N \lambda_{ik}} Y_{\delta}(j)$ and $\sum_{i=1}^N \frac{\lambda_{ij}}{\sum_{k=1}^N \lambda_{ik}} X_{\delta}(j)$ are the mean cost and operating time until replacement averaging on the degraded state j .

Overall, we have

$$Y_{\delta}(i) = \begin{cases} (c^{CI} + a_i) \int_0^{\infty} P_{ii}(u) du + \sum_{i=1}^N \frac{\lambda_{ij}}{\sum_{k=1}^N \lambda_{ik}} Y_{\delta}(j), & \text{if } D(i) = C \\ \sum_{j=1}^N P_{ij}^M (C_{ij}^M + Y_{\delta}(j)), & \text{if } D(i) = M \\ C^R, & \text{if } D(i) = R \end{cases} \quad (19)$$

$$X_{\delta}(i) = \begin{cases} \int_0^{\infty} P_{ii}(u) du + \sum_{i=1}^N \frac{\lambda_{ij}}{\sum_{k=1}^N \lambda_{ik}} X_{\delta}(j), & \text{if } D(i) = C \\ \sum_{j=1}^N P_{ij}^M X_{\delta}(j), & \text{if } D(i) = M \\ 0, & \text{if } D(i) = R \end{cases} \quad (20)$$

3. EXAMPLE THROUGH HYPOTHETIC SYSTEM

In this section, a hypothetical system is studied to illustrate the impact of different policies on the system's total cost and number of maintenance. The system consists of twenty one states (1-21), which represents the system degradation levels in ascending order. State 1 is the state of no degradation (best performance) and state 21 is the state of total failure (worst performance). For simplicity, we only consider degradation in the sense that at any moment the system only degrades to the next degraded state (with a fixed degradation rate $\lambda_{i,i+1}$) or experienced a shock so that it fails immediately (with a failure rate λ_{iN}). From the assumption that the state transition is a continuous time Markov process, we have the set of Kolmogorov forward equations as shown in Eqn (21):

$$\frac{dP_{ij}(t)}{dt} = \sum_{k=i}^{j-1} \lambda_{kj} P_{ik}(t) - \sum_{k=j+1}^N \lambda_{jk} P_{ij}(t) \quad (21)$$

Here the first term on the right of the equation refers to the degradation process from state i , and the second term on the right refers to the further degradation process from state j . Eqn (21) can be re-written as follows:

$$\frac{d}{dt} \begin{bmatrix} P_{ii}(t) \\ P_{i,i+1}(t) \\ \dots \\ P_{iN}(t) \end{bmatrix} = Q_i \begin{bmatrix} P_{ii}(t) \\ P_{i,i+1}(t) \\ \dots \\ P_{iN}(t) \end{bmatrix} \quad (22)$$

where $Q_i = \begin{bmatrix} -\sum_{k=i+1}^N \lambda_{ik} & 0 & \dots & 0 \\ \lambda_{i,i+1} & -\sum_{k=j+1}^N \lambda_{jk} & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \end{bmatrix}$

With Eqn (22), $P_{ij}(t)$ can be calculated numerically given that the initial probability $t = 0$ is $[P_{i1}, P_{i,i+1}, \dots, P_{iN}]^T = [1, 0, \dots, 0]^T$. The system state is then randomly generated in Monte Carlo simulation based on the probability $P_{ij}(t)$. A numerical example of Markov process with detail derivation can be found in (Ming Tan & Raghavan, 2008).

Due to the loss caused by degradation, namely lower productivity and higher recourse consumption, the cost of operation is increasing with the degradation levels. The degradation and failure rate, operation and maintenance cost for different states are hypothetically assumed and given in TABLE.I. In this study, we want to investigate the impact of failure-induced cost C^f on the optimal maintenance policies. When failure occurs, it will induce a further cost such as production delay, human and asset lost, etc. The total cost of the system maintenance at failure is the summation of system's replacement cost and the failure-induced cost.

For illustration purpose on the computation of the conditional probability of the post-maintained j state given pre-maintained state i , we further considered a special system in this case study. The system is assumed to consist of n identical sub-systems in parallel, in which a system state i represents the condition that $(i - 1)$ subsystems are operating. With this special system, an analytical form of the maintenance probability P_{ij}^M can be derived.

The imperfect maintenance is characterized using the maintenance quality represented by a parameter p_m , which is the probability that a subsystem can be recovered to as new by maintenance actions. The value of p_m of a subsystem can be estimated using the method of determining the restoration factor RF described in (Ming Tan & Raghavan, 2008). Since our system consists of n identical sub-systems in parallel and all the failed sub-systems has equal probability p_m to be recovered at each maintenance, the probability of post-maintained state j is the probability that $i - j$ sub-systems are recovered and thus one can use the binomial distribution to compute the P_{ij}^M as follows.

$$P_{ij}^M = \binom{i-1}{j-1} (1-p_m)^{j-1} p_m^{i-j} \tag{12}$$

It follows that the expected post-maintained state and its variance are both linearly increasing with the pre-maintained state i , i.e. $E(j) = p_m + (1 - p_m)i$ and $var(j) = (i - 1)p_m(1 - p_m)$. These indices indicate that as the system is more degraded, it is more difficult to maintain the system to the initial condition and the consistence of the maintenance quality decreases. For a general system, the optimization algorithm is still applicable as long as the maintenance probability P_{ij}^M is given. The

investigation for such a general system is beyond the scope of the present work.

SS	DR	FR	OC	MC
1	0.4966	0.0082	2.7183	31.4942
2	0.5016	0.0091	2.8577	31.6111
3	0.5066	0.0099	3.0042	31.7371
4	0.5117	0.0108	3.1582	31.8736
5	0.5169	0.0118	3.3201	32.0216
6	0.5221	0.0129	3.4903	32.1826
7	0.5273	0.0141	3.6693	32.3587
8	0.5326	0.0155	3.8574	32.5531
9	0.5379	0.0169	4.0552	32.7705
10	0.5434	0.0185	4.2631	33.0183
11	0.5488	0.0202	4.4817	33.3091
12	0.5543	0.0221	4.4715	33.6631
13	0.5599	0.0242	4.9531	34.1154
14	0.5655	0.0265	5.207	34.7253
15	0.5712	0.0291	5.4739	35.59511
16	0.5769	0.0317	5.4746	36.9004
17	0.5827	0.0347	6.0496	38.945
18	0.5886	0.0381	6.3598	42.2541
19	0.5945	0.0416	6.6859	47.7363
20	0.6005	0.0455	7.0287	56.9556
21	0.6065	0.0498	0	72.6683

SS : System State
 DR : Degradation Rate (per month)
 FR : Failure Rate (per month)
 OC : Operation Cost Rate (\$.000/month)
 MC : Maintenance Cost (\$.000)

Table 1. System State's Maintenance Cost & Degradation Rate

4. MONTE CARLO SIMULATION RESULT & DISCUSSION

Using different values of the failure-induced cost C^f and maintenance quality p_m , the optimal maintenance plans are derived for each of the above-mentioned maintenance policy.

Monte Carlo simulation is run for each derived maintenance policy so as to investigate the impact of failure-induced cost and maintenance quality on the system's total operation-maintenance cost, number of maintenance and number of failure. For each value of C^f and p_m , the simulation is repeated for 500 random samples. The total system runtime is assumed to be 120 months.

4.1. Impact of Failure-Induced Cost

The failure-induced cost is assumed to range from 20 to 1000 (\$.000). For SI-CBM and CI-CBM, the maintenance quality is kept at $p_m = 0.8$.

Figure 2 shows the changes of mean value of the total maintenance-operation cost of different maintenance policies vs. failure-induced cost. The total maintenance-operation cost is the summation of all operation, maintenance and replacement cost:

$$\text{replacement cost} = \# \text{replacement} \times C^R$$

$$\text{maintain cost} = \sum_{i=1}^{N-1} \#(\text{maintain at state } i) \times C_i^M$$

$$\text{operation cost} = \sum_{i=1}^{N-1} (\text{duration at state } i) \times a_i$$

It appears that the mean value of the total cost has a linear relation with respect to the failure-induced cost.

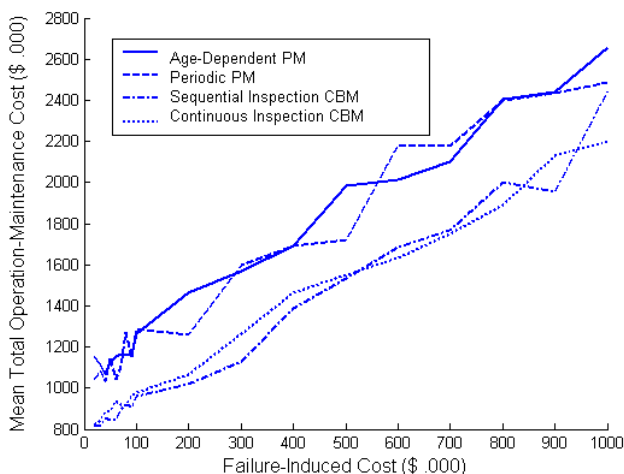


Figure 2. Mean value of total Operation-Maintenance Cost of various maintenance policies vs. failure-induced cost

One can also see that CBM policies have a clear advantage over traditional PM policies in term of cost reduction. At low failure-induced cost, utilizing CBM policies can save up to $(1100 - 800)/1100 = 27\%$ of the total cost under PM policies.

Figure 3 shows the normalized standard deviation (NSTD) curve of the total cost under different maintenance policies. This NSTD is the standard deviation of the total cost from 500 samples of Monte Carlo run divided by its mean value $\bar{\sigma} = \sigma/\mu$. The standard deviation appears to have a linear relation to the failure-induced cost. At low C^f , the NSTD values under CBM policies are approximately equal to the NSTD under PM policies as 10%. However, as C^f

increase, the NSTD under CBM policies increases dramatically up to 70% for SI-CBM and 62% for CI-CBM while it is less than 50% for PM policies cases. This is due to the increase number of imperfect maintenance under CBM as the failure-induced cost increases. As a result of rising failure-induced cost, the optimal CBM policies have to increase the number of maintenance in order to reduce the number of failure.

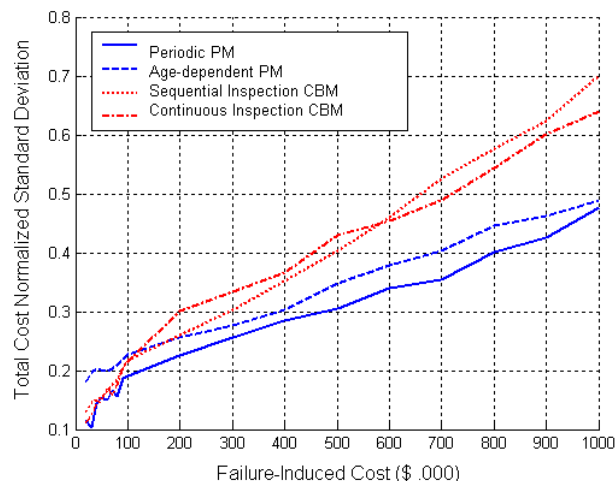


Figure 3. Cost normalized standard deviation under different maintenance policies vs. Failure-induced cost

Figure 4 shows that the mean number of failure decreases exponentially as the failure-induced cost rise. It is the effect of optimal maintenance policy, which tends to reduce the number of failure as the failure-induced cost increase. However, the mean number of failure only decreases to a certain value for each maintenance policy. This lower bound value is lower for CBM policies than PM policies by 25%, which proves that CBM is more advantageous than PM in preventing system failure.

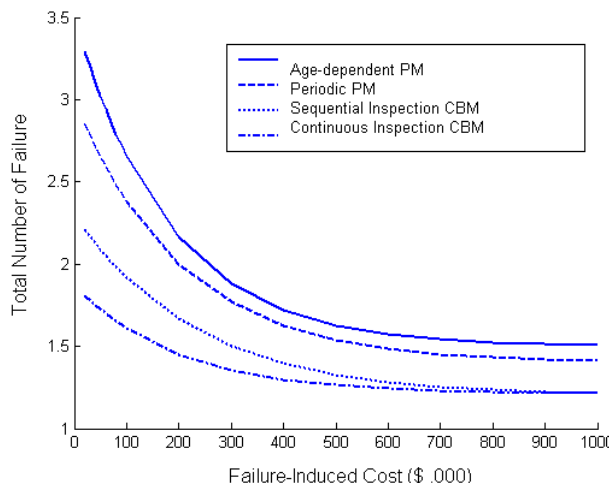


Figure 4. Mean number of failure vs. Failure-induced cost

In summary, one can see that SI-CBM and CI-CBM have a clear advantage in term of cost and failure reduction over Age-dependent and periodic PM. However, imperfect maintenance causes the total cost of CBM to vary significantly, especially at high failure-induced cost due to higher number of maintenance needed for the optimal policy. This large variation in cost may render the financial budgeting for using CBM difficult.

4.2. Impact of maintenance quality

In this case, the failure-induced cost is kept at 100 (\$.000) while the maintenance quality is ranging from $p_m = 1$ (perfect maintenance) to $p_m = 0.6$.

SI-CBM and CI-CBM policies are investigated to study the impact of maintenance quality. Figure 5 shows the total cost of SI-CBM and CI-CBM under two schemes: optimal policies and the policies assuming perfect maintenance ($p_m = 1$). Under the CBM policies that assume perfect maintenance while the maintenance is actually imperfect, the total cost increase dramatically as p_m decrease. As p_m close to 1, the difference between optimal CBM and the one assuming perfect maintenance is negligible as expected. However, the difference increase significantly when $p_m = 0.6$ and beyond. At $p_m = 0.6$, the optimal CI-CBM can save up to $((1500 - 1000))/1500 = 33\%$ of the total cost comparing to the policy assuming perfect maintenance. The cost under CBM policies assuming perfect maintenance eventually rise up to infinity as p_m approaches zero since at $p_m = 0$, maintenance take no effect. The total cost under both optimal CI-CBM and SI-CBM also tend to saturate as p_m decreases. This is due to the fact that maintenance is gradually ruled out due to its poor quality (referring to figure 6). Thus the saturated value is corresponding to the CBM policy that does not utilize maintenance.

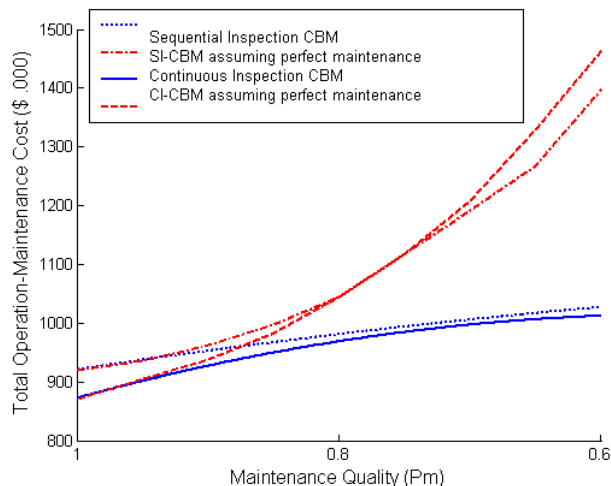


Figure 5. Total operation-Maintenance Cost vs. Maintenance Quality with $C^f = 100$

Figure 6 shows the mean number of maintenance changes with respect to the maintenance quality. The plots under SI-CBM and CI-CBM follow the same trend. When the maintenance quality gets worse, the mean number of maintenance increases to cover for the imperfection. However, at low values of p_m , the number of maintenance drops dramatically to zero as maintenance is too ineffective, and hence our optimization process for maintenance will try to reduce the number of maintenances.

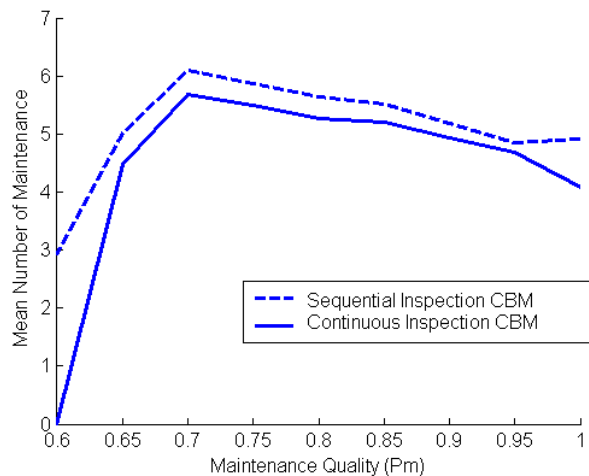


Figure 6. Mean number of maintenance vs. Maintenance Quality, $C^f = 100$

From this study, we can see that there is a threshold for maintenance quality, under which, the maintenance is no longer effective and should be changed to preventive replacement. The quality of maintenance must be carefully taken into account when making a maintenance policy since a poor maintenance quality can lead to a large portion in overall cost.

5. CONCLUSION

In this work, we study different maintenance policies for a multistate system. Four maintenance policies are investigated, namely age-dependent and periodic preventive maintenance, sequential and continuous inspection condition-based maintenance (CBM). The system has a state dependent degradation rate during its operation, and it also suffers shock failure which makes it fails immediately with a state dependent failure rate. The failure is assumed to induce further cost and maintenance is assumed to be imperfect. The maintenance policies are optimized correspondingly.

Monte Carlo simulation shows that CBM is more advantageous in term of cost and failure reduction than Age-dependent and periodic PM. On the other hand, the maintenance cost under CBM is less consistent than under PM, which renders the budgeting difficult. We also illustrate the important of maintenance quality since a poor maintenance quality can lead to a large waste in maintenance cost. It can be proven that the maintenance quality must be higher than a threshold to be worth carrying out.

One issue for CBM to be effectively applied is to have accurate inspection. Besides, CBM also need a dynamics logistic supply of spare parts, which may further cause some time delay between inspection and maintenance. Hence, for the future work, we will consider the inspection quality and time delay due to supply limit in our model.

In our paper, the Monte-Carlo simulation is run for two parameters, but varying only one parameter at a time. A matrix of multi-variables will be studied for a future work to better understand the trade-offs between different quantities. These will permit understanding strategies based on which one can practice non-CBM methods on some components versus CBM on others.

REFERENCES

- Barlow, R.E., Proschan, F., & Hunter, L.C. (1996). *Mathematical theory of reliability* (Vol. 17): Society for Industrial Mathematics.
- Berg, M., & Epstein, B. (1976). A modified block replacement policy. *Naval Research Logistics Quarterly*, 23(1), 15-24.
- Brown, M., & Proschan, F. (1983). Imperfect repair. *Journal of Applied Probability*, 20(4), 851-859.
- Chan, J.K., & Shaw, L. (1993). Modeling repairable systems with failure rates that depend on age and maintenance. *Reliability, IEEE Transactions on*, 42(4), 566-571.
- Drenick, RF. (1960). The failure law of complex equipment. *Journal of the Society for Industrial and Applied Mathematics*, 8(4), 680-690.
- Grall, A., Dieulle, L., Bérenguer, C., & Roussignol, M. (2002). Continuous-time predictive-maintenance scheduling for a deteriorating system. *Reliability, IEEE Transactions on*, 51(2), 141-150.
- Kijima, M. (1989). Some results for repairable systems with general repair. *Journal of Applied Probability*, 26(1), 89-102.
- Kijima, M., Morimura, H., & Suzuki, Y. (1988). Periodical replacement problem without assuming minimal repair. *European Journal of Operational Research*, 37(2), 194-203.
- Lam, CT, & Yeh, RH. (1994). Optimal maintenance-policies for deteriorating systems under various maintenance strategies. *Reliability, IEEE Transactions on*, 43(3), 423-430.
- Levitin, G., & Lisnianski, A. (2000). Optimization of imperfect preventive maintenance for multi-state systems. *Reliability Engineering & System Safety*, 67(2), 193-203.
- Lu, S., Tu, Y.C., & Lu, H. (2007). Predictive condition based maintenance for continuously deteriorating systems. *Quality and Reliability Engineering International*, 23(1), 71-81.
- Malik, MAK. (1979). Reliable preventive maintenance policy. *AIEE transactions*, 11(3), 221-228.
- Martorell, S., Sanchez, A., & Serradell, V. (1999). Age-dependent reliability model considering effects of maintenance and working conditions. *Reliability Engineering & System Safety*, 64(1), 19-31.
- Ming Tan, C., & Raghavan, N. (2008). A framework to practical predictive maintenance modeling for multi-state systems. *Reliability Engineering & System Safety*, 93(8), 1138-1150.
- Moustafa, MS, Maksoud, EY, & Sadek, S. (2004). Optimal major and minimal maintenance policies for deteriorating systems. *Reliability Engineering & System Safety*, 83(3), 363-368.
- Murthy, DNP, Atrens, A., & Eccleston, JA. (2002). Strategic maintenance management. *Journal of Quality in Maintenance Engineering*, 8(4), 287-305.
- Nakagawa, T. (1984). Optimal policy of continuous and discrete replacement with minimal repair at failure. *Naval Research Logistics Quarterly*, 31(4), 543-550.
- Nakagawa, T. (1986). Periodic and sequential preventive maintenance policies. *Journal of Applied Probability*, 23(2), 536-542.
- Nakagawa, T., & Yasui, K. (1987). Optimum policies for a system with imperfect maintenance. *Reliability, IEEE Transactions on*, 36(5), 631-633.
- Ohnishi, M., Kawai, H., & Mine, H. (1986). An optimal inspection and replacement policy for a deteriorating system. *Journal of Applied Probability*, 23(4), 973-988.
- Pham, H., & Wang, H. (1996). Imperfect maintenance. *European Journal of Operational Research*, 94(3), 425-438.
- SHEU, S.H., KUO, C.M., & NAGAGAWA, T. (1993). Extended optimal age replacement policy with minimal repair. *RAIRO. Recherche opérationnelle*, 27(3), 337-351.
- Tomasevicz, C.L., & Asgarpoor, S. (2009). Optimum maintenance policy using semi-Markov decision processes. *Electric Power Systems Research*, 79(9), 1286-1291.
- Wang, H., & Pham, H. (1996). A quasi renewal process and its applications in imperfect maintenance. *International journal of systems science*, 27(10), 1055-1062.

MINH DUC LE received the B.Eng degree in 2008 from School of Electrical & Electronics Engineering (EEE), Nanyang Technological University (NTU), majoring in control theory and automation. He was recipient of ABB book prize for excellent academic result and Nanyang research scholarship award. Currently, he is pursuing his PhD at the Division of Circuit and System, EEE, NTU on statistical reliability and maintenance modeling. His research interests include statistical reliability and maintenance, stochastic system control.



CHER MING TAN was born in Singapore in 1959. He received the B.Eng. degree (Hons.) in electrical engineering from the National University of Singapore in 1984, and the M.A.Sc. and Ph.D. degrees in electrical engineering from the University of Toronto, Toronto, ON, Canada, in 1988 and 1992, respectively. He joined Nanyang

Technological University (NTU) as an academic staff in 1997, and he is now an Associate Professor in the Division of Circuits & Systems at the School of Electrical and Electronic Engineering (EEE), Nanyang Technological University (NTU), Singapore. His current research areas are reliability data analysis, electromigration reliability physics and test methodology, physics of failure in novel lighting devices and quality engineering such as QFD. He also works on silicon-on-insulator structure fabrication technology and power semiconductor device physics.

Dr. Tan was the Chair of the IEEE Singapore Section in 2006. He is also the course-coordinator of the Certified Reliability Engineer program in Singapore Quality Institute, and Committee member of the Strategy and Planning Committee of the Singapore Quality Institute. He was elected to be an IEEE Distinguished Lecturer of the Electron Devices Society (EDS) on Reliability in 2007. He is also the Faculty Associate of Institute of Microelectronics (IME) and Senior Scientist of Singapore Institute of Manufacturing Technology (SIMTech). He was also elected to the Research Board of Advisors of the American Biographical Institute and nominated to be the International Educator of the Year 2003 by the International Biographical Center, Cambridge, U.K. He is now appointed as a Fellow of the Singapore Quality Institute (SQI).

He is currently listed in *Who's Who in Science and Engineering* as well as *Who's Who in the World* due to his achievements in science and engineering.

Decision and Fusion for Diagnostics of Mechanical Components

Renata Klein¹, Eduard Rudyk², and Eyal Masad³

^{1,2,3}*R.K. Diagnostics, P.O.B. 66, Misgav Industrial Park, 20179, Israel*

Renata.Klein@rkdiagnostics.co.il

Eddie.Rudyk@rkdiagnostics.co.il

Eyal.Masad@rkdiagnostics.co.il

ABSTRACT

Detection of damaged mechanical components in their early stages is crucial in many applications. The diagnostics of mechanical components is achieved most effectively using vibration and/or acoustical measurements, sometimes accompanied by oil debris indications. The paper describes a concept for fusion and decision for mechanical components, based on vibro-acoustic signatures. Typically in diagnostics of complex machinery, there are numerous records from normally operating machines and few recordings with damaged components. Diagnostics of each mechanical component requires consideration of a large number of features. Learning classification algorithms cannot be applied due to insufficient examples of damaged components. The proposed system presents a solution by introducing a hierarchical decision scheme. The proposed architecture is designed in layers imitating expert's decision reasoning. The architecture and tools used allow incorporation of expert's knowledge along with the ability to learn from examples. The system was implemented and tested on simulated data and real-world data from seeded tests. The paper describes the proposed architecture, the algorithms used to implement it and some examples.

1. INTRODUCTION

In diagnostics and prognostics, the decision is the process that determines the probability that a certain component, module or system is in a healthy state. In order to reach the decision, health indicators from a variety of sources related to a component are combined. In the implementation described herein a multi-layer approach is used. At each layer the features of a similar nature are combined.

Two levels of decision can be identified: component-level and system-level decision.

Component-level decision generates a single decision for each component. This is a complex decision as there are many different sources of information, sometimes

contradicting, that should be taken into account.

During the system-level decision, the health of each component is translated into recommendations for maintenance operations. This level of decision should incorporate root-cause analysis (RCA) type of logic. For example, let's assume that an abnormal behavior was observed in components C1 (a massive gearwheel) and C2 (an anomaly virtual component). During system-level decision, and knowing the system dynamics, it can be concluded that the actual component that requires maintenance operation is C3 (a pinion gear), which is connected to both the big gearwheel C1 and to the indication on the anomaly C2. It can also be concluded that the pinion gear C3 and the gearwheel C1 are, for example, part of module B1 and the most efficient maintenance operation is full replacement of module B1 and not only the faulty component.

In this paper our focus is on component-level decision making. Information on the health of a single component is collected from several sources and should be integrated into a single decision. The component can be monitored by multiple sensors in several operating conditions. For each sensor and operating condition multiple health indicators can be calculated.

2. VIBRATION ANALYSIS

Analysis of vibration signals is performed in several stages. The following processing stages are implemented according to the OSA/CBM layers (MIMOSA) (see Figure 1).

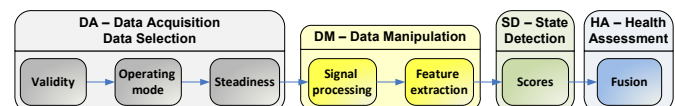


Figure 1. Vibration analysis processing stages

2.1 Data Selection

After data sampling the first step of processing is examination of the acquired data and selection of data appropriate for analysis. The data is screened in several

Renata Klein et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any

stages (see grey blocks in Figure 1): validity, operating conditions and steadiness check. Data selection is a part of the OSA/CBM data acquisition layer (DA).

The goal of the validity stage is to filter out invalid or corrupted sections of data such as sensor disconnection, saturation, spikes and others.

The next stage of data selection is recognition of predefined operating modes. Operating modes are frequently repeating conditions during system regular operation that enhance the manifestation of the damaged components (for instance when the components are loaded) and satisfy specific requirements for data analysis.

The last stage of data selection contains stationarity checks of the analyzed signals.

2.2 Data Manipulation

The OSA/CBM data manipulation (DM) layer in the current architecture is covered by signal processing and feature extraction.

In the case of vibro-acoustic data, signal processing is the most complex and computationally intensive task implicating sophisticated flows of algorithms including many transformations from one domain of analysis to another (Klein, Rudyk, Masad, Issacharoff, 2009b, Antoni & Randall, 2002, and Klein, Rudyk, Masad, 2011). During signal processing the data is transformed into different signatures (instances of a domain) that enhance manifestation of damaged components while filtering out the excitations from other components. Signal processing is done on sections of raw data selected in the data acquisition stage.

Feature extraction is a process in which the signatures are compared with signatures representing the population of 'healthy' systems. Results of the feature extraction are condition indicators (features) of the 'health' status of specific failure modes of a mechanical component. These indicators organized as a function of time are called trends.

The feature extraction process typically calculates and collects a large number of health indicators for different components of the system under test. The failure modes of a type of component are manifested in the relevant signatures according to a characteristic pattern.

The typical failure modes of a bearing are damages to inner and outer races, rotating elements, or cage. The pattern of each failure mode of a bearing can be described by several harmonics of characteristic frequencies (also known as bearing tones or pointers) with sidebands representing the amplitude modulation. More details can be found in Klein, et al. (2009a), Klein et al. (2011), Bhavaraju, Kankar, Sharma, Harsha, 2010, Li, Sopon, He, 2009, and Hariharan, Srinivasan, 2009.

2.3 Decision

The stages after feature extraction are part of the state and health assessment (SA and HA) OSA-CBM layers. A decision regarding the health status of a component is taken per run or flight of the machine.

The inputs to the decision process are normalized features. During the normalization process the distance of a feature from the distribution of the same feature in normally operating machines is calculated. Practically during normalization the Mahalanobis distance is calculated.

The decision at each stage is generated as a probability to be in one of pre-defined states, for instance three states representing component health status: 'Normal', 'Low', and 'High' indicating respectively a normally operating component, a component with a small deviation from normal, and a component with a large deviation from normal. An additional state should be considered to represent missing or incomplete information when the decision cannot be taken. In the presented application this state is named 'Unknown'. A set of the 4 probabilities corresponding to the different states is called decision vector. The decision vector generated per run is stored in a trend of decisions.

3. ARCHITECTURE OF THE DECISION AND FUSION

A single feature or health indicator is a function of component type, sensor, operating mode, processing domain, pointers (harmonics and sidebands), and type of indicator. For example, processing domain can be orders, location – first harmonic of the shaft, and indicator – energy. To obtain the decision for a component it is therefore required to undergo the following stages: combination of indicators, pointers, processing domains, operating modes, and sensors.

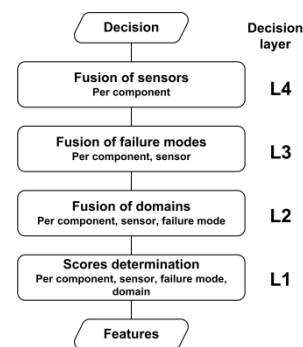


Figure 2. Layers of decision

At the first decision stage features coming from different operating modes and pointers are merged. This process is called scoring and will be denoted L1. The second decision stage (L2) merges processing domains. During the third stage (L3) of decision all the failure modes are merged. At the next and final decision stage the information from all the

sensors is joined (L4). Figure 2 shows a schematic representation of decision layers.

The architecture of the process (layers hierarchy) imitates the way an expert makes a decision. At the first stage (L1) the expert inspects a single spectrogram such as time-orders spectrogram. The expert checks the behavior of the several pointers corresponding to a failure mode as a function of the operating conditions. Depending on the component and failure mode under observation, processing domain and sensor the expert can decide whether the energy levels indicate damage.

On the next stage (L2) an expert seeks additional evidences for presence of the specific failure mode based on other processing domains such as envelope (usually by inspecting the time-orders spectrogram of the envelope). Evidences from several processing domains can strengthen or weaken the indications based on the component and failure mode under observation.

After examination of different domains all the failure modes of the component will be considered. Evidences from all the failure modes are inspected and again can weaken or strengthen the final decision. As the damage progress other failure modes may also rise due to suboptimal component operation. For example, a damage of the bearing outer race might cause a damage of the bearing roller elements.

When several sensors are used to perform the diagnostics of the component the final stage integrates their decisions. Based on relative location between the component and sensors some logic can be implemented to dismiss false positives. An example of such logic can be to take weighted voting between the sensors where the weight is proportional to the distance between the component under observation and the sensor. Such that more proximate sensors have a higher weight, but indications from several distant sensors will also be considered as indications of damage.

The scoring layer (L1) is different from the other decision layers (L2-L4). The inputs for this layer are normalized features and the output is a decision vector. In all other layers the inputs and the outputs of a decision layer are decision vectors.

4. SCORING LAYER (L1)

The first stage of the decision process is the scoring. In this stage the various features that were extracted for a certain failure mode (energy, confidence¹, pointer-location etc.) are

¹ Confidence is a feature which represents a distance of a pattern (harmonics of the carrier and corresponding sidebands) from the population of healthy machine signatures – ‘score P’ in Klein et al., 2011.

combined into a single number which may be regarded as the probability for a failure.

Features associated with the same failure mode (e.g. an outer-race pattern with sidebands and without sidebands) are joined together, and results from different operating modes are analyzed together and joined into a single result.

4.1 Scoring algorithm guidelines

The main guidelines for the development of the scores algorithm are presented below. The feature extraction process and the definition of the specific features for bearings are described in Klein et al. (2011). Note that confidence levels and pointer locations² are relevant only to bearings scores.

1. In bearing scores, if the confidence is too low the respective energy levels should be disregarded. If the confidence is high the respective energy levels should be more significant.
2. Consistently high energy and/or confidence levels should be more significant than sporadic high energies and confidence, since the latter may be caused by noise.
3. Consistency is particularly important in a feature produced in approximately similar conditions, and (in bearings) ones which have close pointer locations.
4. The final score will be a decision vector.

4.2 Algorithm description

The general scores algorithm can be separated into 5 steps as described below. The 1st stage is relevant only to bearing scores.

4.2.1 Confidence filtering (for bearing scores)

In order to accommodate the 1st guideline, we multiply the energy of each pattern by a decrease factor which is a function of the respective confidence $(c, e_1, \dots, e_N) \rightarrow (f(c)e_1, \dots, f(c)e_N)$ where f is a continuous monotonically ascending function with values in $[0,1]$. If f is properly configured, low confidence will lead to low energy levels.

4.2.2 Energy conversion

The 2nd guideline means that the affect of the energy on the score should be subadditive³. We therefore convert the energy values into new values using a continuous

² The algorithm selects the location, with the highest corresponding z-score. This selection represents the most probable location of the peaks.

³ A subadditive function is a function φ that $\varphi(x+y) < \varphi(x) + \varphi(y)$.

monotonically ascending function $\varphi: \mathbb{R} \rightarrow \mathbb{R}$, which is subadditive and in fact strictly subadditive above a certain threshold $E_0 > 0$. Below E_0 φ will be zero, so that low energies do not contribute to the score. Choosing the right function is a matter of assessing the distribution of energy values. For example a simple logarithmic function may be used.

For $e < E_0$ $\varphi(e)$ may be some small negative constant (instead of zero). φ may also be smoothed around E_0 to prevent edge effects. After all we need φ to be subadditive only for large values.

$$(u_1, \dots, u_N) = \varphi(f(c)e_1, \dots, f(c)e_N) \quad (1)$$

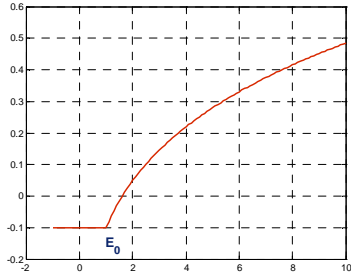


Figure 3. An example of a monotonically ascending function φ which is strictly subadditive above $E_0=1$

4.2.3 Interaction between record fragments

In the data-selection stage the recording was fragmented into intervals with similar operation modes. In each fragment the vibrations were assumed to be stationary. Each fragment was processed separately. Now we wish to compare the features extracted from various fragments and look for consistency. According to guideline 3, we need to measure proximity of conditions and pointer locations (pointer locations are only relevant to bearings, since other components have fixed pointer locations which may be determined from their geometry). We construct a metric d based on measures of RPM, load, and other operation parameters, as well as pointer shifts, which may indicate if energies are related to the same origin. Using this metric we can determine the amount of correlation we may expect between the fragments.

This correlation may then be used to increase or decrease energy levels.

$$v_{kn} = u_{kn} + h(d(k, j), u_{jn}, u_{kn}) \quad n = 1, \dots, N \quad (2)$$

where k and j are two distinct fragments.

4.2.4 Initial score estimation

In this step we turn energy levels v_{kn} into probability. This is done by a configurable fuzzy filter. In accordance with guideline 4, we use several fuzzy filters p_{knm} .

$$p_{knm} = s_m(v_{kn}) \quad (3)$$

4.2.5 Merging scores of fragments

Now we merge the scores of different record fragments. We regard the different fragments as though they were independent measurements⁴ P_{nm} .

$$P_{nm} = 1 - \prod_{k=1}^K (1 - p_{knm}) \quad (4)$$

4.2.6 Merging pointer scores

Merging pointers is a simple matter of applying a statistical function g (such as mean, percentile etc.) on the scores produced in the previous step.

$$score_m = g(P_{1m}, \dots, P_{Nm}) \quad (5)$$

The function g is applied to the results of all the patterns associated with the current failure mode. Thus, the scores of the various patterns are also merged.

4.3 Algorithm illustration

The results of the algorithm on two sets of simulated data are provided below. The first set represents features of bearing with damaged outer race OR (Figure 4), and the second set represents features of a healthy bearing with some abnormal energy levels that may occur due to feature overlap (Figure 5). Both sets of simulated features included outer race energy level for BPFO and its harmonics (ORS1 and ORS2), energies of sidebands around a bearing fault frequency peak (OR1-OR6), and confidence levels for the outer race expected pattern. The labels on y axis of both figures represent different operating conditions (fragments).

In both Figure 4 and Figure 5 when comparing graphs (a) and (b) the energies corresponding to the fragments with low confidence were decreased considerably, whereas energies corresponding to high confidence levels remained intact. On the next stage (c) consistently high levels in adjacent fragments were increased (see Figure 4). After the last stage (d) the energies that were well below the threshold yielded low probabilities. Comparing raw features (a) and the final scores (d) it can be observed that in the simulated damage (Figure 4) the scores are high whereas in the simulation of healthy bearing (Figure 5) the scores are small thus illustrating the capability of the algorithm to reduce false alarms.

⁴ The fragments merged can be considered independent because they represent separated segments of time and usually different operating modes with different load.

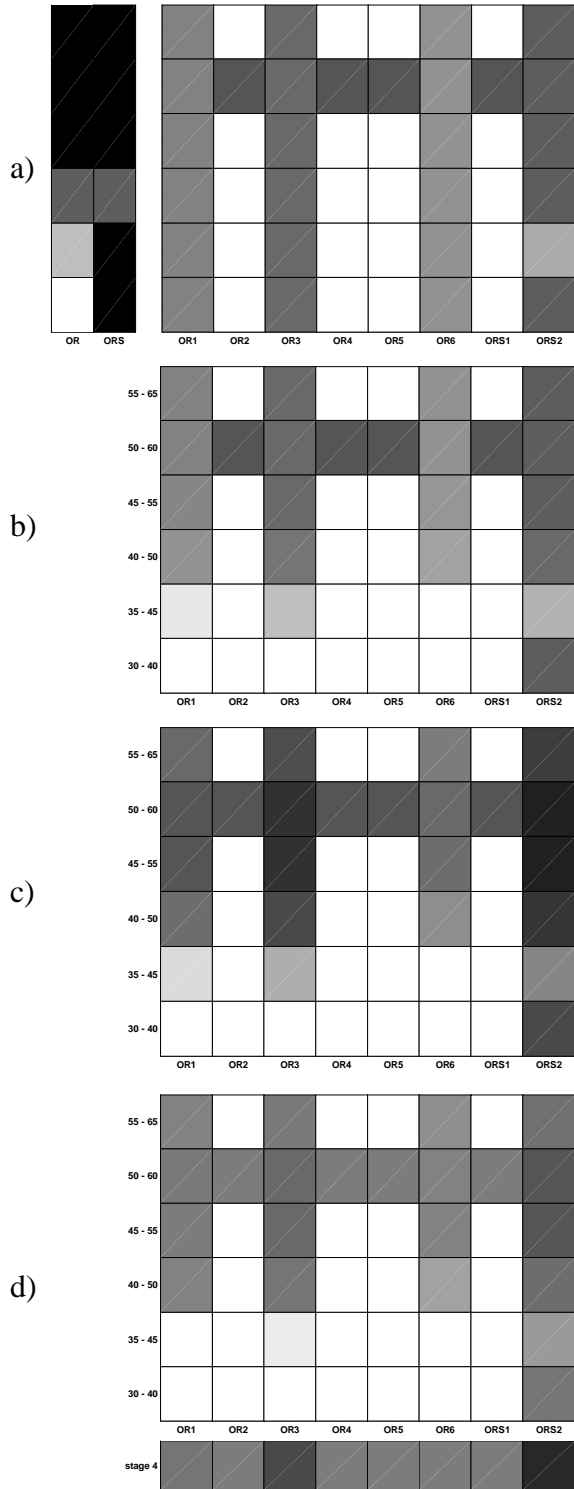


Figure 4. Results of damage in OR: a) raw features, on the left side – confidence and on the right – energy levels in logarithmic scale; b) energy levels after confidence filtering \vec{u} ; c) \vec{u} after correlation of fragments \vec{v} ; d) probability of abnormal behavior, before (upper graph) and after (lower graph) merge of record fragments.

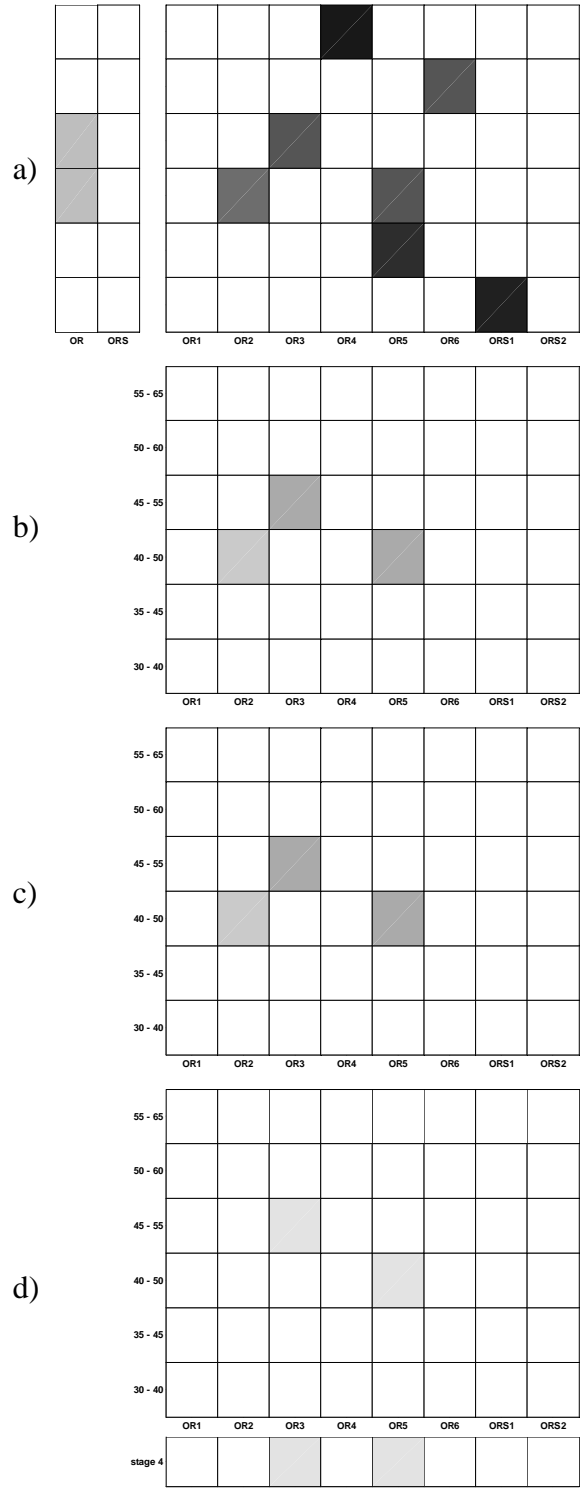


Figure 5. Results of healthy bearing: a) raw features, on the left side – confidence and on the right – energy levels in logarithmic scale; b) energy levels after confidence filtering \vec{u} ; c) \vec{u} after correlation of fragments \vec{v} ; d) probability of abnormal behavior, before (upper graph) and after (lower graph) merge of record fragments.

5. FUSION LAYERS (L2-L4)

Each decision layer (L2-L4) can be implemented by a different decision model. All decision models share an identical interface and allow a plug-and-play behavior.

In the current implementation 2 types of models were used: worst-case scenario and Bayesian network (Neapolitan, 2003).

5.1 Bayesian network model

The Bayesian network model allows definition networks of arbitrary complexity. The network is initialized with a corresponding conditional probability table (CPT). This table defines the effect of each combination of inputs on the respective output.

The model allows manual assign of CPT or learning of expected behavior using examples.

5.2 Worst-case scenario (WCS) model

The WCS model receives several decision vectors as input. The input vector with the highest deviation from the normal is selected as output of the model.

One subject that should be specifically addressed is the case of non-zero 'Unknown' probability. It is clear that if one of the inputs contains non-zero probability in an abnormal state (indicating some kind of deviation from normal behavior), the 'unknown' state should be ignored. Otherwise if all other inputs indicate completely 'normal' behavior 3 options should be considered:

1. To generate a 'normal' decision,
2. To generate an 'unknown' decision,
3. To generate a combination between 'normal' and 'unknown' states by assigning non-zero probability to each.

Each option has its own logic and should be considered depending on the application.

5.3 Model selection

Selection of the decision model for each decision layer is based on the level of mutual correlation between the merged decisions. If high level of correlation is expected between the merged decisions then it is beneficial to use the Bayesian network model. This model can incorporate complex interconnections between the elements and provide means for more sophisticated decision-making. For example, it is plausible that different processing domains (layer L2) will provide indications of declining health of component. Thus multiple weak indications may intensify the decision that the component's health is declining. In contrast, if only a single weak indication was received it

may be dismissed as no other supporting factors were detected.

On the other hand, if minor or no correlation is expected between the input elements then a WCS model is more appropriate. It actually states the health of the combination is the same as the health of the weakest (highest probability of damage) element in that combination. For example, in the L4 layer a fusion of sensors is performed. At early stages of fault development only the closer sensors will be able to detect a shift from the normal. Depending on the sensors locations as the fault development progresses more distant sensors may or may not detect some discrepancy also. So in case of insufficient information on correlation between sensors and transmission path (component-sensor) a WCS decision may be selected.

Decision modules used at each layer and corresponding parameters can be defined for each component separately based on available information and component specificity.

In current application the L2 layer (fusion of domains) is implemented by Bayesian network model. Layers L3-L4 are using WCS model.

6. ANALYSIS OF REAL DATA

Data used in this section originates from PHM '09 data challenge (Klein et al., 2009b).

The PHM09 marked data set included 280 recordings of 4 seconds, measured on the gearbox described in Figure 6, using two vibration sensors and a tachometer. All the bearings were similar. Some of the signals were recorded when the gearbox was in 'spur' configuration, and others when it was in 'helical' configuration. Data were collected at 30, 35, 40, 45 and 50 Hz shaft speed, under high and low loading.

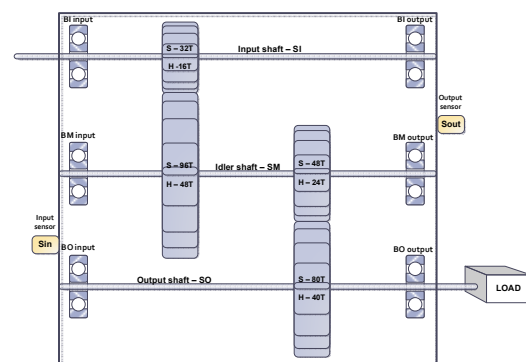


Figure 6. Challenge apparatus: spur (S) and helical (H) configurations.

The records used in the following analysis are listed in Table 1.

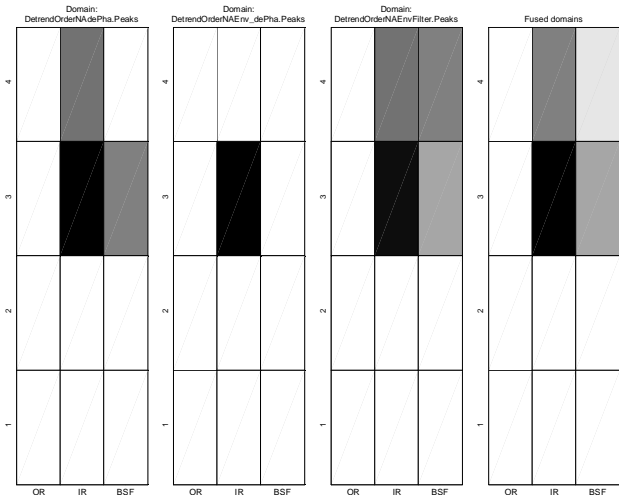


Figure 7. bA1, sensor Sin (L2)

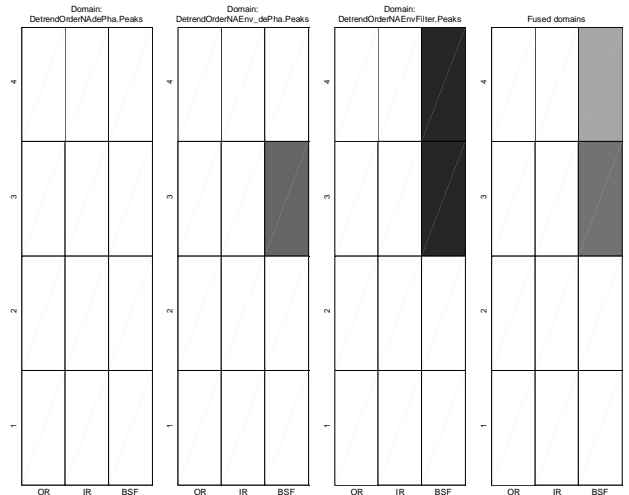


Figure 10. bB1, sensor Sin (L2)

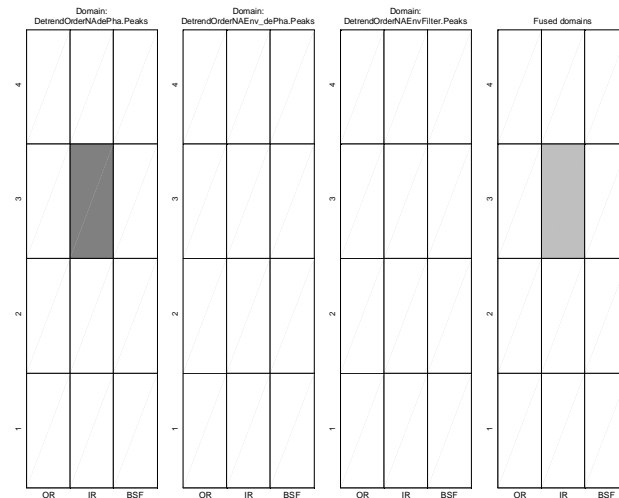


Figure 8. bA1, sensor Sout (L2)

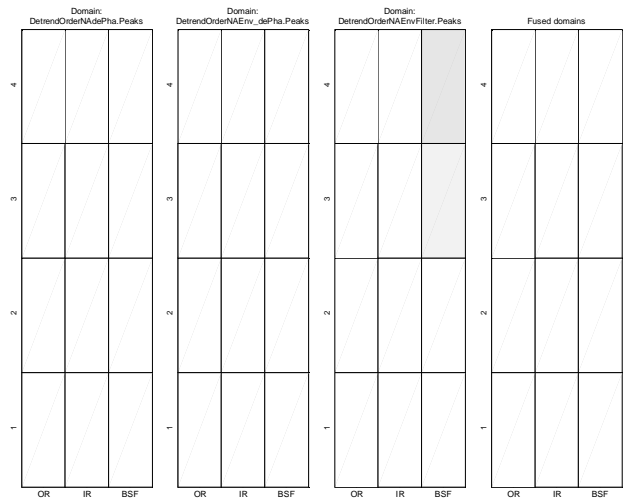


Figure 11. bB1, sensor Sout (L2)

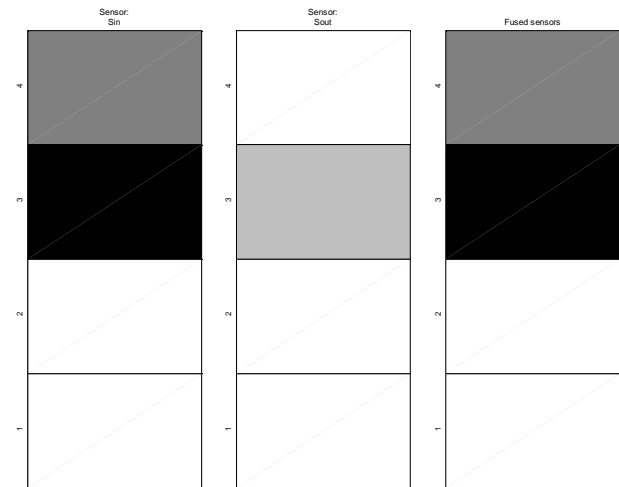


Figure 9. bA1 final decision (L4)

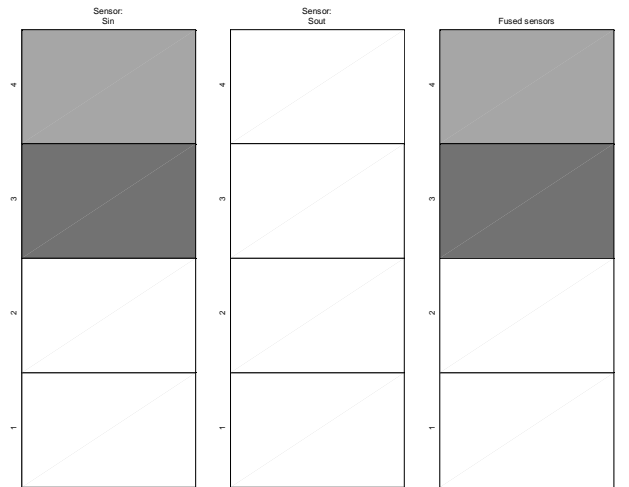


Figure 12. bB1 final decision (L4)

	Name	bA1	bB1	Other damages
1	Spur 1	Good	Good	Good
2	Spur 2	Good	Good	Gear
3	Spur 6	Inner race	Ball	Gear, Shaft
4	Spur 8	Good	Ball	Shaft

Table 1. Analyzed records and corresponding bearing damages (bA1, bB1).

Recordings from 2 sensors (called Sin and Sout) were provided. Both bearings are identical and located closer to sensor Sin. Bearing bA1 was mounted on the input shaft and bearing bB1 on the idler shaft. The corresponding bearing tones overlap since the idler shaft rotates at a third of the rotating speed of the input shaft (see Klein et al., 2009b).

Results of fusion layers are presented in Figure 7-Figure 12. Due to space limitations the score results are not presented herein. All the graphs present probabilities for damage in a gray scale color map (white represents zero probability and black a probability of 1). Conclusions maybe derived on probabilities of damage for components (L4) or failure modes (L2) according to a specific sensor. For practical purposes the decision on component probability of damage (L4) is the most relevant.

Figure 7, Figure 8, Figure 10 and Figure 11 show results of domain fusion (layer L2). Three domains corresponding to the leftmost subplots were fused. The domains that were considered were: order of the dephased signal, order of the envelope of the band-pass filtered signal, and order of the envelope of the dephased signal (see Klein et al., 2009b). The fusion result is displayed on right subplot of each figure. The columns of each subplot correspond to the bearing failure modes (IR – inner race, OR – outer race, BSF – ball), and the rows correspond to the different records as described in Table 1.

In Figure 7 incorrect indications of ball damage can be observed. This is due to the bearing tones overlap mentioned beforehand. The third harmonic of ball spin frequency (BSF) of bB1 coincides with BSF of bA1. In the case of the PHM'09 challenge apparatus the discrimination between these bearings is problematic. In practical cases this situation is rare.

Figure 9 and Figure 12 present results of failure mode and sensor fusion (layers L3 and L4 respectively). On the leftmost subplots results of layer L3 (failure mode fusion) are displayed. Each subplot corresponds to a single sensor. The rightmost subplot represents the result of layer L4 (sensor fusion) which is actually the final decision.

All damages were recognized correctly. All recordings from undamaged bearings were classified correctly as well. Moreover, the probabilities for sensor Sin were significantly higher compared to the probabilities for sensor Sout. This

may be due to the fact that the bearings are located closer to sensor Sin.

It should be noted that the damages in other components did not affect the decisions for the bearings bA1, bB1.

7. CONCLUSIONS

Hierarchical architecture of knowledge based system for decision and fusion was presented. The architecture was implemented using an original scoring algorithm and Bayesian belief networks.

The hierarchy and algorithm design was inspired by vibration expert reasoning. The system allows incorporation of expert knowledge along with ability to learn from examples.

The architecture was tested with both simulated and real data and displayed good discrimination between damaged and healthy mechanical components. Detection of the damage in bearings was not affected by damages in shafts and/or gears.

In the future the system should be checked on more extensive data collections. Implementation of additional decision models such as neural networks and other types of classifiers may be also considered. As well the condition probability tables of the Bayesian networks can be determined automatically based on examples.

REFERENCES

- Antoni, J., Randall, R. B., (2002, April), Differential Diagnosis of Gear and Bearing Faults, *Journal of Vibration and Acoustics*, Vol. 124 pp. 165-171.
- Bhavaraju, K. M, Kankar, K., Sharma, S. C., Harsha, S. P., (2010). A Comparative Study on Bearings Faults Classification by Artificial Neural Networks and Self-Organizing Maps using Wavelets, *International Journal of Engineering Science and Technology*, Vol. 2(5), 2010, pp. 1001-1008.
- García-Prada, J., C., Castejón, C., Lara, O. J., (2007). Incipient bearing fault diagnosis using DWT for feature extraction, *12th IFTOMM World Congress*, Besançon (France), June18-21, 2007.
- Hariharan, V., Srinivasan, PSS. (2009). New Approach of Classification of Rolling Element Bearing Fault using Artificial Neural Network, *Journal of Mechanical Engineering*, Vol. ME 40, No. 2, December 2009, Transaction of the Mech. Eng. Div., The Institution of Engineers, Bangladesh, pp. 119-130.
- Klein, R., Rudyk, E., Masad, E., Issacharoff M., (2009a). Emphasizing bearings' tones for prognostics, *The Sixth International Conference on Condition Monitoring and Machinery Failure Prevention Technologies*, pp. 578-587.
- Klein, R., Rudyk, E., Masad, E., Issacharoff M., (2009b). Model Based Approach for Identification of Gears and

Bearings Failure Modes, *International Journal of Prognostics and Health Management*.

Klein, R., Rudyk, E., Masad, E. (2011). Methods for diagnostics of bearings in non-stationary environment, *CM2011-MFPT2011 Conference Proceedings*. June 20-22, Cardiff, UK.

Li, R., Sopon, P., He, D., (2009). Fault features extraction for bearing prognostics, *Journal of Intelligent Manufacture*, DOI 10.1007/s10845-009-0353-z.

MIMOSA, OSA-CBM V3.1L: Open Systems Architecture for Condition-Based Maintenance, www.mimosa.org

Neapolitan, R.E., (2003). Learning Bayesian Networks, *Prentice Hall Series in Artificial Intelligence*, Prentice Hall (April 6, 2003), ISBN-13: 978-0130125347.

Ocak, H., Loparo, K. A., (2001). A new bearing fault detection and diagnosis scheme based on hidden Markov modeling of vibration signals, *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001 Proceedings (ICASSP '01)*. May 07-11, Salt Lake City, UT, USA.

analysis and topological dynamics. In the last 4 years, Eyal is an algorithm developer at "R.K. Diagnostics".

Renata Klein received her B.Sc. in Physics and Ph.D. in the field of Signal Processing from the Technion, Israel Institute of Technology. In the first 17 years of her professional career she worked in ADA-Rafael, the Israeli Armament Development Authority, where she managed the Vibration Analysis department. In the decade that followed, she focused on development of vibration based health management systems for machinery. She invented and managed the development of vibration based diagnostics and prognostics systems that are used successfully in combat helicopters of the Israeli Air Force, in UAV's and in jet engines. Renata was a lecturer in the faculty of Aerospace Engineering of the Technion, where she developed and conducted a graduate class in the field of machinery diagnostics. In the last three years, Renata is the CEO and owner of "R.K. Diagnostics", providing R&D services and algorithms to companies who wish to integrate Machinery health management and prognostics capabilities in their products.

Eduard Rudyk holds a B.Sc. in Electrical Engineering from Ben-Gurion University, Israel, M.Sc. in Electrical Engineering and MBA from Technion, Israel Institute of Technology. His professional career progressed through a series of professional and managerial positions, leading development of pattern recognition algorithms for medical diagnostics and leading development of health management and prognostics algorithms for airborne platforms, such as UAV's and helicopters. For the last 4 years Eduard is the director of R&D at "R.K. Diagnostics".

Eyal Masad received his B.Sc., M.Sc. and Ph.D. degrees from the Faculty of Mathematics in the Technion, Israel Institute of Technology. His research topics were in the fields of Machine learning, Information theory, nonlinear

Defect source location of a natural defect on a high speed-rolling element bearing with Acoustic Emission

B Eftekharnjad¹, A. Addali² and D Mba²

¹Renewable Energy Systems Ltd, Kings Langley, WD4 8LR United Kingdom

²School of Engineering, Cranfield University, Bedford, England, MK43 0AL

Email: A.Addali@cranfield.ac.uk

Abstract

The application of Acoustic Emission (AE) technology for machine health monitoring is gaining ground as powerful tool for health diagnosis of rolling element bearings. The successful application of AE to life prognosis of bearings is very dependent on the ability of the technology to identify and locate a defect at its earliest stage. Determining source locations of AE signals originating in real time from materials under load is one of the major advantages of the technology. This paper presents results which highlight the ability of AE to locate naturally initiated defects on high-speed roller element bearing in-situ. To date such location has only be successfully demonstrated at rotational speeds of less than 100 rpm.

1. Introduction

The rolling element bearing is the most common part of rotating machines. The continued interest in condition-based maintenance of industrial assets has lead to a growing interest in monitoring of rolling bearings. The application of Acoustic Emission (AE) in monitoring the rolling element bearings has grown in popularity over the past few decades [1]. To date most of the published work has studied the applicability of AE technology in detecting seeded faults artificially introduced on the bearing. Yoshioka [2] was one of the earliest researchers who studied the

applicability of AE in detecting naturally degraded bearings. Later, Elforjani et al [3] conducted an experiment aimed at building on Yoshioka's work. Their results showed the effectiveness of AE in detecting the onset of bearing failure, identifying the circumferential location of the defect on the race at very early stages of degradation, and the diagnostic potential of enveloping AE signatures. Although conclusive, this research was not representative of the broad operation range of bearings as the test was undertaken at a slow rotational speed (72 rpm). The results presented in this paper aims to complement the work of Elforjani [3, 4] by experimentally investigating the use of AE for detecting and locating the natural pitting of a bearing rotating at 1500 rpm in which significantly higher background AE operating noise is expected.

2. Experimental Setup

The test rig used in this experiment is displayed in **Figure 1**. The bearing test rig has been designed to simulate varying operating conditions and accelerate natural degradation. The chosen bearing for this study was an SKF single thrust ball bearing, model number SKF51210. To ensure accelerated failure of the race the standard grooved race was replaced with a flat race, model number SKF 81210TN. This caused a point contact between the ball elements and the race resulting in faster degradation of the race and early initiation of sub-surface fatigue cracks. The

Eftekharnjad et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

load on the test bearing was applied by a hand operated hydraulic pump (Hi-Force No: HP110-Hand pump-Single speed-Working Pressure: 700 BAR). The flat race was fitted onto the loading shaft in a specifically designed housing. This housing was constructed to allow for placement of AE sensors directly onto the race. Modifications were made to the support of the flat bearing race so as to allow positioning of the AE sensors, see Figure 2. The placement of the AE sensors was such that it facilitated the identification of the source of AE during operation. The motor on the rig operated at 1500rpm and the number of rolling element in the test bearing was 14 and the ball pass frequency (BPF) was 175Hz.

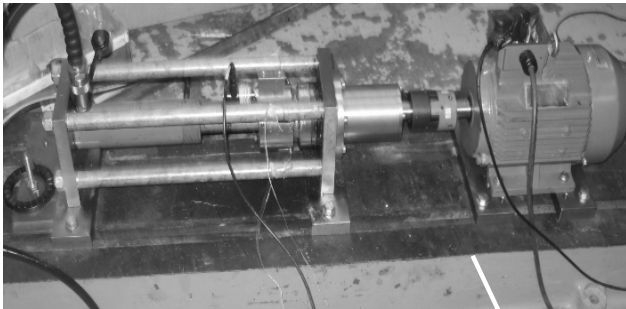


Figure 1 Test rig assembly

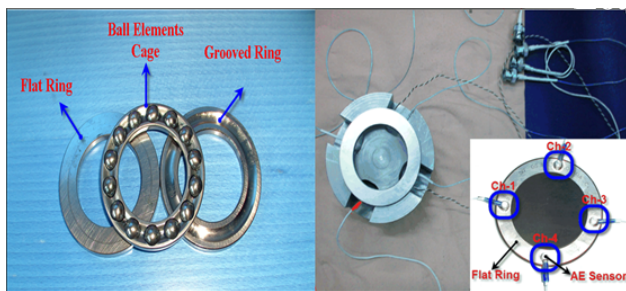


Figure 2 Test bearing and sensor arrangement on the flat race.

The AE acquisition system employed commercially available piezoelectric sensors (Physical Acoustic Corporation type 'PICO') with an operating range of 200–750 kHz at temperatures ranging from 265 to 1770C. The AE sensors were connected to a data acquisition system through a preamplifier (40dB gain). The

system was set to continuously acquire AE absolute energy (atto-Joules) over a time constant of 10 ms at a sampling rate of 100 Hz. The absolute energy is a measure of the true energy and is derived from the integral of the squared voltage signal divided by the reference resistance (10 k-Ohms) over the duration of the AE signal. For these tests a fixed sample length (250msec) of AE waveforms were captured every 60 seconds at 2 MHz sampling frequency. Throughout the test AE HITs were also acquired. An Acoustic Emission HIT is normally described by several parameters such as threshold, duration, counts and rise time. The AE signal duration is the time between the first and last amplitude threshold crossing while the rise time is the time between the start of the HIT and the instant at which the maximum amplitude of HIT is reached, see figure 3. Also, a single AE event can be produced by number of AE HITs. In addition to this, the timing parameters employed for defining an event during these experiments included the HIT definition time (HDT), HIT lockout time (HLT) and peak definition time (PDT) and these were set at 500 μ sec, 500 μ sec and 100 μ sec respectively. Correctly setting the PDT will result in an accurate measurement of peak amplitude while the appropriate definition of HDT will ensure that each signal generated from the structure is reported as one HIT. As it defines the period over which a HIT can be acquired. With an accurate setting of HLT spurious measurement during the signal decay will be avoided [5]; essentially it defines the period between successive HITs; its second function is to inhibit the measurement of reflections. In addition, an accelerometer (ISO BASE Endevco 236 with repose between 10 and 8000 Hz) was mounted on the flat race housing and vibration measurements were acquired at sampling of 10 kHz at three-minute intervals using a NI-6009 USB analog to digital data acquisition card.

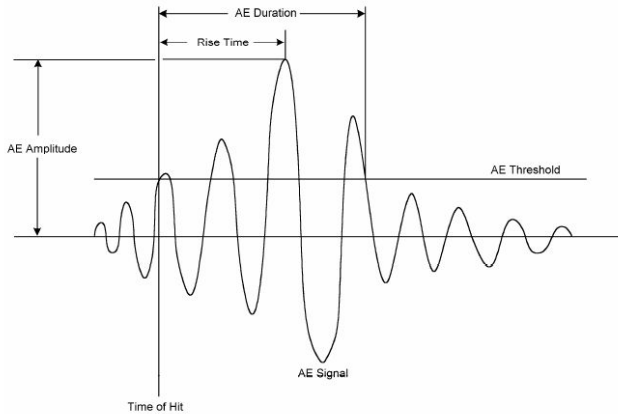


Figure 3 Schematic representation of an AE Hit [5]

3. Test procedure

For the purpose of this experiment the following procedure was undertaken to determine the subsurface stresses on the test bearing and thereby estimate the time, or number of cycles, prior to a surface defect on a track. Theories employed for this procedure, particularly for the flat race, included the Hertzian theory for determining surface stresses and deformations, the Thomas and Hoerhsh theory for subsurface stress, and the Lundberg and Palmgren theory for fatigue evaluation. For the grooved race the standard procedure, as described by BS 5512,1991, was employed for determining dynamic load rating. The theoretically determined life was calculated to be approximately 16 hours though the actual test duration was significantly longer. The test rig was allowed to operate until a spall was induced on the flat race and figure 4 shows the developed defect upon the termination of the tests. At this time abnormal vibration levels were registered and the rig was stopped. A load of approximately 50000N was applied on the bearing throughout the test. The test was stopped at 278 minutes though the AE measurement failed after 220 min due to excessive temperatures experienced on the bond holding the sensor to the race.

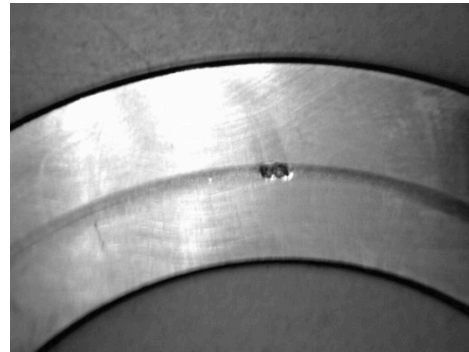


Figure 4 Defect on the outer race at termination of bearing test

4. Results and Discussion

4.1 Real time monitoring of the Vibration and AE levels

The overall trends of Acoustic Emission activity and the vibration r.m.s noted for the duration of both tests are presented in Figure 5. There was an initial rise in AE and vibration levels at the very start of the tests. This is associated with the run-in period. After this period both vibration and AE levels remained level for approximately 40mins after which a noticeable increase in AE was again observed from 40mins of operation though vibration levels remained constant. The drop in vibration levels at 50mins into operation was due to a glitch in the vibration recording system that was fixed immediately. Comparing the overall trend of vibration and AE r.m.s it is evident that the AE is more sensitive in monitoring the progression of the defect. This was because the AE level began to increase continuously much earlier than vibration levels. It must be noted that these are accelerated failure tests and the difference in time between these techniques (AE and vibration) in identification of the defect will most certainly be much longer for non-accelerated test conditions; further highlighting the increased sensitivity of the AE technology.

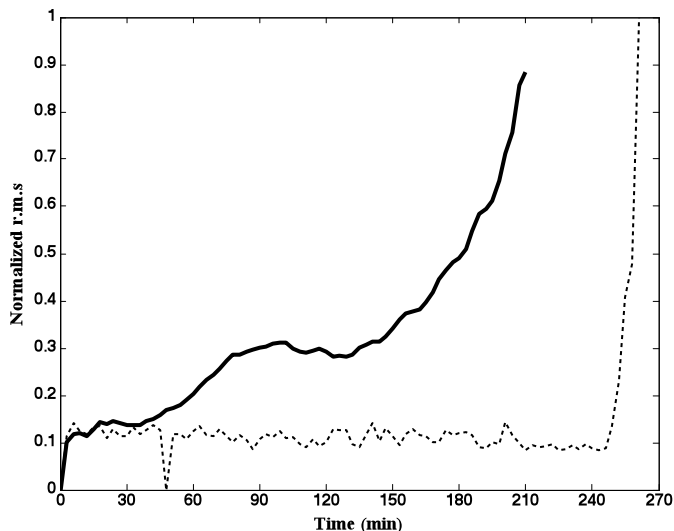


Figure 5 Overall AE (--) and vibration (....) r.m.s levels

The AE and vibration waveforms upon the termination of the test are presented in Figure 6. Evident were AE burst spaced at 175 Hz that corresponds to the bearing defect frequency, though not evident on the vibration plot. Also values of Kurtosis and Crest Factor¹ (CF) associated with AE signal are significantly higher than corresponding levels of vibration (Crest factor and Kurtosis values of 13 and 14.2 respectively for AE, and, 0.2 and 2.6 respectively for vibration), see Figure 6. This reiterates the diagnostic advantage of AE over vibration; as it is more sensitive to damage detection [1]. A time-frequency plot of a section of AE wave associated with a surface defect showed a broad frequency range (100 to 600 kHz), see figure 7. This shows the significant high frequency content of AE associated with the bearing defect. A Gabor wavelet transform was employed to determine the time-frequency spectrum. For the wavelet analysis *AGU-Vallen Wavelet* software, developed by Vallen System GmbH, was employed [6]. Given this well-established view that AE is more sensitive than vibration, the aim of this paper is not to re-iterate the obvious but to assess the applicability of AE to locate the position of the

¹ The CF defined as the ratio of the peak value divided by the signal r.m.s.

growing defect in-situ on a high speed bearing in comparison to slow speed bearing tests that have already shown defect location with AE.

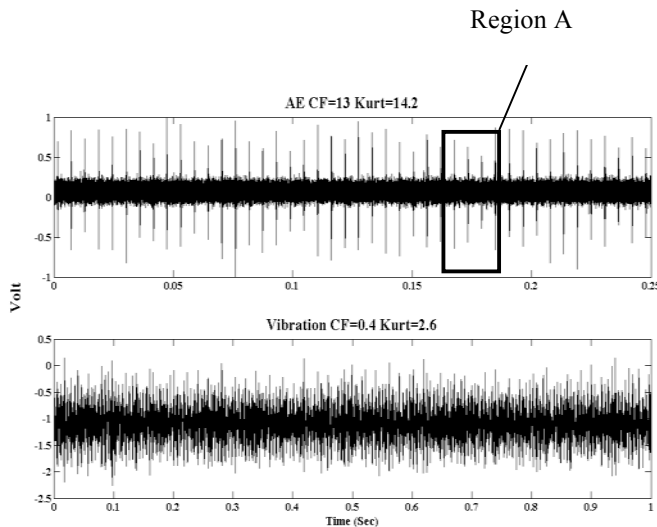


Figure 6 Acoustic Emission and vibration waveforms associated with the damaged bearing

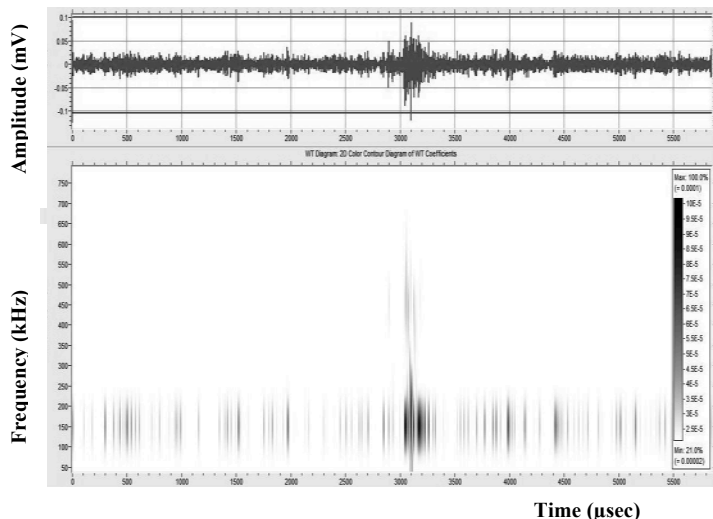


Figure 7 Time-frequency plot associated with a single AE burst from region A of Figure 6

4.2 Defect source location

The most common method for source location involves employing differences in time of arrival of waves at the receiving sensor. Given the actual

location of the AE sensor and the wave velocity for the bearing material, the location of the AE source can be determined. The sensor positions on the race allow linear location of the source to be calculated which involves linear interpolation between the coordinates of two adjacent sensors based on the differences in arrival time at the receiving sensors. Simulated AE sources on the test bearing race (Hsu-Nelson) showed the dominant frequency content of AE's recorded to be approximately 300 kHz which corresponds to a velocity of 4000 m/s for the symmetric zeroth lamb wave mode (S_0) on steel at 1.8 mm MHz (0.3 MHz, and 6 mm thick race). This velocity was used for all source location investigations and prior to the onset of testing several Hsu-Nielsen sources were made at various positions on the surface to establish the accuracy at this velocity and specific threshold level. Results were within 4% of the exact geometric location of the Hsu-Nielsen sources. For this investigation, a threshold of 70 dB was set and whenever the threshold was exceeded, the location of the source is computed and identified. Further, any AE event detected above this threshold is assigned to the geometric position (source); this is a cumulative process and as such a fixed source will have the largest contributory events in a cumulative plot.

Figure 8 presents such a cumulative plot detailing location results for the test at three chosen operating times. The x-axis represents the circumferential distance between each sensor; the position of each sensor is detailed on each of the plots in figure 8. The y-axis of figure 8 details the number of AE events captured during the test. Observations showed that at 120mins and 180mins into operation the recorded events suggested activity in the vicinity of sensor 4, however by 206mins into operation, a large number of AE events were registered between sensor-1 and -2 suggesting the development of surface damage. The location of this abnormally high concentration of AE events matched the location of damage upon the termination of test, see figure 4. The events noted earlier in the test

are attributed to spurious AE activity. Interestingly, the identification of the defect location become evident from the cumulative plots at approximately 200 minutes into operation even though AE levels had seen rising from 60 minutes into the test. The inability to identify the location much earlier into the test, unlike observations at the lower rotational speeds [3,4,7], is attributed to the higher operating background noise that makes identification of AE HITs more difficult. A direct comparison of AE operating noise at 72 rpm and 1500 rpm was noted to be 52 dB and 70 dB respectively. Such high operational noise level (70dB) could make source location significantly more challenging than at low rotational speeds. To enhance the ability to identify the defect location earlier would require advanced noise cancellation techniques.

5. Conclusion

The applicability of AE for source location of bearing defect in-situ has been demonstrated. Threshold levels above operating background levels have been shown to be sufficiently suitable for differentiating AE time of arrival intervals. This conclusion has been derived based on results from tests on a few experiments. Whilst the probability of having four AE sensors placed on a bearing race is limited it can be employed as a quality control tool for bearing manufacturers or applied on bespoke critical bearings.

6. References

- [1] Mba, D., and Rao, R. B. K. N., 2006, "Development of Acoustic Emission Technology for Condition Monitoring and Diagnosis of Rotating Machines: Bearings, Pumps, Gearboxes, Engines, and Rotating Structures," *Shock and Vibration Digest*, 38(1) pp. 3-16.
- [2] T. Yoshioka, 1992, "Application of Acoustic Emission Technique to Detection of Rolling Bearing Failure," *J. Soc. Tribologists Lubrication Eng*, 49.

[3] Elforjani, M., and Mba, D., 2009, "Assessment of Natural Crack Initiation and its Propagation in Slow Speed Bearings," *Nondestructive Testing and Evaluation*, 24(3) pp. 261.

[4] Elforjani, M., and Mba, D. [2010], "Accelerated Natural Fault Diagnosis in Slow Speed Bearings with Acoustic Emission," *Engineering Fracture Mechanics*, .

[5] Physical Acoustic Co., Pci2-Based AE system user manual

[6] <http://www.vallen.de/>

[7] Elforjani, M., and Mba, D., 2008, "Observations and Location of Acoustic Emissions for a Naturally Degrading Rolling Element Thrust Bearing," *Journal of Failure Analysis and Prevention*, 8(4) pp. 370-385.

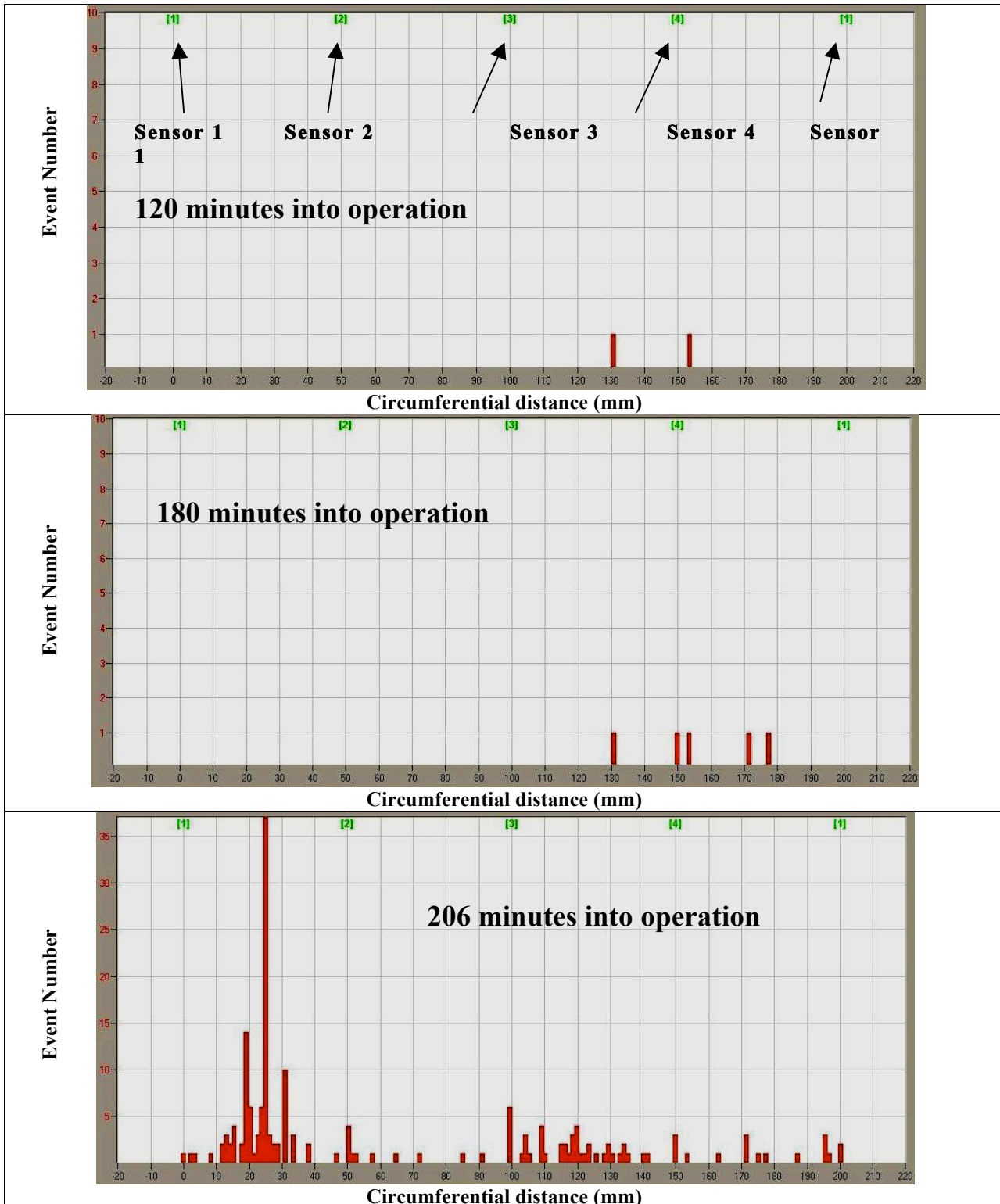


Figure 8 Acoustic Emission events against sensor position at different time intervals

Deriving Bayesian Classifiers from Flight Data to Enhance Aircraft Diagnosis Models

Daniel L.C. Mack¹, Gautam Biswas¹, Xenofon D. Koutsoukos¹, Dinkar Mylaraswamy², and George Hadden²

¹ *Vanderbilt University, Nashville, TN, 37203, USA*
daniel.l.mack@vanderbilt.edu
gautam.biswas@vanderbilt.edu
xenofon.koutsoukos@vanderbilt.edu

² *Honeywell Aerospace, Golden Valley, MN 55422, USA*
dinkar.mylaraswamy@honeywell.com
george.d.hadden@honeywell.com

ABSTRACT

Online fault diagnosis is critical for detecting the onset and hence the mitigation of adverse events that arise in complex systems, such as aircraft and industrial processes. A typical fault diagnosis system consists of: (1) a reference model that provides a mathematical representation for various diagnostic monitors that provide partial evidence towards active failure modes, and (2) a reasoning algorithm that combines set-covering and probabilistic computation to establish fault candidates and their rankings. However, for complex systems reference models are typically incomplete, and simplifying assumptions are made to make the reasoning algorithms tractable. Incompleteness in the reference models can take several forms, such as absence of discriminating evidence, and errors and incompleteness in the mapping between evidence and failure modes. Inaccuracies in the reasoning algorithm arise from the use of simplified noise models and independence assumptions about the evidence and the faults. Recently, data mining approaches have been proposed to help mitigate some of the problems with the reference models and reasoning schemes. This paper describes a Tree Augmented Naïve Bayesian Classifier (TAN) that forms the basis for systematically extending aircraft diagnosis reference models using flight data from systems operating with and without faults. The performance of the TAN models is investigated by comparing them against an expert supplied reference model. The results demonstrate that the generated TAN structures can be used by human experts to identify improvements to the reference model, by adding (1) new causal links that relate evidence to faults, and different pieces of evidence, and (2) updated thresholds and new monitors that facilitate the derivation of more precise evidence from the sensor data. A case study shows that this improves overall reasoner performance.

This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

1. INTRODUCTION

An important challenge facing aviation safety is early detection and mitigation of adverse events caused by system or component failures. Take an aircraft, which consists of several subsystems such as propulsion, avionics, bleed, flight control, and electrical; each of these subsystems consists of several dozen interacting components within and between subsystems. Faults can arise in one or more aircraft subsystems; their effects in one system may propagate to other subsystems, and faults may interact. To detect these faults, an onboard fault diagnosis solution must be able to deal with these interactions and provide an accurate diagnostic and prognostic state for the aircraft with minimal ambiguity.

The current state of online fault diagnosis is focused on installing a variety of sensors onboard an aircraft along with reasoning software to automatically interpret the evidence generated by them, and infer the presence of faults. One such state of the art system is the Aircraft Diagnostic and Maintenance System ADMS (Spitzer, 2007) that is used on the Boeing B777. ADMS can be broadly categorized as a model-based diagnoser that separates system-specific knowledge and the inferencing mechanism.

Consider characteristics of some typical faults arising in aircraft subsystems. Turbine blade erosion is a natural part of turbine aging and wearing of the protective coating due to microscopic carbon particles exiting the combustion chamber. As the erosion progresses over time, it starts to affect the ability of the turbine to extract mechanical energy from the hot expanding gases. Eventually this fault manifests itself as increase in fuel flow and gradual degradation of engine performance. This causal propagation of faults is usually known to a domain expert and captured mathematically using a static system reference model. As evidence gets generated by aircraft installed sensors, a reasoning algorithm “walks” the relevant causal paths and infers the current state of the aircraft—in this case, turbine erosion of the propulsion engine.

The ADMS uses a fault propagation system reference model that captures the interactions between aircraft components under various operating modes. A Bayesian belief propagation network together with the

Bayesian update rule provides an ideal framework for onboard diagnostic reasoning using the reference model. It provides the necessary transparency for certification as a safety system, while allowing the subsystem manufacturer to encode proprietary fault models. The generation of this reference model is mostly a manual process, and often the most tedious step in the practical development and deployment of an ADMS. While most of the knowledge about fault propagation can be derived from earlier aircraft designs, upgrades to component design (for example using active surge control rather than passive on-off surge prevention) create gaps in the knowledge base. As the engineering teams “discover” the new knowledge from an operating fleet, they are translated into expert heuristics that are added to the specific aircraft model, rather than applying systematic upgrades to the overall reference model that was generated at design time.

Many of the shortcomings of the ADMS can be attributed to incomplete and incorrect information in the system reference model. In other words, there is a missing link in making systematic upgrades and increments to the reference model as vast amount of operational data is collected by operating airlines. We look at this problem as a “causal structure discovery” problem. Specifically, learning causal structures in the form of a Bayesian Network built for classical fault diagnosis, wherein the nodes represent system faults and failures (causes) and available diagnostic evidence (symptoms) (Pearl, 1988). Unlike associations, Bayesian networks can be used to better capture the dependencies among failures (failure cascade from one subsystem to another) and evidence cascade (failure mode in one system triggering a symptom in a nearby component). We adopt this approach, and develop a data mining approach to updating existing reference models with new causal information.

This paper presents a case study, an adverse event surrounding an in-flight shutdown of an engine, which was used to systematically augment an existing ADMS reference model. Section 2. describes the basic principles and the constituents of a model-based onboard fault reasoner. Section 3. describes the problem statement that formally defines the model augmentation to be derived using operational data. Next, section 4. describes the available historic data surrounding the adverse event. Section 5. briefly discusses the challenges in taking operational data and transforming it into a form that can be used by the data mining algorithms. Section 6. then discusses these data mining algorithms employed for constructing the diagnostic classifiers as Tree-Augmented Naïve Bayesian Networks (TANs). Section 7. presents experimental results of this case study to show how a human expert could utilize the classifier structure derived from flight data to improve a reference model. Metrics are defined for evaluating classifier performance, and a number of different experiments are run to determine when improvements can be made in the existing model. Section 8. presents a summary of the approach, and outlines directions for future work for diagnostic and prognostic reasoning using the data mining algorithms.

2. BACKGROUND ON REFERENCE MODELS AND REASONERS

Model-based strategies that separate system-specific knowledge and the inferencing mechanism are preferred

for diagnosing large, complex, real-world systems. An aircraft is no exception to this, as individual component suppliers provide system-specific knowledge that can be represented as a bipartite graph consisting of two types of nodes: failure modes and evidence. Since this knowledge acts as a baseline for diagnostic inferencing, the term “reference model” is also used to describe this information. The set F captures all *distinct* failure modes defined or enumerated for the system under consideration. A failure mode $fm_i \in F$ may be occurring or not occurring in the system, which is indicated by a 1 (occurring) or 0 (not occurring) state. Often a -1 unknown state is also included in the initial state description. The following are shorthand notations regarding these assertions.

$$\begin{aligned} fm_i = 0 &\Leftrightarrow \text{The failure mode is not occurring} \\ fm_i = 1 &\Leftrightarrow \text{The failure mode is occurring} \end{aligned} \quad (1)$$

Every failure mode has an a priori probability of occurring in the system. This probability is given by $P(fm_i = 1)$. Failure modes are assumed to be independent of one another, i.e., given any two failure modes fm_k and fm_j , $P(fm_k = 1 | fm_j = 1) = P(fm_k = 1)$.

To isolate and disambiguate the failure modes, component suppliers also define an entity called “evidence” that is linked to sensors and monitors in the system. The set DM denotes all distinct diagnostic monitors defined for the system under consideration. A diagnostic monitor associated with $m_j \in DM$, can either *indict* or *exonerate* a subset of failure modes called its ambiguity group. In other words, each monitor m_i in the system is labeled by three mutually exclusive values allowing a monitor to express indicting, exonerating or unknown support for the failure modes in its ambiguity group. The notations are described in equation (2).

$$\begin{aligned} m_i = 0 &\Leftrightarrow \text{Exonerating evidence} \\ m_i = 1 &\Leftrightarrow \text{Indicting evidence} \\ m_i = -1 &\Leftrightarrow \text{Unknown evidence} \end{aligned} \quad (2)$$

An ideal monitor m_j fires only when one or more failure modes in its ambiguity group are occurring. Given the fact that the i^{th} failure mode is occurring in the system, d_{ji} denotes the probability that monitor m_j will provide indicting evidence under this condition.

$$d_{ji} \Leftrightarrow P(m_j = 1 | fm_i = 1), \quad (3)$$

d_{ji} is called the detection probability of the j^{th} monitor with respect to failure mode fm_i . A monitor may fire when none of the failure modes in its indicting set are present in the system. *False alarm probability* is the probability that an indicting monitor fires when its corresponding failure modes in its ambiguity group are not present in the system. That is,

$$\epsilon_j \Leftrightarrow P(m_j = 1 | fm_i = 0, \forall fm_i \in \text{Ambiguity Set}) \quad (4)$$

Designing a monitor often requires deep domain knowledge about the component or subsystem, but the details of this information may not be important from the reasoner’s viewpoint. A more abstract view of the monitor is employed in the reasoning algorithm. This abstraction is shown in Figure 1. With few exceptions, most

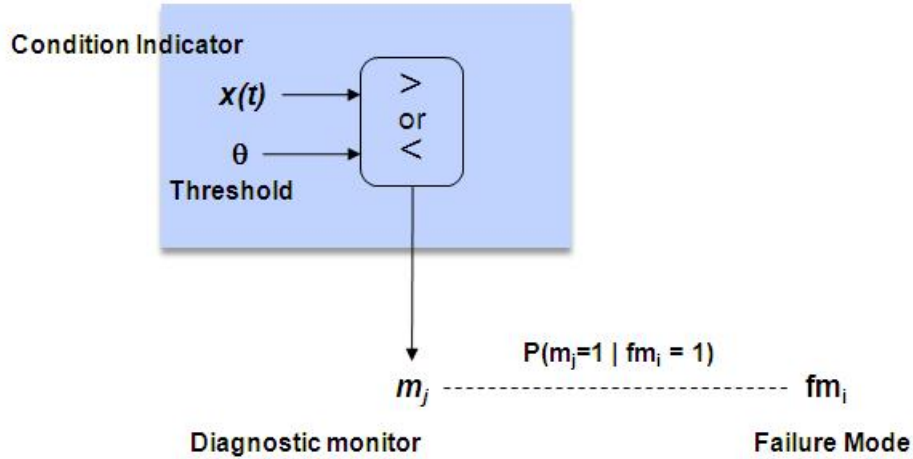


Figure 1: Abstraction of Diagnostic monitor

diagnostic monitors are derived by applying a threshold to a time-series signal. This signal can be a raw sensor value or a derived quantity from a set of one or more sensor values. We call this a condition indicator (CI) and denote it as $x(t)$. Assuming a pre-defined threshold value θ , we set $m = 1 \Leftrightarrow x(t) \leq \theta$. A diagnostic monitor may specify the underlying condition indicator and the threshold or simply provide the net result of applying a hidden threshold.

Figure 2 illustrates an example reference model graphically, with fault modes (hypotheses) as nodes on the left, and diagnostic monitors $m_i \in DM$ on the right. Each link has an associated detection probability, i.e., conditional probability $P(m_j = 1 | fm_i = 1)$. In addition, fault nodes on the left contain the *a priori* probability of fault occurrence, i.e., $P(fm_i)$. Probabilities on the DM nodes indicate the likelihood that a particular monitor would indicate a fault in a nominal system. Bayesian methods are employed to combine the evidence provided by multiple monitors to estimate the most likely fault candidates.

The reasoner algorithm (called the W-algorithm) combines an abductive reasoning scheme with a forward propagation algorithm to generate and rank possible failure modes. This algorithm operates in two steps: (1) *Abductive reasoning step*: Whenever a diagnostic monitor m_1 fires, it provides either indicting (if $m_1 = 1$) or exonerating (if $m_1 = 0$) evidence for the failure modes in its ambiguity set, $AG = \{fm_1, fm_2, \dots, fm_k\}$. This step assumes that the firing of a DM implies at least one of the faults in the ambiguity set has occurred; and (2) *Forward reasoning step*: For each fm_i belonging to AG , this step calculates all other diagnostic monitors that may fire if any of the failure modes are indeed occurring. These are called the evidence of interest. Let m_2, m_3, \dots denote this evidence of interest set. Some of these monitors may be indicting evidence, for example $m_2 = 1$ or they may be exonerating evidence, for example $m_3 = 0$. The reasoning algorithm calculates the joint probability $P(fm_1 = 1, m_1 = 1, m_2 = 1, m_3 = 0, \dots)$ of a specific failure mode fm_1 occurring in the system. As additional monitors fire, the numeric values of these probabilities increase or decrease, till a specific failure mode

hypothesis emerges as the highest-ranked or the most likely hypothesis. The reasoning algorithm can generate multiple single fault hypotheses, each hypothesis asserting the occurrence of exactly one failure mode in the system.

The reasoning algorithm may not succeed in reducing the ambiguity group to a single fault element. This can happen for various reasons: (1) incompleteness and errors in the reference model; (2) simplifying assumptions in the reasoning algorithm; and (3) missing evidence (monitors) that support or discriminate among fault modes. For example, a modest aircraft has over 5000 monitors and failure modes; estimating the detection probabilities, d_{ji} , even for this aircraft is a challenging offline design task. Errors in d_{ji} , and more specifically missing the link between a monitor and a failure mode (incompleteness) can adversely affect the reasoner performance. Further, to keep things simple for the reasoner, a modeler may assume that the firing events for monitors are independent. This eliminates the need for the modeler to provide joint probability values of the form, $P(m_j = 1, m_k = 1 | fm_i = 1)$ (say for 4000 monitors and 1000 faults the modeler would have to provide 1.56×10^{10} probability values), and instead approximate it as $P(m_j = 1 | fm_i = 1) \times P(m_k = 1 | fm_i = 1)$. This reduces the total number of probabilities to 4×10^6 , which is still a large number but orders of magnitude less than the number required for the joint distributions and the order of the joint distributions grow exponentially when additional monitors fire. Designing a good set of monitors is yet another challenging task. For example, the modeler may have overlooked a new monitor m_p that could have differentiated between failure modes fm_1 and fm_2 .

Given the complexity of large systems such as an aircraft, incompleteness in the reference model is expected. As one collects enough operational data, some of these gaps can be addressed. The collection of assumptions made about the fault hypotheses and monitors results in the probability update function for each fault hypothesis, $fm_i \forall i \in F$, being computed using a Naïve Bayes model, i.e., $P(fm_i | m_j, m_k, m_l \dots) = \alpha \times P(m_j, m_k, m_l \dots | fm_i) = \alpha \times P(m_j | fm_i) \times$

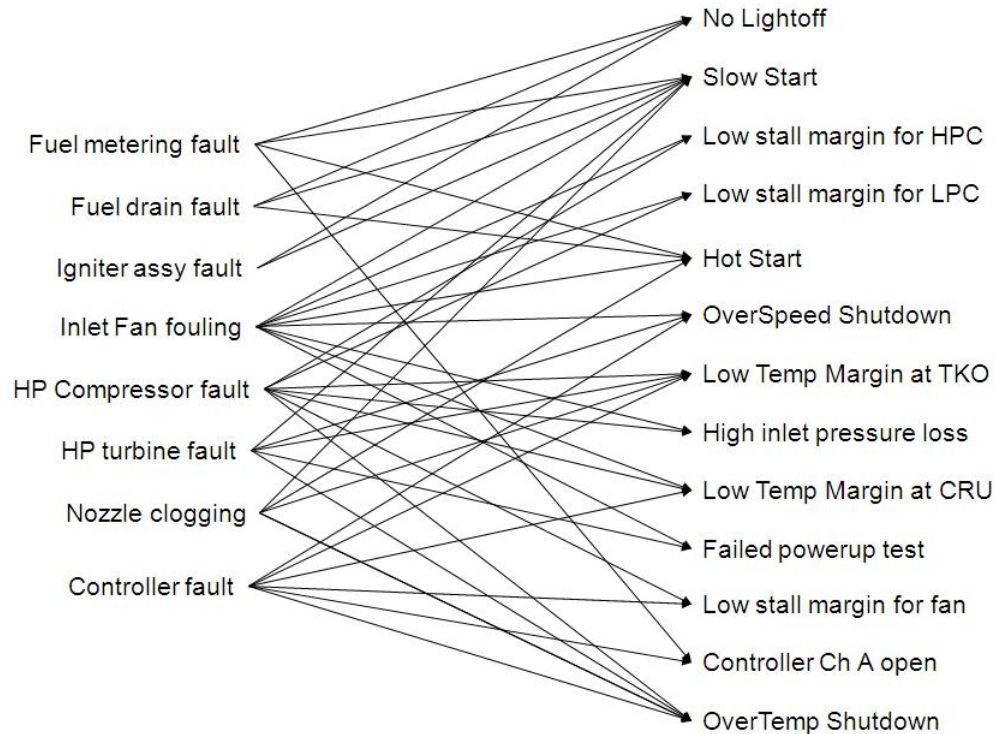


Figure 2: Example Reference Model

$P(m_j|f m_i) \times P(m_j|f m_i) \times \dots$ where α is a normalizing constant. The direct correspondence between the reference model and the simple Bayesian structure provides opportunities to use a class of generative Bayesian model algorithms to build structures that are relevant for diagnostic reasoning from data. These newly learned structures can then be used with a systematic approach for updating the system reference model. The following work focuses on this systematic approach.

3. PROBLEM STATEMENT

The combination of the reference model and reasoner when viewed as a single fault diagnoser can be interpreted as a Noisy-OR classifier, which is a simplified form of a standard Bayesian Network. These networks that model diagnostic information (i.e., monitor-fault relations) can be built from data itself as a practical application of data mining. A number of Machine Learning techniques for building Bayesian networks from data have been reported in the literature (Friedman, Geiger, & Goldszmidt, 1997), (Cheng, Greiner, Kelly, Bell, & Liu, 2002), (Grossman & Domingos, 2004). For example, state-based hidden Markov Models (HMMs) (Smyth, 1994) and even more general Dynamic Bayesian Network (DBN) (Dearden & Clancy, 2001), (Lerner, Parr, Koller, & Biswas, 2000), (Roychoudhury, Biswas, & Koutsoukos, 2008), (Verma, Gordon, Simmons, & Thrun, 2004) formulations can be employed to capture the dynamics of aircraft behavior and effects of faults on system behavior and performance. However, rather than addressing the problem as a traditional data mining problem, it is approached as an

application that works to extend an existing ADMS. In other words, the output of the data mining algorithms have to be designed to provide information that supplements existing expert-generated reference models, as opposed to providing replacement forms of the reference model with corresponding reasoner structures. This technique can be construed as a method for supporting human experts, by having a human in the loop to interpret the findings generated by the data mining algorithm and make decisions on how to modify and update the existing reference model and reasoner structures. By including humans, who verify and integrate the model enhancements, we turn the verification into a straightforward task for the human to either approve the selected changes, or ignore them.

A systematic approach to the data mining task in this context is to discover elements of the reference model that can be augmented by comparing the results of a data driven model produced from using a learning algorithm against an existing structure extracted from the reference model. The extraction of this existing structure from the Reference Model begins by isolating a specific failure mode. The failure mode chosen is often guided by the available data where the mode was active. Isolating a single failure mode from the reference model and the monitors that indict the mode produces a tree structure where the class node describes the binary presence of that fault. The indicators (the leaves of the tree) have probabilities for the indictment of the mode and false alarm rates from the reference model that can be used to construct the complete probabilistic space. This structure and probabilistic information is the classical defini-

tion of a Naïve Bayes classifier. With data and the use of algorithms to build a Bayesian structure, the model can be leveraged to improve this very specific structure. Limiting it to a structure that preserves the general tree of the Naïve Bayes classifier eases the transition of information from the learned structure back to the original model. This is balanced with our desire to add limited causality into the network to help the expert understand if there are health indicators which are correlated. This information could be added back in limited ways to the isolated structure, without requiring the entire reference model to be changed from the Noisy-OR model. The structure that is chosen for learning is a Tree Augmented Naïve Bayesian network, which we will discuss in more detail in section 6.

The information added back to the reference model falls into three areas:

1. **Update Monitors** Update the threshold θ associated with a diagnostic monitor. The idea is to make the monitor i more sensitive to failure mode j (so that it can be detected earlier, if it is present) without sacrificing the false alarm rate, ϵ_j for the monitor.
2. **Add Monitors to Failure Mode** Update the isolated structure by adding a monitor that helps indict the failure mode. Specifically this could take two forms: (a) creating a new monitor with the requisite probabilistic information, and adding a new d_{ji} to associate it with the failure mode, and (b) assigning a non-zero number for d_{ji} if the link did not already exist with a previously created monitor.
3. **Create Super Monitors** Creating new monitors that combine existing monitors, say m_i and m_j such that the combination of monitor indictments asserts stronger evidence for a specific failure mode f_{m_k} . That is, calculate a stronger value for $P(m_i = 1, m_j = 1 | f_{m_k} = 1)$ which is greater than $P(m_i = 1 | f_{m_k} = 1) \times P(m_j = 1 | f_{m_k} = 1)$.

In addition to establishing areas of improvement found by comparing the data mining results with the reference model, the computational complexity of the data mining algorithms should be manageable, so that they can be used as exploratory analysis tools by the domain experts. Potentially, the experts may apply a successive refinement process by requesting a number of experimental runs, each using a specific data set from an operating fleet of aircraft, and the results from the n^{th} experiment augments or confirms the reference model from the $(n - 1)^{th}$ experiment. This will result in a continuous learning loop wherein historical observations from the fleet are analyzed systematically to understand the causal relations between failure modes and their manifestations (monitors). In addition, building models from the data may also reveal unknown (or currently unmodeled) dependencies among failure modes that are linked to the adverse event situations under consideration. Over time, this learning loop will increase the accuracy and time to detection (while reducing false positives) in the diagnostic reasoner.

The next step is to explore and pre-process the available aircraft flight data for the data mining task. The pre-processing plays a major role in determining the nature of information derived from the data, and, using prior

knowledge of the aircraft domain the pre-processing algorithms can be tailored to avoid the “needle-in-the-haystack” search problem.

4. AIRCRAFT FLIGHT DATA

It is important to extract flight data of the right type and form that will potentially help to find and validate new diagnostic information. Since the goal is early and reliable detection of an evolving fault in the aircraft, it is important that the data set formed for analysis span several contiguous flights. This set should also include multiple aircraft to account for aircraft-to-aircraft variations and the heterogeneity of flight conditions and flight paths. Our data set comes from a fleet of aircraft belonging to a regional airline from North America. The fleet consisted of 30+ identical four engine aircraft, each operating 2–5 flights each day. This work examines data spanning three years of the airline’s operations.

The Aircraft Condition Monitoring System (ACMS) is an airborne system that collects data to support fault analysis and maintenance. The Digital ACMS Recorder (DAR) records airplane information onto a magnetic tape (or optical) device that is external to the ACMS. This data is typically stored in raw, uncompressed form. The DAR can record information at a maximum rate of 512 12-bit words per second via a serial data stream modulated in either Harvard Bi-Phase or Bi-Polar Return-to-Zero code. The recorded data is then saved permanently to a compact flash card. The ACMS can be programmed to record parameter data from the propulsion subsystem, the airframe, the aircraft bleed subsystem, and the flight management system at a maximum rate of 16 Hz. We apply our initial data retrieval and pre-processing algorithms to this raw time-series data that was made available to us in the form of multiple CDs.

A second source of information we referenced for this study was adverse event annotations that is available in a FAA developed Aviation Safety Information Analysis and Sharing (ASIAS) database system. The ASIAS database is a collection of adverse events reported by various airline operators. On searching this database for the time period of the flight data available to us revealed that an engine shutdown event had occurred for one of the aircraft in our list. On one of the flights of this aircraft, the third engine (out of four) aboard the aircraft shutdown automatically. As a result, the flight crew declared an emergency situation and returned back to the airport where the flight originated. Fortunately, there were no casualties or serious injuries. For this study, we decided to focus on this failure mode, mainly because the on board reasoner or the mechanics who serviced the aircraft were unable to detect any anomalies in the engine till the adverse event occurred.

In more detail, an adverse event such as an engine shutdown typically evolves as a sequence of anomalous events and eventually leads to a situation, such as over heating, that causes the shutdown. For this case study, our objective was to analyze the ACMS flight data from the aircraft prior to adverse event with the goal of defining anomalies in the system monitors that were not defined for the existing ADMS. The primary intent was to use these anomalous monitor reports to detect and isolate the root cause for the failure as early as possible, so that the onset of the adverse event could be avoided.

Investigation of the airline maintenance crew reports after the adverse event revealed that the root cause for this adverse event was a faulty fuel metering hydro-mechanical unit (Fuel HMA) in the third engine, which was the engine that shut down. The fuel metering unit is a controller-actuator that meters fuel into the combustion chamber of the engine to produce the desired thrust. Given that we now knew the time of the adverse event and the root cause for the particular engine failure, knowledge of the fuel metering unit implied that this was a slowly evolving (i.e., *incipient*) fault that could very likely start manifesting about 50 flights *before* the actual engine shutdown adverse event. Therefore, we extracted the $[-50, 0]$ flight interval for the analysis, where 0 indicates the flight number for which the adverse event occurred and -50 indicates 50 flights before this one. We assumed that the particular aircraft under study had one faulty and three nominal engines. We collected relevant data (discussed below) for all of the engines, and then ran Bayesian classifiers to discover the differences between the faulty and the nominal engines.

As we discussed earlier, all of the aircraft for the regional airline were equipped with ADMS. The diagnoser receives information at pre-defined rates from the diagnostic monitors. In this case study, we also assume that we had access to the sequence of condition indicators (CI) values that were generated. As discussed earlier, the binary monitor output is produced by applying a pre-defined threshold to the CI values. In our data mining analysis, we use the CI's as features, and compare the thresholds derived by the classifier algorithms against the thresholds defined in the reference model. The following condition indicators and diagnostic monitors were available from this aircraft flight dataset.

StartTime This CI provides the time the engine takes to reach its idling speed. Appropriate threshold generates the *no start* diagnostic monitor.

IdleSpeed This CI provides the steady state idling speed. Appropriate threshold generates the *hung start* diagnostic monitor.

peakEGTC This CI provides the peak exhaust gas temperature within an engine start-stop cycle. Appropriate threshold generates the *overtemp* diagnostic monitor.

N2atPeak This CI provides the speed of the engine when the exhaust gas temperature achieves its peak value. Appropriate threshold generates the *overspeed* diagnostic monitor.

timeAtPeak This CI provides the dwell time when the exhaust gas temperature was at its peak value. Appropriate threshold generates the *overtemp* diagnostic monitor.

Liteoff This CI provides the time duration when the engine attained stoichiometry and auto-combustion. Appropriate threshold generates the *no lightoff* diagnostic monitor.

prelitEGTC This CI provides the engine combustion chamber temperature before the engine attained stoichiometry. Appropriate threshold generates the *hot start* diagnostic monitor.

phaseTWO This CI provides the time duration when the engine controller changed the fuel set-point

schedule. There are no diagnostic monitors defined for this CI.

tkoN1, tkoN2, tkoEGT, tkoT1, tkoPALT These CIs provide the fan speed, engine speed, exhaust gas temperature, inlet temperature and pressure altitude, respectively, averaged over the time interval when aircraft is operating under take off conditions. There are no diagnostic monitors defined for these CIs.

tkoMargin This CI provides the temperature margin for the engine during take off conditions. Appropriate threshold generates the *medium yellow* and *low red* diagnostic monitors.

Rolltime This CI provides the time duration of the engine's roll down phase. Appropriate threshold generates the *abrupt roll* diagnostic monitor.

resdTemp These CI provide the engine exhaust gas temperature at the end of the engine's roll down phase. Appropriate threshold generates the *high rtemp* diagnostic monitor.

N2atDip, dipEGTC These CIs provide the engine speed and the exhaust gas temperature at the halfway point in the engine's roll down phase. There are no diagnostic monitors defined for these CI.

N2cutoff These CI provide the rate of change of the engine speed at the halfway point in the engine's roll down phase. There are no diagnostic monitors defined for these CI.

This large volume of CI data (multiple aircraft, multiple flights) provides opportunities to study aircraft engines in different operating scenarios in great depth and detail. However, the data as extracted from the raw flight data DAR files was not in a form that could be directly processed by our classification algorithms. We had to develop data curation methods to generate the data sets that could be analyzed by the machine learning algorithms.

5. DATA CURATION

An important requirement for the success of data driven techniques for knowledge discovery is the need to have relevant and well-organized data. In our study, well-organized implied getting rid of unwanted details, being able to structure the data on a timeline (having all of the CI's aligned in time, and the monitor output inserted into the time line as a sequence of events), and applying filtering algorithms to the noisy sensor data. Relevance is an important concept, since it is necessary to extract sequences of data that contain information about the particular situation being modeled. For example, if the goal is to design a classifier that can identify a faulty situation from one in which there is no fault, it is important that the classifier be provided with both nominal and faulty data, so that it can derive the discriminating features from the data. Further, the systems under study are complex, and they operate in different modes and under different circumstances. This information is likely to be important for the classification task, so the data needs to be appropriately annotated with this information. It is clear that unreliable data is unlikely to provide useful information to an already effective reference model. Our (and others) experiences show that the data curation task is often

more time consuming and sometimes quite difficult (because of noisy, missing, and unorganized data) as compared to the data mining task, which involved running a classifier or a clustering algorithm on the curated data. A good understanding of the nature of the data and how it was acquired is critical to the success of the data mining task.

In our study, the DAR files represented single flights encoded in binary format. As a first step, we organized several thousands of these files by the aircraft tail number. For each aircraft, the data was then organized chronologically using the time stamp associated with the particular flight. Since the case study involves an engine shutdown situation, the data was further classified based on the engine serial number so that the data associated with each engine could be easily identified.

For practical reasons, given the size of the data, and the need to extract specific sub-sequences for the data mining task, we designed a relational database to create an organized representation for the formatted data. This made it easier to access the relevant data for different experimental analyses. For a data analysis session, step 1 involved formulating data base queries and collecting the extracted data segments into the form required by the classification algorithm. Step 2 was the data curation step. A primary task at this step was removing all extraneous non-flight data. For this analysis, all ground-test information (data generated when the maintenance crew ran a test when the aircraft was on the ground) was defined as anomalous and removed during the cleansing step. Step 3 involved running the classification algorithms.

6. TREE AUGMENTED NAÏVE BAYESIAN NETWORKS

The choice of the data driven techniques to apply to particular problems is very much a function of the nature of the data and the problem(s) to be addressed using the data. The extracted portion of the reference model discussed earlier can be modeled as a Naïve Bayes reasoner. The independence assumptions of the model may also be systematically relaxed to capture more discriminatory evidence for diagnosis. There are several interesting alternatives, but one that fits well with the isolated structure is the Tree Augmented Naïve Bayesian Method (Friedman et al., 1997) abbreviated as the TAN algorithm. The TAN network is a simple extension to the Naïve Bayes network formulation. The Root (the fault mode) also known as the class node is causally related to every evidence node. In addition, the independence assumption for evidence nodes is relaxed. An evidence node can have at most two parents: one is the class node, the other can be a causal connection to another evidence node. These constraints maintain the directed acyclic graph requirements and produce a more nuanced tree that captures additional relationships among the system sensors and monitors. At the same time, the learning algorithm to generate the parameters of this structure is computationally simpler than learning a general Bayes net structure.

The TAN Structure can be generated in several different ways. One approach uses a greedy search that constrains the graph from building “illegal” edges (i.e., a node having more than one parent from the evidence

nodes)(Cohen, Goldszmidt, Kelly, Symons, & Chase, 2004). Another procedure, sketched out in Algorithm 1, builds a Minimum Weighted Spanning Tree (MWST) of the evidence nodes and then connects the fault mode to all of the evidence nodes in the tree (Friedman et al., 1997). We use this algorithm in our work. A standard algorithm (e.g., Kruskal’s algorithm(Kruskal, 1956)) is applied to generate the MWST. The edge weight computation for the tree structure utilizes a log-likelihood criterion, such as the Bayesian likelihood value (Chickering, Heckerman, & Meek, 1997) and the Bayesian Information Criterion (BIC) (Schwarz, 1978). If the values are naturally discrete or they represent discretized continuous values, the Bayesian likelihood metric is preferred. This is a simple metric, which calculates the likelihood that two variables are dependent. The BIC is better suited for data sets whose features are derived from continuous distributions (like a Gaussian Normal). For either measure, the values are calculated for every pair of evidence nodes and stored in a matrix. Note that the value calculated for node i to node j is different for the value calculated for node j to node i . Therefore, the directed edges of the MWST represent the implied direction of causality, and the derived structure includes preferred causal directions (and not just correlational information).

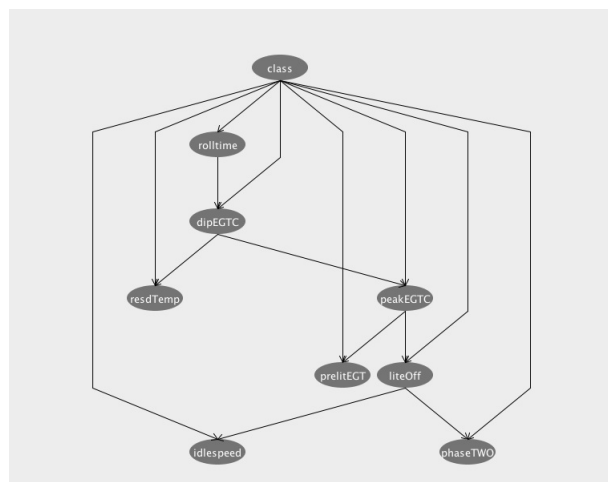


Figure 3: Example TAN Structure

An example TAN structure is illustrated in Figure 3. The root node, labeled class, is the fault hypothesis of interest. The other nodes represent evidence supporting the particular fault hypotheses. For the structure in Figure 3, rolltime, a monitor associated with the shutdown phase of the aircraft is the anchor evidence node in the TAN structure, called the *observation root node*. Like a Naïve Bayesian classifier, the fault hypothesis node (class) is linked to all of the relevant monitor nodes that support this hypothesis. Dependencies among some of the monitors, e.g., rolltime and dipEGTC, are captured as additional links in the Bayesian network. Note that the TAN represents a static structure; it does not explicitly capture temporal relations among the evidence. The choice of the observation root node is important; in some ways, it represents an important monitor for the fault hypothe-

sis, since it is directly linked to this node. This means the distribution used in the observation root node (whether it be a discrete CPT, or a continuous distribution) is conditioned only on the priors of the class distribution. The rest of the MWST structure is also linked to this node. All other conditional probability tables (CPTs) generated for this TAN structure include the class node and at most one other evidence node. The choice of the observation root node may determine the overall structure of the tree, but for the same data, the procedure for identifying the stronger causal links should not change in any significant way, i.e., the strong links will appear on all TANs, irrespective of the choice of the observation root node.

Algorithm 1 TAN Algorithm Using MWST

```

1: INPUT: Dataset D of N Features and a label C
2: INPUT: Observational Root Node FRoot
3: INPUT: CorrelationFunction
4: OUTPUT: TAN Structure with Adjacency Matrix,
   ClassAdjMat, describing the Structure
5: OUTPUT: Probability Values ProbVec for each
   Node {Note: Corr is a matrix of the likelihood
   that feature i is causally related to feature j (dif-
   ferent values can be found for i to j and j to i)}
   {Count(Node, ClassAdjMat, D) is a counting func-
   tion, that takes the Data, the Class, the Full Adj-
   acency Matrix of the TAN and for the Node finds
   either the CPT for discrete-valued features, or the
   set of means and covariances to describe the Gaus-
   sian Normal Distributions of the Node for continu-
   ous valued variables.} {AdjMat describes the par-
   ents so that correct data slices can be isolated and
   used in the counting.}
6: for featurei = 0 to featurei = N do
7:   for featurej = 0 to featurei = N do
8:     if featurei ≠ featurej then
9:       Corr(i, j) = CorrelationFunction(fi, fj, D)
10:    end if
11:   end for
12: end for
13: AdjMat = MWST(Corr, FRoot) { Build a Minimum
   Weighted Spanning Tree using the Correlation Mat-
   rix and the Root chosen}
14: for featurei = 0 to featurei = N do
15:   ClassAdjMat(featurei, C) = 1 {Connect ev-
   ery feature to the Class Node to build the TAN}
16: end for
17: ProbVec(C) = Count(C, ClassAdjMat, D) {Estimate
   the parameters, starting with the class}
18: for featurei = 0 to featurei = N do
19:   ProbVec(featurei) = Count(featurei, ClassAdjMat, D)
20: end for
21: RETURN: (AdjMat, ProbVec)

```

When the inference is used to assign a faulty or nominal label to the observed flight data, the result will be biased towards one class (fault) over another based on the CI value of the observation root node. This shift also changes some causal relationships and may impact how the counting algorithm for parameter estimation groups the data and produces probabilities for the evidence. In a later section we discuss how these choices can be used by the domain expert to make effective improvements to

the reference model for the AHM.

This choice of the observation root node, as shown in Algorithm 1 is an input parameter to the algorithm. This choice is normally based on a ranking computed using a heuristic, such as the highest BIC value. The algorithm in Weka (Hall et al., 2009) builds TANs with every feature as the root node of the MWST. It compares the generated structures, using a scoring metric such as the log-likelihood for the training data. The structure with the best score is then chosen as the classifier structure. Another approach could use domain knowledge to choose this node. For example, using expert knowledge of the system one may choose the sensor that is closest to the fault under consideration because it is not likely to be causally dependent on other sensors. The implication in the classifier is that it will be closest to indicating a fault.

Consider the example TAN shown in Figure 4. When the data for constructing the TAN is extracted from flights just before the adverse event occurred, the root node chosen by the Weka scheme is idlespeed. This node connects to the rest of the MWST, which in this case is the starttime feature, to which the rest of the feature nodes are connected. Using data from flights that were further away (before) from adverse event occurrence, the Weka algorithm picked PeakEGTC as the root node. This is illustrated in TAN structure in Figure 5. However, the derived causal link from idlespeed to starttime to a large group of nodes is retained at the bottom right of Figure 5. The similarities and shifts in the TAN structures from different segments of data typically informs the domain expert about the underlying phenomena due to the fault that is captured by the monitors. We discuss this in greater detail when we develop our case study in Section 7..

6.1 Implementations Used for Building TANs

Two different implementations can be employed for the TAN algorithms used in the experiments. The first is one that attempts to maintain the continuous nature of the features and build Gaussian Normal distributions for the nodes. It is implemented in MATLAB using the Bayesian Network Toolkit (Murphy, 2011).

The second method from the Weka (Hall et al., 2009) toolkit, uses a discretization algorithm which looks to bin each of the features into sets that unbalance the classes to provide the best possible split. For this case study, it produced more accurate classifiers, however, there were situations where it created a number of very fine splits in the feature values to define all of the class structures. The result was excessive binning, which produced very large conditional probability tables. When considering a more general view, methods that produce excessive binning are likely to be less robust to noise. Therefore, one has to consider these trade offs when choosing between these approaches.

7. EXPERIMENTS

To evaluate the data mining approach and demonstrate its ability to improve a diagnoser reference model, we conducted a case study using real flight data from the regional airline described earlier. We defined three standard metrics: (1) classification accuracy (2) false positive rate, and (3) false negative rate to systematically evaluate the TAN learning algorithm. Starting from the

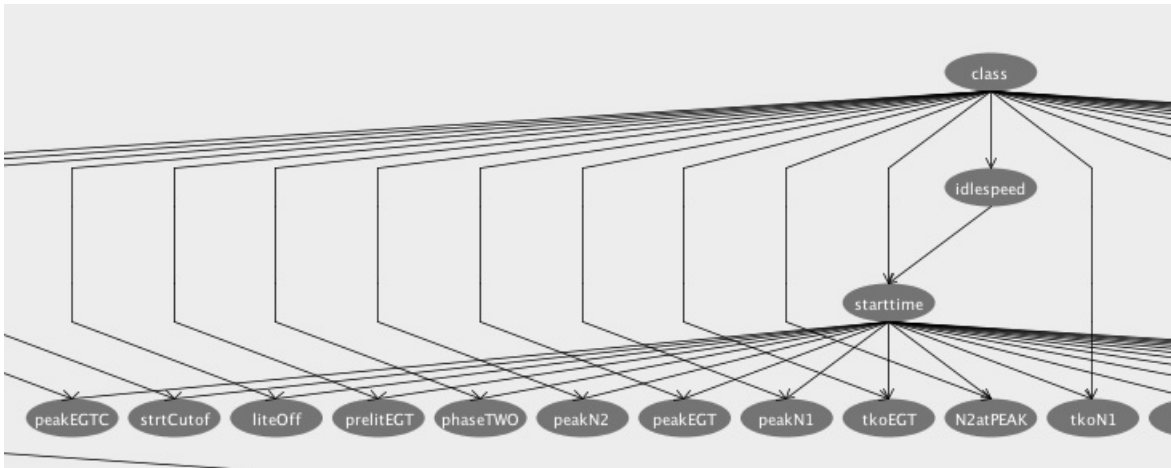


Figure 4: TAN Structure with idlespeed as observation root node

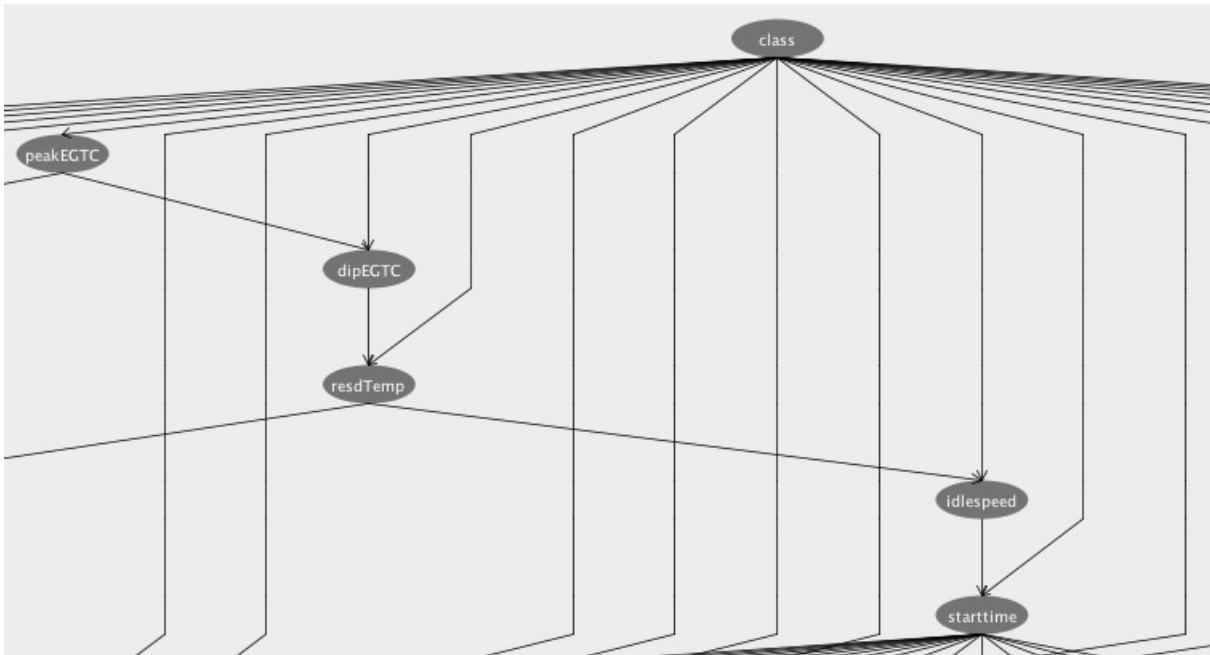


Figure 5: TAN Structure with peakEGTC as observational root node

flight in which the adverse event occurred for a particular aircraft, we used the earliest time to detection as another metric to evaluate the improvement in the system reference model after it has been updated with new information generated from the TAN classifier structure. The particular case study discussed here was an aircraft, where an overheated engine caused the engine to shut-down. This was considered to be a serious adverse event, and the pilots decided to return to the originating airport. By comparing the data from the faulty engine against the three other engines on the aircraft, which operated normally, starting from 50 flights before the adverse event, we were able to generate additional monitor information that reliably pointed to a FuelHMA problem (i.e., leak at the fuel meter) 40 flights before the actual incident. We break up the case study into three experiments and discuss their details next.

7.1 Experiment 1

The first task was to investigate the effectiveness of the generated classifier structures in isolating the fault condition using the condition indicator information derived from the flight data. We used condition indicators (CIs) rather than the health indicators (HIs) in this analysis because they make fewer assumptions about the nature of the data. We hypothesized that the original expert-supplied thresholds for the HIs were set at conservative values to minimize the chances for false alarms, and our derived classifier structures could potentially provide better thresholds without sacrificing the accuracy and false alarm metrics. This would lead to faster detection times.

From the ASIAs database, we extracted the aircraft and flight number in which the adverse event occurred. The report also indicated the nature of the fault that caused the adverse event, and knowledge of the fault provided context to our domain expert as to when this fault could be detected in the engine system. Our expert surmised that the fault would most likely start manifesting about 50 flights prior to the adverse event. The initial dataset that we then formulated consisted of the calculated CIs for all 50 of the identified flights. Each engine has its set of monitors, therefore, we had four sets of CIs, one for each engine. For analysis, we considered two ways for organizing this data. The first approach combined the four engine CI's as four separate features associated with one data point (i.e., one flight). Since we had 25 different CIs, this meant the dataset consisted of 50 data points, with each data point defined by 100 features. The second approach looked at each engine as a separate data point. Therefore, we formed four datasets, each with with 50 data points and 25 features. From the problem definition, it was clear that one of the four engines was faulty, and other three were most likely nominal for the 50 flights that we were analyzing. Therefore, we chose the latter approach for representing the data. To label the dataset appropriately for the classification study that we applied in this experiment, the three engines of the aircraft that showed no abnormalities (1, 2, and 4) were labeled as nominal, and the data points corresponding to engine 3, where the shutdown incident occurred, was labeled as faulty.

The classifiers were trained and evaluated using 10-Fold Cross validation (180 samples for training, and 20 for testing) with the nominal engine data being agnos-

tic of which engine produced the data. All of the CIs described in 4. were used as features in the classifier algorithm. The TAN generation algorithm from the Weka toolkit were used to derive the necessary classifiers. The fact that the discretized representation of the conditional probabilities were employed made it easier to find the threshold values for the diagnostic monitors that were linked to each CI. This is discussed in greater detail in Section 7.3. The derived TAN structure is illustrated in Figure 6.

The classification accuracy for this TAN structure was high, the average accuracy value was 99.5% with a .7% false positive rate and 0% false negative rate. These initial results were encouraging and to better understand them, the experiment was extended to confirm that the classifier results were attributed to the evolving fault in engine 3 and it was not just an artifact of the differences between the different engine characteristics. The above experiment was repeated with the training data including one of the nominal engines (1, 2, or 4) and the faulty engine, 3. The other two nominal engines were used as the test data. If the classifier split the remaining nominal engine data between the nominal and faulty classes derived, this would indicate that its structure more likely an artifact of engine placement on the aircraft. This experiment was repeated two more times, each time using a different nominal engine providing the training data and the other two being used as the test data. For all 3 experiments, the fault classification accuracy remained high, indicating that the classifier was truly differentiating between the fault and no-fault conditions.

7.2 Experiment 2

The positive results from the classification task led to the next step, where we worked with a domain expert to determine which of the CIs in the classifier provided the best discrimination between the faulty and nominal conditions. This information would provide the necessary pointers to update the current reference model. As a first step, the expert examined the TAN created using data from the 50 flight set used in Experiment 1. The expert's attention was drawn to the complex relationship between certain pairs of CI's during different phases of the flight:(1) rolltime and diPEGTC during the Shutdown phase, and (2) PeakEGTC and Starttime from the Startup phase. The expert concluded that there was a likely dependence between the shutdown phase of flight n and the startup of the next flight, $n + 1$. The reasoning was that an incomplete or inefficient shutdown in the previous flight created situations where the startup phase of the next flight was affected. The expert hypothesized that this cycle of degradation from previous shutdown to the next startup resulted in the fault effect becoming larger and larger, and eventually it would impact a number of CIs of the faulty engine.

This phenomena was investigated further by designing an experiment to track how the causal structure and accuracy of the classifiers derived from different segments of data. The different segments were chosen as intervals of flights before the flight with the adverse event occurrence as shown in Table 1. The 50 flights were divided into 5 bins of 10 flights each. A test set was constructed from the remaining 40 flights (data with nominal and faulty labels) as well as the samples of CIs from engine 3 after it had been repaired (after engine 3 was repaired,

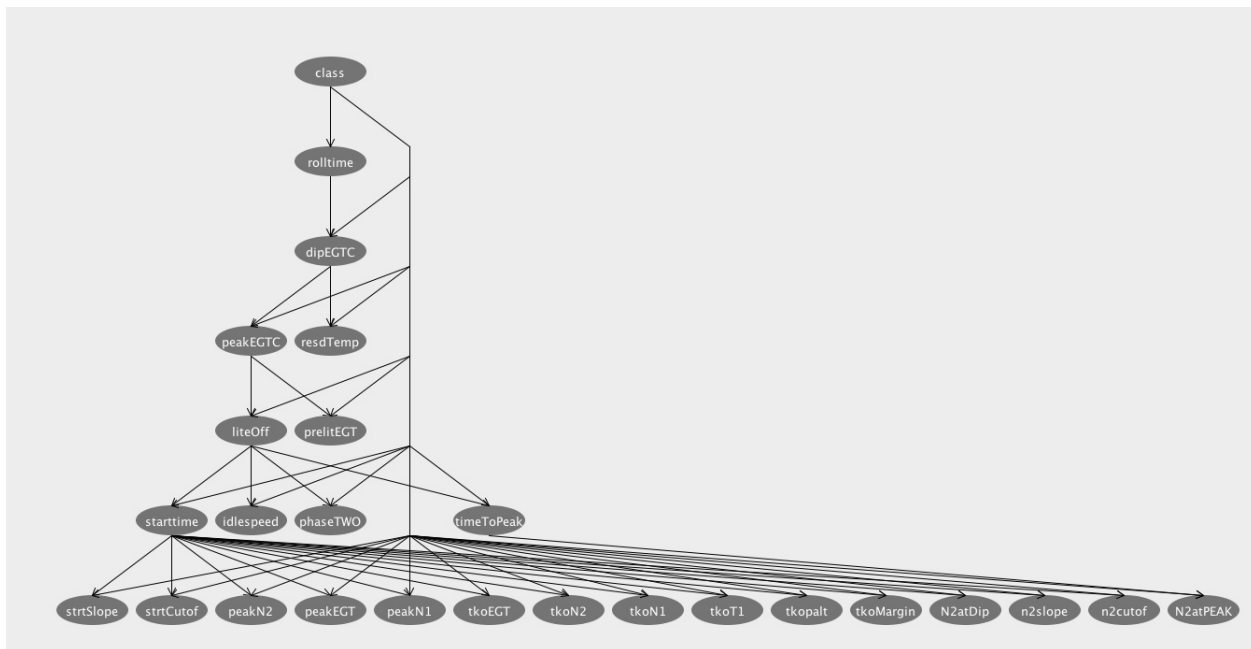


Figure 6: TAN Structure Generated using Data from all 50 Flights

Bin	Training Flights	Acc.on Holdout Set	FP%	Obs. Root Node	Children of ORN	Notes
1	1 to 10	97.65%	2.30%	IdleSpeed	StartTime	Thresholds Chosen from this Bin due to low FP
2	11 to 20	93.90%	5.70%	peakEGTC	liteOff,dipEGTC	peakEGTC Important Node
3	21 to 30	94.65%	5.30%	peakEGTC	liteOff,dipEGTC	peakEGTC Important Node
4	31 to 40	96.62%	3.50%	startTime	peakEGTC	Links startTime and PeakEGTC
5	41 to 50	96.06%	4.10%	liteOff	phaseTwo,RollTime	Links Startup and Rolldown CI

Table 1: Accuracy, False Positive Rate, Observational Root Node and Immediate Child Node for Classifiers Created from different data segments

no anomaly reports were generated by mechanics, and no further adverse events were reported for engine 3 in the ASIAs database, therefore, it was considered to be nominal). Table 1 shows the accuracy and false positive rate (FP%) metrics reported for the five experiments. The observation root node, and its immediate child in the generated TAN structures are also shown.

The conventional wisdom was that the accuracy and false positive metrics would have the best values for the classifiers generated from data close to the adverse event, and performance would deteriorate for the TAN structures derived from bins that were further away from the incident. The actual results show partial agreement. The bin 1 experiment produced the highest accuracy and lowest false positive rate, but the next best result is produced for TAN classifiers generated from the bin 4 data. This prompted the domain expert to study the bin 1 to bin 4 TANs more closely. The expert concluded that two CIs, *startTime* and *peakEGTC* showed a strong causal connection for bin 4, and *startTime* was highly ranked for the bin 1 TAN. On the other hand, *PeakEGTC* was the root node for bins 2 and 3. This study led the domain expert to believe that a new monitor that combined *startTime* and *peakEGTC* would produce a reference model with better detection and isolation capabilities. The process of designing and testing the diagnoser with the new monitor is described as Experiment 3.

7.3 Experiment 3

Working closely with the data mining researchers, the domain expert used the framework in the problem statement to reconcile the results from Experiment 2 to suggest explicit changes in the reference model; this included: (1) updates to the threshold values that specified the diagnostic monitors (i.e., updated HIs from the CIs), (2) a new monitor that could be added to the current reference model and (3) the addition of a “Super Monitor” to the reference model.

The CPTs generated by the learned classifier were defined as discretized bins with split points that could be interpreted as thresholds. Looking at the bins, the lowest false positive rate occurred in bin 1. For the observation root node, the thresholds were updated using the results for bin 1 by comparing the split values with the original thresholds. For the remaining nodes, their causality with respect to the observation parent was removed by marginalizing to remove that particular variable. Once marginalization is applied, the CPT lists the probability values at the nominal versus faulty split points. The domain expert studied these split values and made decisions on whether the new split values should update the thresholds in the reference model. The trade off was to improve the accuracy of fault detection without introducing too much noise (uncertainty) into the decision process.

Studying the TAN structures provided additional information to the domain expert. When the *slowStart* HI fired, the expert discovered that this was not because the *startTime* during start up was slow; sometimes the fault occurred when the *startTime* was too fast. This implied a new HI could be added to the failure mode that now examines if *startTime* is under a threshold and too fast. The addition of this HI (called *fastStart*) to the reference model would be to speed up detection by adding new evidence to indict the fault.

Experiment 2 also showed a causal relationship appearing between *startTime* and *peakEGTC*. The domain expert suggested adding this as a “super monitor”. This new HI would combine information from the *fastStart* HI and the *HighTemp* HI to identify the *fuelHMA* fault in the reference model. In other words, if both monitors fired, then this new monitor would also fire directly implicating the fault hypothesis. In other words, joint occurrence of these two monitors provides stronger evidence of the fault than if one considers the effect of the two monitors individually. For example, in the original structure that showed a possible relationship between monitors in flight *N* and flight *N+1*, the causality might cause this new monitor to fire only when the two HI involved fire in that explicit sequence, flight *n* and flight *n + 1*. Not only does this super monitor combine the results from other monitors, but it also indicates cyclic behaviors that again provide very useful diagnostic information. In general, these “super monitors” could model complex interactions thus increasing the overall discriminability properties of the reasoner. The consequence of using a super monitor, is that the usefulness of the two monitors used in the construction are lost. These are removed from the links to the failure mode being examined (however they remain for any other failure mode). In this situation, the just created monitor for fast startTimes would be removed as well as the *HighTemp* HI in place of a new super monitor for a fast start and a high engine temperature on start up.

To show that this new super monitor and the updated thresholds produce better results, multiple traces of the monitor output were made for the 50 flight data set. This included 10 nominal flights after the problem was caught and corrected. The first trace is only a recording of the original monitors designed by the expert. The second trace includes the new monitors (both the *fastStart* monitor and “super monitor”) derived by the data mining analyses, as well the updated information (thresholds). Run separately, they can be analyzed to determine if the reasoner finds the fault sooner in the trace and indicates that maintenance is more than likely needed for the aircraft.

These results from the reasoner simulations are shown in Figures 7 and 8. The traces illustrate the reasoner’s inferences at different flight numbers before the actual incident occurrence. This analysis demonstrates how far before the adverse event the reasoner would reliably detect the fault, and potentially generate a report that would lead to preventive maintenance, and, therefore, avoidance of the adverse event. With the original reference model the reasoner was unable to disambiguate between three potential fault candidates at any point leading up to the event. All of the fault candidate hypotheses required more evidence to support the isolation task. This would not avoid the unfortunate shutdown and the emergency return to the originating airport. Figure 8 shows the reasoner trace for the new reference model. Using the updated thresholds and the new super monitor (which is derived from a new monitor itself and one original monitor) suggested by the data mining algorithms led to a correct isolation of the fault, i.e., the *fuelHMA* problem. In this case, the reasoner originally hypothesized five fault conditions: four of these were linked to the faulty engine and one was a vehicle level hypothesis. As further monitor information became available, fuel metering re-

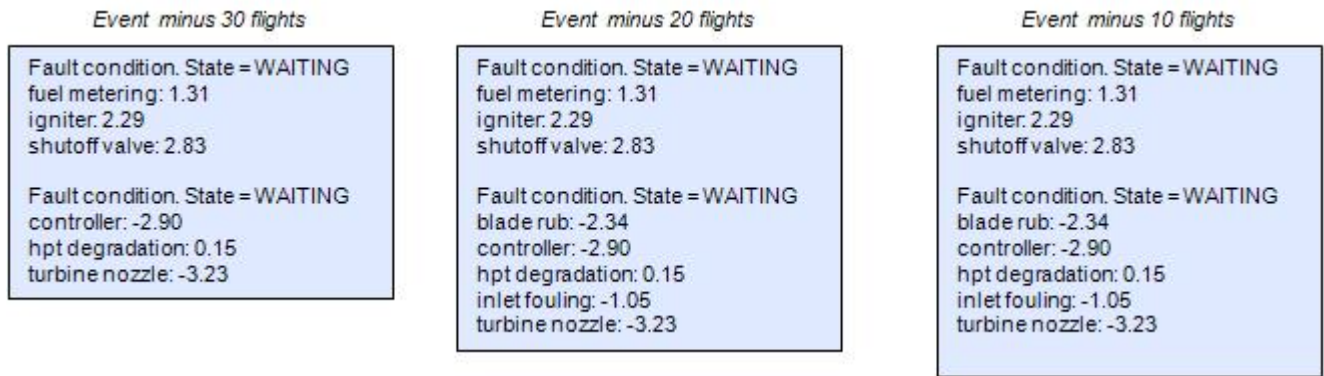


Figure 7: Trace of the Reasoner on the Original Reference Model

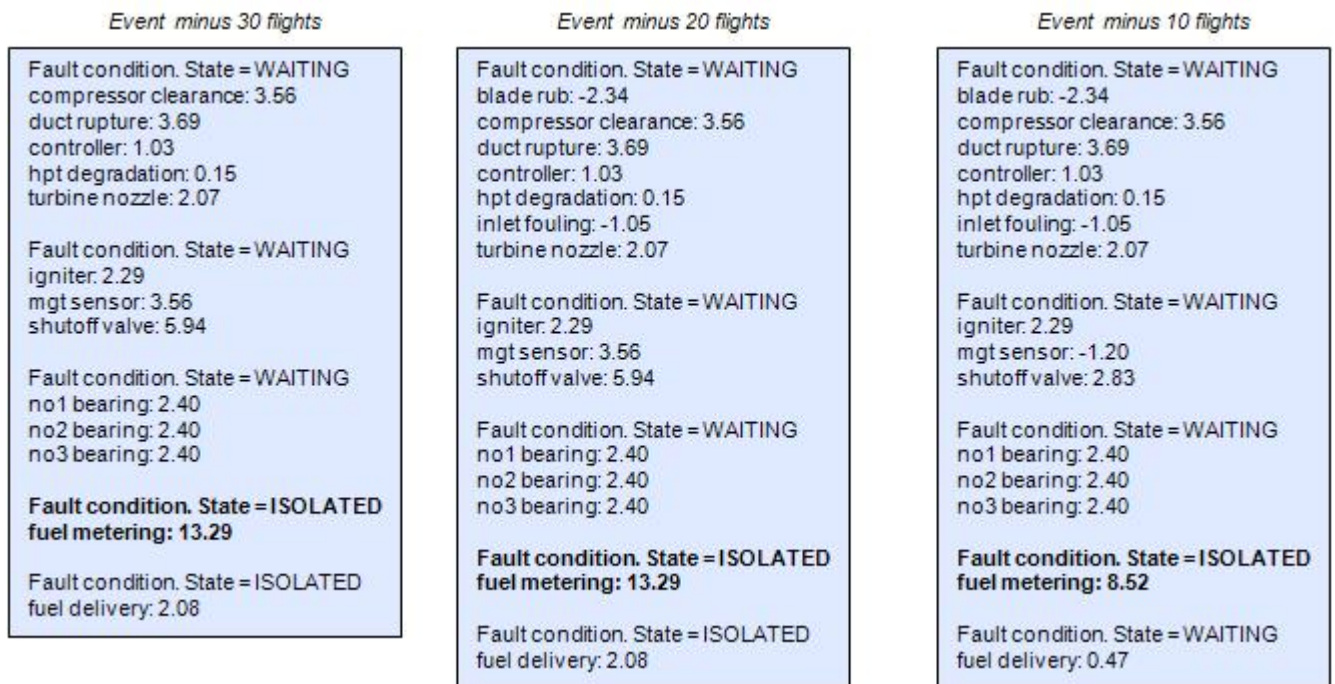


Figure 8: Trace of the Reasoner with the improved Reference Model

mained the only plausible candidate, and the fault hypothesis was established unambiguously. The fact that this isolation by the reasoner occurred 30 flights before the incident is significant because it gave sufficient advanced warning to the maintenance crews to fix the problem before an emergency situation developed.

This case study provides encouraging results in the area of diagnoser model improvement through data mining. It indicates that it may be possible to uncover new information about the relationship between components on a vehicle and how they can be harnessed to improve diagnostic reasoning. Not only can it help isolate faults, but also potentially catch them earlier in the cycle. These three experiments provide a general direction to assisting a domain expert in improving their work, and giving them access to new or missing information.

8. CONCLUSIONS AND FUTURE WORK

The overall results from this case study generated positive results and show the promise of the data mining methodology and the overall process that starts from data curation and ends with systematic updates to and verification of the system reference model. The results presented clearly demonstrate that the data mining approach is successful in: (1) discovering new causal relations in the reference model, (2) updating monitor thresholds, and (3) discovering new monitors that provide additional discriminatory evidence for fault detection and isolation. Experiment 3 demonstrated that the new knowledge leads to better diagnoser performance in terms: (1) early detection, and (2) better discriminability. An immediate next step will be to generalize this methodology by applying it to other adverse event situations. In the longer term, to further validate this work, we plan to advance this research in a number of different directions.

- Validation of the approach and classifier structures generated by looking at additional engine data sets from other flight data that report the same and related adverse events. To establish the robustness of the work, it is important to extend the analysis to looking at multiple occurrences of the same adverse event, and to compare the thresholds, relations, and monitor structures generated by the extended data analysis.
- Extension of the analysis methodology beyond single systems and subsystems. A rich source of information about fault effects involves looking at the interactions between subsystems, especially after fault occurrence begins to manifest. Of particular interest is looking at cascades of monitors and cascades of faults. In this framework, studying the response of the avionics systems under different fault conditions would be very useful.
- Advance our data mining techniques to extract causal relations between avionics and other subsystems, as well as study correlations between the combined avionics and engine features and adverse vehicle events, such as in-flight engine shutdowns and bird strikes. Understanding what features of a flight differentiate situations when a pilot starts compensating for what may be a slowly degrading component that originates from a bird strike will help us gain a better understanding of how to

monitor the actual operation of the aircraft with its subsystems under various conditions. This also presents interesting problems from the application and development of machine learning algorithms to utilize in this data mining problem.

ACKNOWLEDGMENTS

The Honeywell and Vanderbilt researchers were partially supported by the National Aeronautics and Space Administration under contract NNL09AA08B. We would like to acknowledge the support from Eric Cooper from NASA; Joel Bock and Onder Uluyol at Honeywell for help with parsing and decoding the aircraft raw data.

REFERENCES

- Cheng, J., Greiner, R., Kelly, J., Bell, D., & Liu, W. (2002). Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence*, 137(1-2), 43 - 90.
- Chickering, D. M., Heckerman, D., & Meek, C. (1997). A Bayesian approach to learning Bayesian networks with local structure. In *In Proceedings of Thirteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann.
- Cohen, I., Goldszmidt, M., Kelly, T., Symons, J., & Chase, J. S. (2004). Correlating instrumentation data to system states: a building block for automated diagnosis and control. In *Proceedings of the 6th conference on Symposium on Operating Systems Design & Implementation - Volume 6* (pp. 16–16). Berkeley, CA, USA: USENIX Association.
- Dearden, R., & Clancy, D. (2001). Particle filters for real-time fault detection in planetary rovers. In *Proc. of the 12th International Workshop on Principles of Diagnosis* (p. 1-6).
- Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian Network Classifiers. *Machine Learning*, 29, 131–163.
- Grossman, D., & Domingos, P. (2004). Learning Bayesian network classifiers by maximizing conditional likelihood. In *Proceedings of the twenty-first international conference on Machine learning* (pp. 46–). New York, NY, USA: ACM.
- Hall, M., Eibe, F., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1), pp. 10-18.
- Kruskal, J., Joseph B. (1956). On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proceedings of the American Mathematical Society*, 7(1), pp. 48-50.
- Lerner, U., Parr, R., Koller, D., & Biswas, G. (2000). Bayesian Fault Detection and Diagnosis in Dynamic Systems. In *Proc. of the 17th Nat. Conf. on Artificial Intelligence* (p. 531-537). San Mateo, CA, USA: AAAI Press.
- Murphy, K. (2011). *Bayesian Net Toolbox @ONLINE*. Available from <http://code.google.com/p/bnt/>
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc.

- Roychoudhury, I., Biswas, G., & Koutsoukos, X. (2008). Comprehensive diagnosis of continuous systems using dynamic Bayes nets. In *Proc. of the 19th International Workshop on Principles of Diagnosis* (p. 151-158).
- Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6.
- Smyth, P. (1994). Hidden Markov models for fault detection in dynamic systems. *Pattern Recognition*, 27(1), pp. 149-164.
- Spitzer, C. (2007). Honeywell Primus Epic Aircraft Diagnostic and Maintenance System. *Digital Avionics Handbook*(2), pp. 22-23.
- Verma, V., Gordon, G., Simmons, R., & Thrun, S. (2004). Real-time fault diagnosis. *IEEE Robotics and Automation Magazine*, 11(2), pp. 56-66.

Design for Fault Analysis Using Multi-partite, Multi-attribute Betweenness Centrality Measures

Tsai-Ching Lu¹, Yilu Zhang², David L. Allen¹, and Mutasim A. Salman²

¹*HRL Laboratories LLC, Malibu, California, 91360, U.S.A*

*tlu@hrl.com
dlallen@hrl.com*

²*General Motors, Warren, Michigan, 48090, U.S.A*

*yilu.zhang@gm.com
mutasim.a.salman@gm.com*

ABSTRACT

As electrical and electronic systems (EES) steadfastly increase their functional complexity and connectedness, they pose ever-growing challenges in fault analysis and prevention. Many EES faults are intermittent, emerging (new faults), or cascading, and cannot be addressed by the traditional component-level diagnostic design. Leveraging the latest advancements in Network Science, we take the holistic approach to model and analyze the highly interrelated in-vehicle EES as layered sub-networks of hardware components, software components, and communication links. We develop multi-partite, multi-attribute betweenness centrality measures to quantify the complexity and maintainability of the layered EES network. We then use the betweenness centrality distribution to identify fault analysis monitoring points and fault-mitigation strategies. The promising results obtained by our initial empirical study of an example in-vehicle EES presents a first step toward network-theory based IVHM.

1. INTRODUCTION

The complexity of the electrical and electronic system (EES) in vehicles has evolved over the years in response to continuously increasing demand for incorporating new electronic control units (ECUs) onto vehicles. These allow for advanced safety, convenient and comfort features, as well as meeting new emission and fuel-economy standards. However, the fast growing number of ECUs and their peripherals has led to complex interactions which can lead to unexpected emerging or cascading failures.

Current state-of-the-art diagnosis and prognosis algorithms typically focus on one aspect of the system which makes it

difficult to capture problems originating from the interaction between and across different system layers: physical level (power or communication), functional level and communication level. Such multi-layer problems are typically addressed after the fact with tedious and error prone manual analysis.

In this paper, we consider in-vehicle EES as an embedded and distributed complex system, subject to the design for fault detection, isolation, and mitigation. Based on recent advancements in Network Science, we develop the layered EES network modeling methodology to capture highly inter-related in-vehicle EES. We develop novel multi-partite and multi-attribute betweenness centrality measures to quantify the importance to which a node has control over pair-wise connections between other nodes in the layered EES network model. We apply multi-partite and multi-attribute betweenness centrality measures to rank and recommend fault detection and isolation monitoring points that cannot be discovered by single layered analysis techniques and conventional betweenness centrality measures. We provide usage-based and random failure simulation strategies for recommending fault isolation and mitigations points for desired diagnostic coverage. We present our initial empirical study toward this network-based approach of IVHM.

We discuss related work in Section 2 and introduce our layered network modeling methodology in Section 3. In Sections 4-6, we describe our multi-partite and multi-attribute betweenness centrality, and their application to fault analysis monitoring. Section 7 provides an example study. We conclude our papers with future research direction in Section 8.

2. RELATED RESEARCH

Our work is related to embedded system, complex system diagnosis, and network science. Struss et. al. (2010)

Lu et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

compiled a special issue on the recent advancements of model-based diagnosis in which (Wang & Provan, 2010) describes the automated benchmark diagnostic model generator, with various domain, topology and system-level behaviors, based on the graphical model approach of network science. The benchmark models generated in (Wang & Provan, 2010) can be provided as the input to our methodology for fault detection, isolation and mitigation analysis.

Simonot-Lion (2009) is another special issue compiling recent advancements in the area of in-vehicle embedded system. Zeng et al., (2009) describes a stochastic analysis framework for the end-to-end latency of distributed real-time systems and demonstrated the experimental results on Controller Area Network (CAN). This work focuses on simulation and analysis of probability distribution for end-to-end latency analysis of active safety functions on vehicles. Our work, on the other hand, focuses on design and diagnosis.

Our proposed new measures for quantifying EES complexity and maintainability is based on betweenness centrality measures in network science. Brandes (2008) gives a comprehensive survey and contrasts most recent variants of betweenness centrality. Our proposed new measures are inspired from our layered EES network; therefore, there is no compatible measures in the state-of-the-art as surveyed in (Brandes, 2008). The measures closest to ours are those described in (Borgatti, 2005; Flom et. al., 2004). However, their works do not consider multi-partite, multi-attributes layered networks. In general, these works focus on social network analysis and has no mentioning of fault-isolation and fault-mitigation analysis.

3. LAYERED NETWORK MODELING

By taking the holistic approach to model in-vehicle EES, we make the following modeling assumptions to construct the layered, multi-partite, multi-attribute network for analyzing an EES system.

1. Each network layer models one aspect of EES; for example, physical network layer represents physical wiring connections of ECUs, functional network layer represents relations of software functions among ECUs, message network layers models message flows among ECUs, and so on.
2. Nodes can be annotated with node types. Designation of node type leads to partitions of nodes where nodes in the same partitions do not have edges; for example, one ECU node is not directly linked to another ECU node, but via Message nodes in a message network layer.
3. Nodes can be annotated with node attributes to represent their special characteristics. Node attributes are usually defined orthogonally to node types; nodes with the same node type may have different node

attributes, and similarly nodes with different types may have the same node attribute. For example, node attributes {Sending, Receiving} can be used to annotate nodes across node types {ECU, Message}.

4. Edges within the layered network can be annotated with edge attributes where the value of an attribute typically represents the types of information flowing between nodes. For example, a feature node may have an edge to another feature node with edge attributes {data, frequency} in the dataflow network.
5. Edges across different layers typically represent dependency or identity relations. For example, the same hardware ECU node may appear in both electrical and physical sub-networks which warrant across layer edges.

Formally, we consider a graph $G=(N,E)$ consists of a nonempty countable set of nodes N and a set of directed or undirected edges $E \subseteq N \times N$. A multipartite graph is a graph where N is divided into nonempty disjoint subsets (called Parts) and no two nodes in the same subset have an edge connecting them. Nodes can be associated with a vector of node attributes N_A ; similarly, edges can be associated with a vector of edge attributes E_A . Part is imposed by topological structure, whereas attribute is primarily augmented for the semantic aspect of a node. A layered, multi-partite, multi-attribute EES network consists of layers of multi-partite, multi-attribute graphs where node types correspond to parts, and edges across layers represent dependency or identity relations for entities in different layers. Figure 1 shows an example layered network of in-vehicle EES layered network.

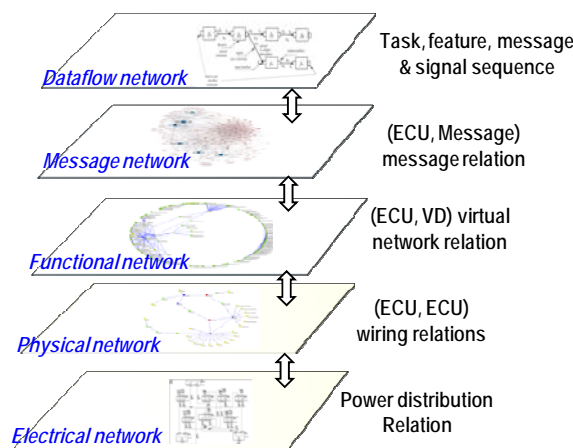


Figure 1: An example layered EES network consists of layers of electrical, physical, functional, message, and dataflow sub-networks; relation within each layers are shown to the right; dependency and identify relations across layers are summarized into double arrows across layers of sub-networks. Note that across layer links are not restricted to neighboring layers only.

4. BETWEENNESS CENTRALITY

Betweenness centrality is defined in social network analysis to quantify the importance to which a node has control over pair-wise connections between other nodes, based on the assumption that the importance of connections is equally divided among all shortest paths for each pair (Freeman, 1978). The *betweenness centrality* $BC(n_i)$ for a node $i \in N$ is defined as follows.

$$BC(i) = \sum_{h \neq i \neq j} \frac{\sigma_{hj}(i)}{\sigma_{hj}},$$

where σ_{hj} is the total number of shortest paths between h and j , and $\sigma_{hj}(i)$ is the number of such shortest paths that pass through the node i . The $BC(i)$ can be scaled between 0 and 1 using $\frac{BC(i)}{(|N|-1)}$ where $|N|$ is the number of nodes in the graph. Correspondingly, the *betweenness centrality* $BC(e)$ for an edge $e \in E$ is defined as the number of shortest paths passing through the edge, i.e., $BC(e) = \sum_{h \neq i \neq j} \frac{\sigma_{hj}(e)}{\sigma_{hj}}$. The $BC(e)$ could be normalized between 0 and 1 using $\frac{BC(e)}{[(|N|-1)(|N|-2)]/2}$.

Recognizing the rich semantics in the layered EES network, we develop novel multi-partite, multi-attribute betweenness centrality to account for node types and attributes in the layered in-vehicle EES network.

4.1. Multi-partite Betweenness Centrality

In the layered EES network, each node and edge can have different types and attributes which warrant further constraints on how betweenness centrality can be defined when considering different semantic meaning of shortest paths in the layered EES network. We propose three different multipartite betweenness centrality measures based on the constraints on node types (parts) in the network.

We first define the *homogeneous* multipartite betweenness centrality $BC_P(i)$ for a node $i \in N_P$, where N_P is a part $N_P \subset N$, is defined as follows:

$$BC_P(i) = \sum_{h \neq i \neq j \in N_P} \frac{\sigma_{hj}(i)}{\sigma_{hj}},$$

where σ_{hj} is the total number of shortest paths between h and j , given that nodes h, i , and j are all in the same part N_P , and $\sigma_{hj}(i)$ is the number of such shortest paths that pass through the node i . This is to constrain the shortest paths such that the starting and ending nodes are the same node types (in the same part) as the one of the intermediate node. For example, an ECU node linked to another ECU node via a gateway ECU with some message nodes along the path.

Next, we define the *bi-mode* multipartite betweenness centrality where the starting and ending nodes are the same part but different from the part of the intermediate node. The

bi-mode multipartite betweenness centrality $BC_{\bar{P}}(i)$ for a node i is:

$$BC_{\bar{P}}(i) = \sum_{h, j \in N_Q \neq N_P, i \in N_P} \frac{\sigma_{hj}(i)}{\sigma_{hj}},$$

where σ_{hj} is the total number of shortest paths between h and j that are in the same part, but are different from the part of the node i , and $\sigma_{hj}(i)$ is the number of such shortest paths that pass through i . One example use of this measure is to consider a message node i sitting on the paths of communications between two different ECU nodes.

We define the *heterogeneous* multipartite betweenness centrality for a node $BC_{\bar{P}}(i)$ for a node i as follows:

$$BC_{\bar{P}}(i) = \sum_{h \in N_o, i \in N_P, j \in N_Q} \frac{\sigma_{hj}(i)}{\sigma_{hj}},$$

where σ_{hj} is the total number of shortest paths between h and j that are in different parts and not in the same part as the node i ($N_o \neq N_P \neq N_Q$), and $\sigma_{hj}(i)$ is the number of such shortest paths that pass through i . This measure assumes that there are at least three parts defined in the network. One example use of such measure could be finding out the betweenness for a node in functional layer and starting and ending nodes are in the layers of message and physical networks.

4.2. Multi-attribute Betweenness Centrality

To account for attributes orthogonal to topological definition of parts, we define *homogeneous* multi-attribute betweenness centrality $BC_A(i, a)$ and *negated* multi-attribute betweenness centrality $BC_{\bar{A}}(i, a)$ for a node $i \in N$ and an attribute $a \in A_N$ as follows:

$$BC_A(i, a) = \sum_{h \neq i \neq j, a(h)=a(i)=a(j)} \frac{\sigma_{hj}(i)}{\sigma_{hj}},$$

where σ_{hj} is the total number of shortest paths between h and j , given that nodes h, i , and j has the same values for the attribute a (i.e., $a(h) = a(i) = a(j)$), and $\sigma_{hj}(i)$ is the number of such shortest paths that pass through i ; and

$$BC_{\bar{A}}(i, a) = \sum_{h \neq i \neq j, a(h)=a(j), a(i) \neq a(j)} \frac{\sigma_{hj}(i)}{\sigma_{hj}},$$

where σ_{hj} is the total number of shortest paths between h and j , given that nodes h and j has the same values for the attribute a (i.e., $a(h) = a(j)$) but they have different values from node i (i.e., $a(i) \neq a(j)$), and $\sigma_{hj}(i)$ is the number of such shortest paths that pass through i .

Similarly, the multi-attribute betweenness centrality, $BC_A(e, a)$ and $BC_{\bar{A}}(e, a)$ for an edge $e \in E$ and an attribute $a \in A_E$, can be defined as those for the nodes. One example

use of multi-attribute betweenness centrality is the attributes of an ECU such as the “role” which can have the attribute value “receiving” or “sending” for different messages.

4.3. Betweenness Centrality Distribution

In addition to quantifying the importance of a target (a node or an edge) in the network, we can compute the betweenness centrality for every target in the network to derive a distribution of betweenness centrality. We can then compute descriptive statistics (e.g., average, percentile, variance, skewness, etc.) to characterize such betweenness centrality distribution. In Figure 2, we show an example of truncated homogeneous betweenness centrality distribution for a functional network.

The betweenness centrality distribution can be used to quantify the complexity, as well as maintainability of a layered EES system. For example, a centralized design of EES may have a more skewed betweenness centrality distribution than the one with distributed design. In Figure 2, we see that $69/584=11.81\%$ nodes have above average betweenness centrality, which give us a quite skewed distribution from the functional network point of view.

To improve system maintainability, more resources can potentially be put into the system to improve the reliability of the targets with high betweenness centrality metric, or to increase diagnostic coverage for targets with low betweenness centrality metric. The betweenness centrality distribution can enable such a trade-off analysis for improving the design of maintainability.

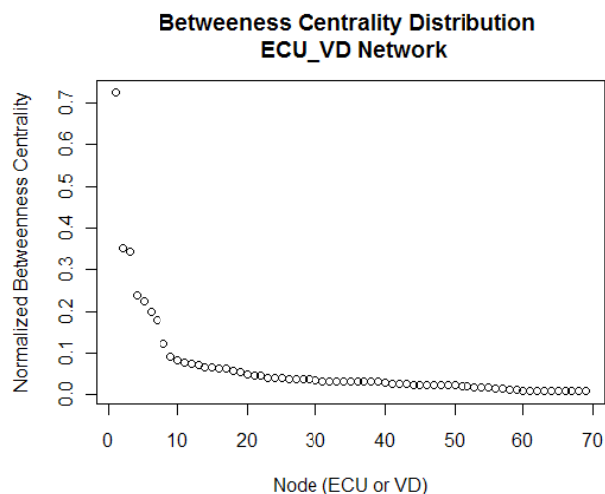


Figure 2: An example distribution of betweenness centrality: the top 69 nodes (out of 584 nodes) have above average betweenness centrality in a functional network.

5. FAULT ANALYSIS MONITORING POINTS

Fault-detection and fault-isolation requires actively monitoring a system in operation. The layered EES network models diverse aspects of the system and the operational status can be thought of as signals and information flow over the network. One may consider using the network model as a platform to simulate the operations of EES; however, it is unlikely to simulate all possible combinations of inputs, especially in the wide ranges of different and unforeseeable operational environments.

Betweenness centrality, as a measure of quantifying the node importance, provides a good basis for ranking where to include fault analysis monitoring points, assuming due to resource constraints not all parts can be monitored. In Figure 3, we show an example network to illustrate this point. The upper panel shows that node *G* has four immediate neighboring nodes whereas node *I* has only three immediate neighbors; however, node *I* is more important than node *G* with respect to betweenness centrality measure ($BC(I)=14 > BC(G)=0.67$). The bottom left panel shows that if node *G* fails, node *I* can still monitor all traffics on the network; however, if node *I* fails, the network is fragmented as shown in the bottom right panel. This warrants the claim that the high betweenness centrality node serves as a better fault monitoring point.

We propose to use the betweenness centrality distribution, in conjunction with the measure of degree neighbors, as the basis for setting up monitoring points for fault detection and isolation with respect to desired fault coverage. The following steps can generate recommended fault analysis monitoring points:

1. Compute betweenness measures (multi-partite, multi-attribute betweenness centrality) to quantify the importance of the nodes in EES;

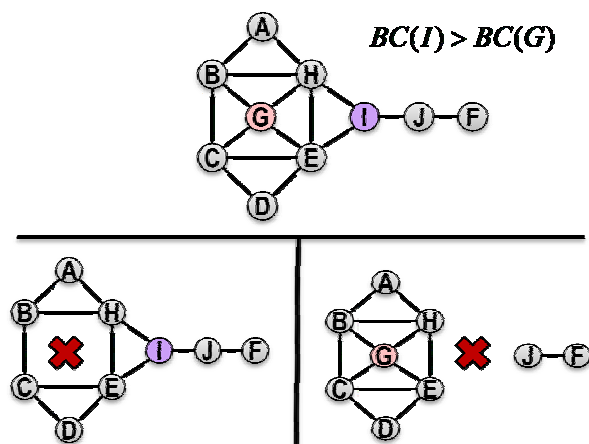


Figure 3: An example to illustrate the usefulness of betweenness centrality for fault analysis monitoring.

2. Apply an adjustable threshold to select nodes as candidate monitoring points (e.g., select nodes with BC measure above x -percentile of BC distribution).
3. Check whether the degree neighbors (e.g. 2nd degree neighbor) of all selected nodes provide the desired coverage of the whole network;
4. If no, go back to Step 2 and adjust the threshold;
5. If yes, recommend the selected nodes as the monitoring points.
4. Check whether the sequential failure simulation has reached the completion criterion (e.g., stop simulation when all edges are removed from the network; or when a certain percentage of survival nodes is remained in the network).
5. If no, go to Step 1;
6. If yes, output the effects of sequential failure for mitigation analysis.

Diagnostic coverage for a given monitoring point is computed via the nodes' degree neighbors, which is system dependent and subject to the observability of failure mechanisms. Figure 4 shows an example of diagnostic coverage for node I as the monitoring point.

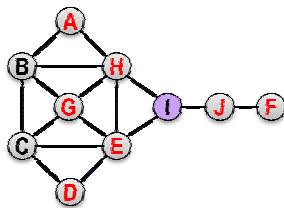


Figure 4: Example diagnostic coverage with a monitoring point at node I and coverage of 2nd degree neighbors (nodes in red fonts).

The diagnostic coverage of the whole network is the union of the diagnostic coverage of all selected monitoring points. It is advised to trade-off between cost and diagnostic coverage in selecting monitoring points for a given system.

6. FAULT MITIGATION ANALYSIS

The purpose of fault-mitigation analysis is (1) to quantify how robust the EES is with respect to different failures, and (2) to identify which surviving nodes can potentially take over the functionality of nodes which have failed.

To support fault-mitigation analysis, we introduce two sequential failure strategies: random failure and usage-based failure strategies to simulate the effects of failures. For usage-based failure strategy, we assume that the usage of a node is in proportion to its betweenness centrality. We can either deterministically fail the node with the largest betweenness centrality by assuming that the most used node is more likely to fail, or randomly select a node to fail.

The steps for fault-mitigation simulation are summarized as follows:

1. Compute betweenness centrality measures (multipartite, multi-attribute betweenness centrality) for all nodes.
2. Select the next node to fail according to the selected failure strategy (random or usage-based failure).
3. Simulate the effect of failures by removing the edges of the selected node.

The output of sequential failure simulation consists of a sequence of failure nodes and their effects in the form of updated betweenness centrality distributions.

We propose two measures to quantify the robustness of EES. First, we propose to quantify network fragmentations that may result in the loss of the ability in executing fault-mitigation operation, using the threshold for dissolving the giant component. Second, we propose to quantify the gradual changes of sequential failures using the mean of the normalized betweenness centrality.

A giant component is a connected sub-network that contains a majority of the entire network nodes. Since nodes in the giant component can all reach each other, this warrants the potential of executing fault-mitigation operations. However, when failures are induced, edges are removed from failed nodes. This may lead to network fragmentation which in turn dissolves the giant component; consequently, fault-mitigation operations may not be able to reach all nodes in the network. Hence we can evaluate the robustness of a layered EES network by considering how many failures are needed for a given failure strategy (usage-based or random) to reach a threshold value of nodes remaining in the giant component.

To quantify the gradual effect of sequential failures before the giant component reaches its dissolving threshold, we propose to use the changes in the means of the normalized betweenness centralities. By definition, nodes in two different fragmented subnets will not have shortest paths between them. This will lead to the decreasing of the mean of the normalized betweenness centrality for the whole network as sequential failures progress.

Since the output of a failure simulation records the effect of each simulated sequential failure, we can make recommendations on which nodes may potentially be burdened to implement fault-mitigation functions of the failed nodes. A simple heuristic is to use the neighbors of the failed node to carry out the function. Such heuristic may not be viable for usage-based strategy, as the failing node is the one that has the highest importance for pair-wised connections. Another heuristic is having every second highest importance node of the survival network fragments carry out the function of the failed node. This heuristic avoids immediate nearest-neighbor failing and at the same time carries out the failed node that needs to sit on many shortest paths.

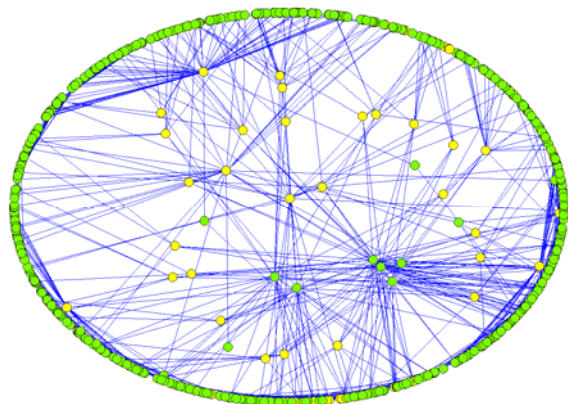


Figure 5: An example functional network depicting relations between virtual devices (green nodes) and ECUs (yellow nodes).

7. AN EXAMPLE STUDY

To demonstrate the values of proposed methods, we show our analysis on an example layered EES network. We first show how consideration of different node types may lead to different views of the importance of a node. We next show the effect of usage-based node failures based on different node types. Finally, we show simulation of sequential failures for fault-mitigation analysis.

We apply multi-partite betweenness centrality on the network depicted in Figure 5. We show the distributions of betweenness centrality for each part in Figure 6 and Figure 7.

We simulate the failures and inspect the changes of betweenness centrality measures. Figure 8 show an example of changes in the distribution of betweenness centrality for failing the top three ECUs. Nodes with increasing betweenness centrality after the failures can be considered as survival nodes that can carry out functions of failed nodes (e.g., Node9 and Node17 in Figure 8).

Betweenness Centrality Distribution Functional Network

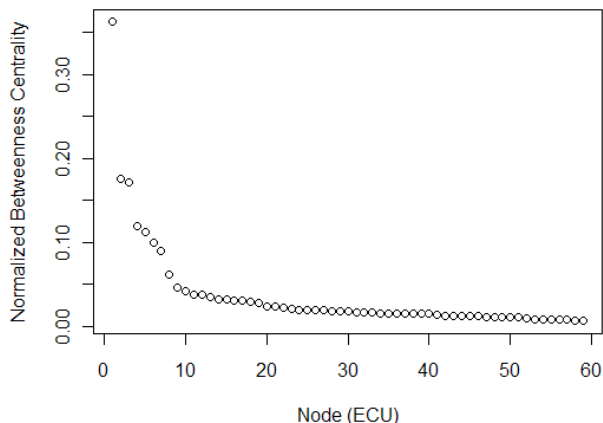


Figure 6: An example distribution of betweenness centrality in functional network. The distribution shows the top 59 ECU (out of 102) with above average betweenness centrality.

Betweenness Centrality Distribution Functional Network

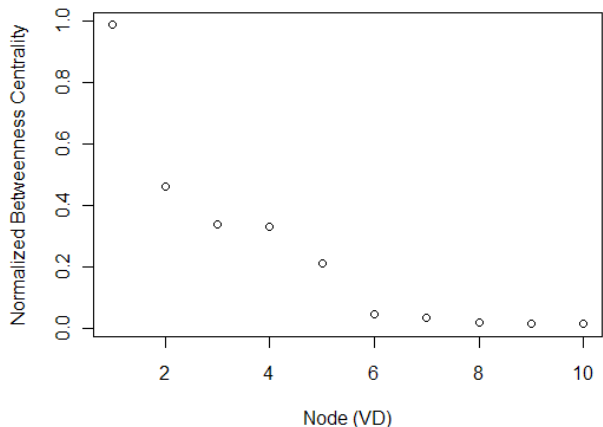


Figure 7: An example distribution of betweenness centrality for VD part in functional network. The distribution shows the top 10 VDs (out of 482) with above betweenness centrality.

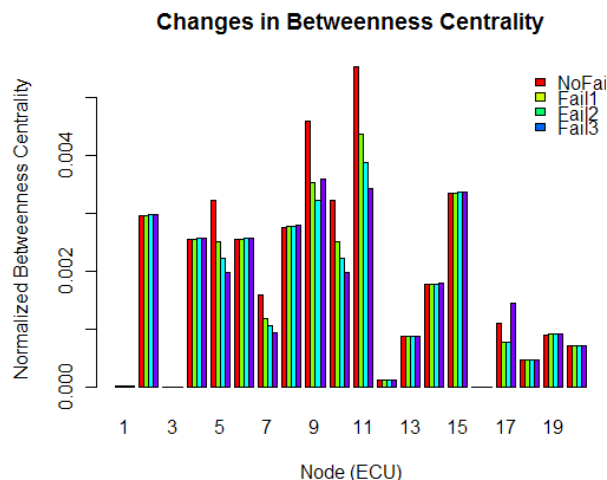


Figure 8: An example of changes in the distribution of betweenness centrality by failing the top 3 ranking ECU. The original ECU betweenness centrality is charted along with their new distribution after each simulated failures (Fail1, Fail2, and Fail3).

8. CONCLUSION

The network-theory based approach reported in this paper provides a first step toward integrated fault detection, isolation, and mitigation analysis capabilities for in-vehicle embedded electrical and electronic systems (EES). We apply layered network modeling over EES to build a layered multi-partite, multi-attribute network which represents physical, structural, functional, and data-flow aspects of in-vehicle EES. We employ two failure strategies to simulate failures and analyze the effects using betweenness centrality measures. We develop novel multi-partite, multi-attribute betweenness centrality to account for the effects of failures and to quantify complexity, maintainability, and robustness of EES. We provided an example to demonstrate our proposed methodology.

ACKNOWLEDGEMENT

The authors would like to thank Thomas Fuhrman, Mark Howell, and Shengbing Jiang for many valuable technical discussions. Special thanks also go to many of our colleagues in Engineering, including Doug Duddles, Sandeep Menon, Ken Orlando, Bob Schwabel, Lars Soderlund, Mike Sowa, and Natalie Wienckowski, for sharing the domain knowledge and shaping the research directions.

REFERENCES

Brandes, U. (2008). On Variants of Shortest-Path Betweenness Centrality and their Generic Computation, *Social Networks* 30(2): 136-145.

- Borgatti, S.P. (2005). Centrality and Network Flow, *Social Networks*: 27, 55-71.
- Flom, P.L., Friedman, S.R., Strauss, S. and Neaigus, A. (2004). A New Measure of Linkage Between Two Sub-networks, *Connections* 26 (1): 62-70.
- Freeman, L.C. (1978). Centrality in Social Networks: Conceptual Clarification, *Social Networks*: 1, 215-239.
- Simonot-Lion, F. (2009). Special Section on In-Vehicle Embedded Systems, *IEEE Transaction on Industrial Informatics*, Vol. 5, No. 4.
- Struss, P., Provan, G., de Kleer, J., and Biswas, G. (2010). Introduction to Special Issue on Model-Based Diagnosis, *IEEE Transaction on System, Man, and Cybernetics – Part A: Systems and Humans*, Vol 40, No 5.
- Wang, J. and Provan, G. (2010). A Benchmark Diagnostic Model Generation, *IEEE Transaction on System, Man, and Cybernetics – Part A: Systems and Humans*, Vol 40, No 5.
- Zeng, H., Di Natale, M., Giusto, P. and Sangiovanni-Vincentelli, A. (2009). Stochastic Analysis of CAN-based Real-Time Automotive Systems, *IEEE Transaction on Industrial Informatics*, Vol. 5, No. 4.

Tsai-Ching Lu received his M.S. degree in Computer Sciences from NYU in 1996, and his M.S. and Ph.D. degrees in Intelligent Systems from University of Pittsburgh in 2000 and 2003 respectively. He joined HRL Laboratories, LLC, in 2003 and currently holds the position of Senior Research Staff Scientist. His research interests include decision making under uncertainty, complex electronic diagnosis and prognosis, and behavior modeling and predictions. He received HRL's Distinguished Inventor Award in 2010.

Yilu Zhang received his B.S., and M.S. degrees in Electrical Engineering from Zhejiang University, China, in 1994, and 1997, respectively; and his Ph.D. degree in Computer Science from Michigan State University in 2002. He joined the R&D center of General Motors Co. at Warren, Michigan in 2002, and currently holds a position of Staff Researcher. Dr. Zhang's research interests include statistical pattern recognition, machine learning, signal processing, and their applications such as integrated system health management and human machine interactions. Dr. Zhang is a Senior Member of IEEE. He is a winner of 2008 "Boss" Kettering Award – the highest technology award in GM – for his contribution to Connected Vehicle Battery Monitor, a remote vehicle diagnostics technology.

David Allen received his M.S. and Ph.D. degrees in Computer Science from the University of California, Los Angeles (UCLA) in 2001 and 2005 respectively. He joined HRL Laboratories, LLC, in 2006 and currently holds the position of Research Staff Scientist. His research interests include artificial intelligence, probabilistic reasoning, decision making under uncertainty, complex systems

analysis, and network science; he has applied them to many domains including integrated system health management, cybersecurity, social network analysis, and behavior modeling. In 2010 he received HRL's Distinguished Inventor Award.

Mutasim Salman is a Lab. Group manager and a Technical Fellow in the Electrical, Controls and Integration Lab. of GM Research and Development Center. He has the responsibility of development and validation of algorithms for state of health monitoring, diagnosis, prognosis and fault tolerant control of vehicle critical systems. He pioneered the work on integrated chassis control in the late eighties that led to the production of GM "Industry First" Stabilitrak1 and then to Stabilitrak3. He had an extensive experience in hybrid vehicle, modeling, control and energy management strategies. He has several GM awards that includes 3 GM prestigious Boss Kettering, 3 McCuen, and 2 President and Chairman Awards. Mutasim received his bachelor's degree in Electrical Engineering from University of Texas at Austin; M.S. and PhD in Electrical Engineering with specialization in Systems and control from University of Illinois at Urbana- Champaign. He also has an Executive MBA. He is IEEE senior member. He holds 21 patents and has coauthored more than 42 refereed technical publications and a book. He joined the GM R&D staff in 1984.

Distributed Damage Estimation for Prognostics based on Structural Model Decomposition

Matthew Daigle¹ Anibal Bregon² and Indranil Roychoudhury³

¹ *University of California, Santa Cruz, NASA Ames Research Center, Moffett Field, CA, 94035, USA*
matthew.j.daigle@nasa.gov

² *Department of Computer Science, University of Valladolid, Valladolid, 47011, Spain*
anibal@infor.uva.es

³ *SGT, Inc., NASA Ames Research Center, Moffett Field, CA, 94035, USA*
indranil.roychoudhury@nasa.gov

ABSTRACT

Model-based prognostics approaches capture system knowledge in the form of physics-based models of components that include how they fail. These methods consist of a damage estimation phase, in which the health state of a component is estimated, and a prediction phase, in which the health state is projected forward in time to determine end of life. However, the damage estimation problem is often multi-dimensional and computationally intensive. We propose a model decomposition approach adapted from the diagnosis community, called possible conflicts, in order to both improve the computational efficiency of damage estimation, and formulate a damage estimation approach that is inherently distributed. Local state estimates are combined into a global state estimate from which prediction is performed. Using a centrifugal pump as a case study, we perform a number of simulation-based experiments to demonstrate the approach.

1. INTRODUCTION

Model-based prognostics approaches capture knowledge of how a system and its components fail through the use of physics-based models that capture the underlying physical phenomena (Daigle & Goebel, 2010b; Saha & Goebel, 2009; Luo, Pattipati, Qiao, & Chigusa, 2008). Model-based prognostics algorithms consist of two parts: (i) *damage estimation*, which is fundamentally a joint state-parameter estimation problem, and (ii) *prediction*, which projects the current joint state-parameter estimate forward in time to determine end of life (EOL). In (Daigle & Goebel, 2011), we developed a prognostics framework using particle filters for the damage estimation step that handles several simultaneously progressing damage processes. However, the approach may not scale

well as the number of damage processes to track (i.e., the dimension of the system) increases.

In this paper, we improve both the scalability and the computational efficiency of the damage estimation task by exploiting structural model decomposition (Williams & Millar, 1998), similar to methods developed within the diagnosis community (Bregon, Pulido, & Biswas, 2009; Roychoudhury, Biswas, & Koutsoukos, 2009; Staroswiecki & Declerck, 1989). In particular, we adopt the possible conflicts (PCs) approach (Pulido & Alonso-González, 2004). PCs decompose a global system model into minimal overdetermined subsystems (local submodels) for fault detection and isolation (Pulido & Alonso-González, 2004). PCs have also been used to formulate smaller estimation tasks for fault identification (Bregon, Pulido, & Biswas, 2009). In general, PCs can be used to automatically decompose a global joint state-parameter estimation task into a set of local estimation tasks that are easier to solve and require less overall computation. We use the PC approach to derive a minimal set of submodels and define a local damage estimation task for each one. Every local estimator computes a local joint state-parameter estimate, represented as a probability distribution. Then, the local estimates are merged into a global estimate from which prediction is performed in the typical way (Daigle & Goebel, 2011).

The models are decomposed into independent submodels by using measured signals as local inputs. Therefore, each local estimator operates independently, and the damage estimation becomes naturally distributed. Clearly then, this approach establishes a formal basis for distributed prognostics. This is in contrast to other proposed distributed prognostics approaches, e.g. (Saha, Saha, & Goebel, 2009), which still treat the prognostics problem as a global one in which only the computation is distributed, whereas we propose to decompose the global problem into a set of local ones for which computation may be trivially distributed.

Daigle et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

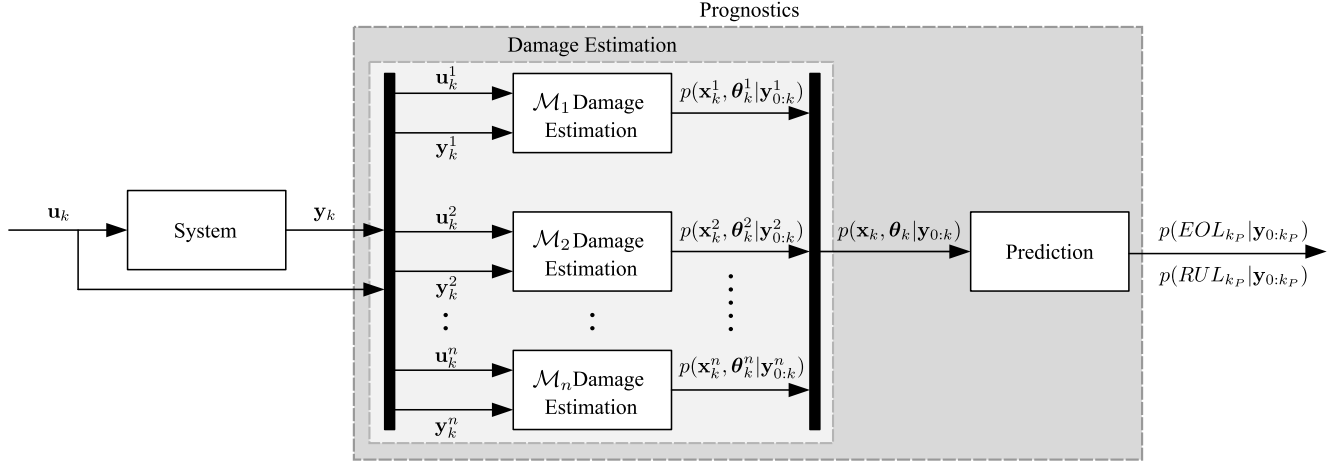


Figure 1. Prognostics architecture.

We demonstrate our prognostics methodology on a centrifugal pump. Centrifugal pumps appear in a variety of domains and often need to be operated for long time periods, hence diagnostics and prognostics become critical to ensuring continued operation that meets performance requirements. We apply our model-based prognostic approach based on structural model decomposition to centrifugal pumps using a number of simulation-based experiments when multiple damage mechanisms are active, and compare to results using the global estimation approach presented in (Daigle & Goebel, 2011).

The paper is organized as follows. Section 2 formally defines the prognostics problem and describes the prognostics architecture. Section 3 describes the modeling methodology and develops the centrifugal pump model for prognostics. Section 4 presents the model decomposition approach and provides results for the pump model. Section 5 describes the particle filter-based local damage estimation method. Section 6 discusses the prediction methodology. Section 7 provides results from simulation-based experiments and evaluates the approach. Section 8 concludes the paper.

2. PROGNOSTICS APPROACH

The goal of prognostics is the prediction of EOL and/or remaining useful life (RUL) of a component. In this section, we first formally define the problem of prognostics. We then describe the model-based prognostics architecture based on structural model decomposition.

2.1 Problem Formulation

In general, we define a system model as

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{f}(t, \mathbf{x}(t), \boldsymbol{\theta}(t), \mathbf{u}(t), \mathbf{v}(t)) \\ \mathbf{y}(t) &= \mathbf{h}(t, \mathbf{x}(t), \boldsymbol{\theta}(t), \mathbf{u}(t), \mathbf{n}(t)),\end{aligned}$$

where $t \in \mathbb{R}$ is a continuous time variable, $\mathbf{x}(t) \in \mathbb{R}^{n_x}$ is the state vector, $\boldsymbol{\theta}(t) \in \mathbb{R}^{n_\theta}$ is the parameter vector, $\mathbf{u}(t) \in \mathbb{R}^{n_u}$

is the input vector, $\mathbf{v}(t) \in \mathbb{R}^{n_v}$ is the process noise vector, \mathbf{f} is the state equation, $\mathbf{y}(t) \in \mathbb{R}^{n_y}$ is the output vector, $\mathbf{n}(t) \in \mathbb{R}^{n_n}$ is the measurement noise vector, and \mathbf{h} is the output equation. The parameters $\boldsymbol{\theta}(t)$ evolve in an unknown way.

The goal is to predict EOL (and/or RUL) at a given time point t_P using the discrete sequence of observations up to time t_P , denoted as $\mathbf{y}_{0:t_P}$. The component must meet a given set of functional requirements. We say the component has failed when it no longer meets one of these requirements. In general, we may capture this boundary on acceptable component behavior using a threshold that is a function of the system state and parameters, $T_{EOL}(\mathbf{x}(t), \boldsymbol{\theta}(t))$, where $T_{EOL}(\mathbf{x}(t), \boldsymbol{\theta}(t)) = 1$ if the system has failed and $T_{EOL}(\mathbf{x}(t), \boldsymbol{\theta}(t)) = 0$ otherwise. Using T_{EOL} , we formally define EOL with

$$EOL(t_P) \triangleq \inf\{t \in \mathbb{R} : t \geq t_P \wedge T_{EOL}(\mathbf{x}(t), \boldsymbol{\theta}(t)) = 1\},$$

i.e., EOL is the earliest time point at which the damage threshold is met. RUL is then

$$RUL(t_P) \triangleq EOL(t_P) - t_P.$$

Due to the noise terms $\mathbf{v}(t)$ and $\mathbf{n}(t)$, and uncertainty in the future inputs of the system, we at best compute only a probability distribution of the EOL or RUL, i.e., $p(EOL(t_P) | \mathbf{y}_{0:t_P})$ or $p(RUL(t_P) | \mathbf{y}_{0:t_P})$.

2.2 Prognostics Architecture

In our model-based approach, we develop detailed physics-based models of components and systems that include descriptions of how faults and damage evolves in time. These models depend on unknown parameters $\boldsymbol{\theta}(t)$. Therefore, damage estimation is fundamentally a joint state-parameter estimation problem. In discrete time k , we jointly estimate \mathbf{x}_k and $\boldsymbol{\theta}_k$, and use these estimates to predict EOL and RUL at desired time points. Here, we assume that prognostics is

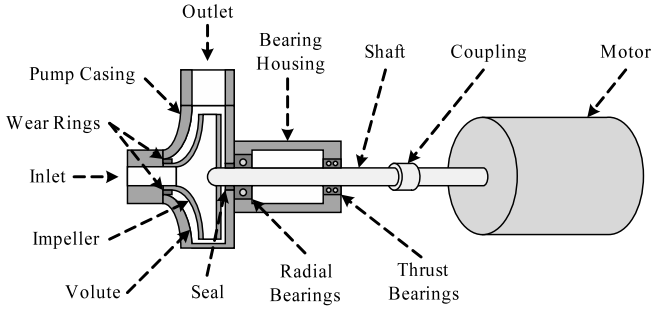


Figure 2. Centrifugal pump.

not aided by a fault diagnosis module, and so we must jointly estimate all possible damage modes.

We employ the prognostics architecture in Fig. 1. The system is provided with inputs \mathbf{u}_k and provides measured outputs \mathbf{y}_k . The damage estimation module takes as input both \mathbf{u}_k and \mathbf{y}_k , and produces the estimate $p(\mathbf{x}_k, \boldsymbol{\theta}_k | \mathbf{y}_{0:k})$. In this work, we decompose the global damage estimation problem into several subproblems based on model decomposition, as shown in Fig. 1. A model decomposition algorithm splits the global model into n submodels, $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_n$. We construct for each submodel \mathcal{M}_i a local estimator that performs damage estimation. Each estimator has input $\mathbf{u}_k^i \subseteq \mathbf{u}_k$ and $\mathbf{y}_k^i \subseteq \mathbf{y}_k$ and produces the local state estimate $p(\mathbf{x}_k^i, \boldsymbol{\theta}_k^i | \mathbf{y}_{0:k}^i)$, where $\mathbf{x}_k^i \subseteq \mathbf{x}_k$ and $\boldsymbol{\theta}_k^i \subseteq \boldsymbol{\theta}_k$. Note that for two submodels \mathcal{M}_i and \mathcal{M}_j , in general it is possible that $\mathbf{x}_k^i \cap \mathbf{x}_k^j \neq \emptyset$ and $\boldsymbol{\theta}_k^i \cap \boldsymbol{\theta}_k^j \neq \emptyset$. The local estimates are merged into the global estimate $p(\mathbf{x}_k, \boldsymbol{\theta}_k | \mathbf{y}_{0:k})$. The prediction module uses this joint state-parameter distribution, along with hypothesized future inputs, to compute EOL and RUL as probability distributions $p(EOL_{k_P} | \mathbf{y}_{0:k_P})$ and $p(RUL_{k_P} | \mathbf{y}_{0:k_P})$ at given prediction times k_P .

3. PUMP MODELING

We apply our prognostics approach to a centrifugal pump, and develop a physics-based model of its nominal and faulty behavior. Centrifugal pumps are used in a variety of domains for fluid delivery. A schematic of a typical centrifugal pump is shown in Fig. 2. Fluid enters the inlet, and the rotation of the impeller, driven by an electric motor, forces fluid through the outlet. Radial and thrust bearings, along with lubricating oil contained within the bearing housing, helps to minimize friction along the pump shaft. Wear rings prevent internal pump leakage from the outlet to the inlet side of the impeller, but a small clearance is typically allowed to minimize friction (a small internal leakage is normal). We review here the main features of the model, and refer the reader to (Daigle & Goebel, 2011) for details.

The state of the pump is given by

$$\mathbf{x}(t) = [\omega(t) \quad T_t(t) \quad T_r(t) \quad T_o(t)]^T,$$

where $\omega(t)$ is the rotational velocity of the pump, $T_t(t)$ is the thrust bearing temperature, $T_r(t)$ is the radial bearing temperature, and $T_o(t)$ is the oil temperature.

The rotational velocity of the pump is described using a torque balance,

$$\dot{\omega} = \frac{1}{J} (\tau_e(t) - r\omega(t) - \tau_L(t)),$$

where J is the lumped motor/pump inertia, τ_e is the electromagnetic torque provided by the motor, r is the lumped friction parameter, and τ_L is the load torque.

We assume the pump is driven by an induction motor with a polyphase supply. A torque is produced on the rotor only when there is a difference, i.e., a *slip*, between the synchronous speed of the supply voltage, ω_s and the mechanical rotation, ω . Slip, s , is defined as

$$s = \frac{\omega_s - \omega}{\omega_s}.$$

The expression for the torque τ_e is derived from an equivalent circuit representation for the three-phase induction motor, based on rotor and stator resistances and inductances and the slip s (Lyshevski, 1999):

$$\tau_e = \frac{npR_2}{s\omega_s} \frac{V_{rms}^2}{(R_1 + R_2/s)^2 + (\omega_s L_1 + \omega_s L_2)^2},$$

where R_1 is the stator resistance, L_1 is the stator inductance, R_2 is the rotor resistance, L_2 is the rotor inductance, n is the number of phases, and p is the number of magnetic pole pairs. The dependence of torque on slip creates a feedback loop that causes the rotor to follow the rotation of the magnetic field. The rotor speed may be controlled by changing the input frequency ω_s .

The load torque τ_L is a polynomial function of the flow rate through the pump and the impeller rotational velocity (Kallesøe, 2005):

$$\tau_L = a_0\omega^2 + a_1\omega Q - a_2Q^2,$$

where Q is the flow, and a_0 , a_1 , and a_2 are coefficients derived from the pump geometry (Kallesøe, 2005).

The rotation of the impeller creates a pressure difference from the inlet to the outlet of the pump, which drives the pump flow, Q . The pump pressure is computed as

$$p_p = A\omega^2 + b_1\omega Q - b_2Q^2,$$

where A is the impeller area, and b_1 and b_2 are coefficients derived from the pump geometry. Flow through the impeller, Q_i , is computed using the pressure differences:

$$Q_i = c\sqrt{|p_s + p_p - p_d| \text{sign}(p_s + p_p - p_d)},$$

where c is a flow coefficient, p_s is the suction pressure, and p_d is the discharge pressure. The small (normal) leakage flow

from the discharge end to the suction end due to the clearance between the wear rings and the impeller is described by

$$Q_l = c_l \sqrt{|p_d - p_s|} \text{sign}(p_d - p_s),$$

where c_l is a flow coefficient. The discharge flow, Q , is then

$$Q = Q_i - Q_l.$$

Pump temperatures are monitored as indicators of pump condition. The oil heats up due to the radial and thrust bearings and cools to the environment:

$$\dot{T}_o = \frac{1}{J_o} (H_{o,1}(T_t - T_o) + H_{o,2}(T_r - T_o) - H_{o,3}(T_o - T_a)),$$

where J_o is the thermal inertia of the oil, and the $H_{o,i}$ terms are heat transfer coefficients. The thrust bearings heat up due to the friction between the pump shaft and the bearings, and cool to the oil and the environment:

$$\dot{T}_t = \frac{1}{J_t} (r_t \omega^2 - H_{t,1}(T_t - T_o) - H_{t,2}(T_t - T_a)),$$

where J_t is the thermal inertia of the thrust bearings, r_t is the friction coefficient for the thrust bearings, and the $H_{t,i}$ terms are heat transfer coefficients. The radial bearings behave similarly:

$$\dot{T}_r = \frac{1}{J_r} (r_r \omega^2 - H_{r,1}(T_r - T_o) - H_{r,2}(T_r - T_a)),$$

where the parameters here take on analogous definitions.

The overall input vector \mathbf{u} is given by

$$\mathbf{u}(t) = [p_s(t) \quad p_d(t) \quad T_a(t) \quad V(t) \quad \omega_s(t)]^T.$$

The measurement vector \mathbf{y} is given by

$$\mathbf{y}(t) = [\omega(t) \quad Q(t) \quad T_t(t) \quad T_r(t) \quad T_o(t)]^T.$$

Fig. 3 shows nominal pump operation. The input voltage (and frequency) are varied to control the pump speed. The electromagnetic torque is produced initially as slip is 1. This causes a rotation of the motor to match the rotation of the magnetic field, with a small amount of slip remaining, depending on how large the load and friction torques are. As the pump rotates, fluid flow is created. The bearings heat up as the pump rotates and cool when the pump rotation slows.

3.1 Damage Modeling

For the purposes of prognostics, the model must include *damage variables* $\mathbf{d} \subseteq \mathbf{x}$ representing the amount of particular forms of damage. The most significant forms of damage for pumps are impeller wear, caused by cavitation and erosion by the flow, and bearing failure, caused by friction-induced wear of the bearings. In each case, we map the damage to a particular parameter in the nominal model, and this parameter becomes a state variable in $\mathbf{d}(t)$. The evolution of these damage variables is described by *damage progression equations*,

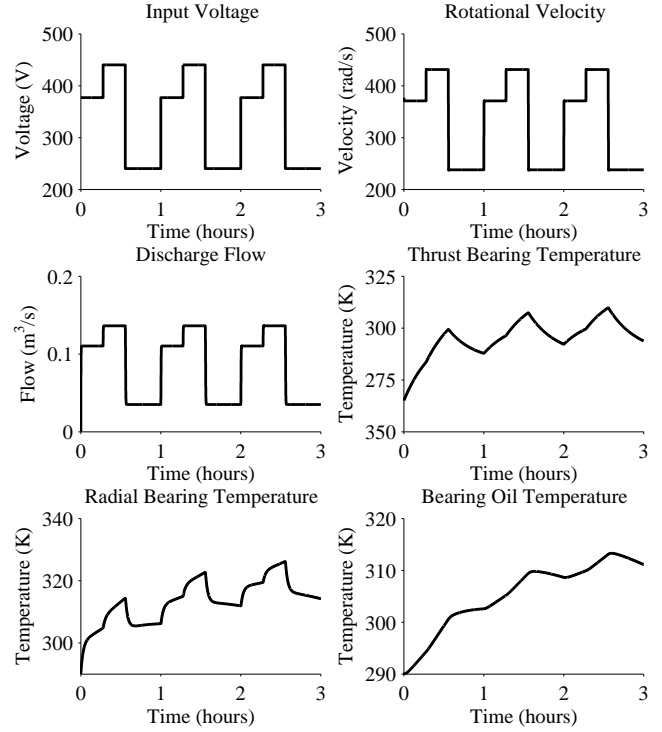


Figure 3. Nominal pump operation.

which are included in the state equation f. These equations are parameterized by unknown *wear parameters* $\mathbf{w} \subseteq \boldsymbol{\theta}$.

Impeller wear is represented as a decrease in effective impeller area A (Biswas & Mahadevan, 2007; Daigle & Goebel, 2011). We use the erosive wear equation (Hutchings, 1992):

$$\dot{A}(t) = -w_A Q_i^2,$$

where w_A is a wear coefficient. A decrease in the impeller area will decrease the pump pressure, which, in turn, reduces the delivered flow, and, therefore, pump efficiency. The pump must operate at a certain minimal efficiency, defining an EOL criteria. We define A^- as the minimum value of the impeller area at which this requirement is met, hence, $T_{EOL} = 1$ if $A(t) < A^-$.

Bearing wear is captured as an increase in friction. Sliding and rolling friction generate wear of material which increases the effective coefficient of friction (Hutchings, 1992; Daigle & Goebel, 2010b, 2011):

$$\dot{r}_t(t) = w_t r_t \omega^2,$$

$$\dot{r}_r(t) = w_r r_r \omega^2,$$

where w_t and w_r are the wear coefficients. The slip compensation provided by the electromagnetic torque generation will mask small changes in friction, but these changes can be observed using the bearing temperatures. Limits on the maximum values of these temperatures define EOL for bearing wear. We define r_t^+ and r_r^+ as the maximum permissible values of the friction coefficients, before the temperature limits

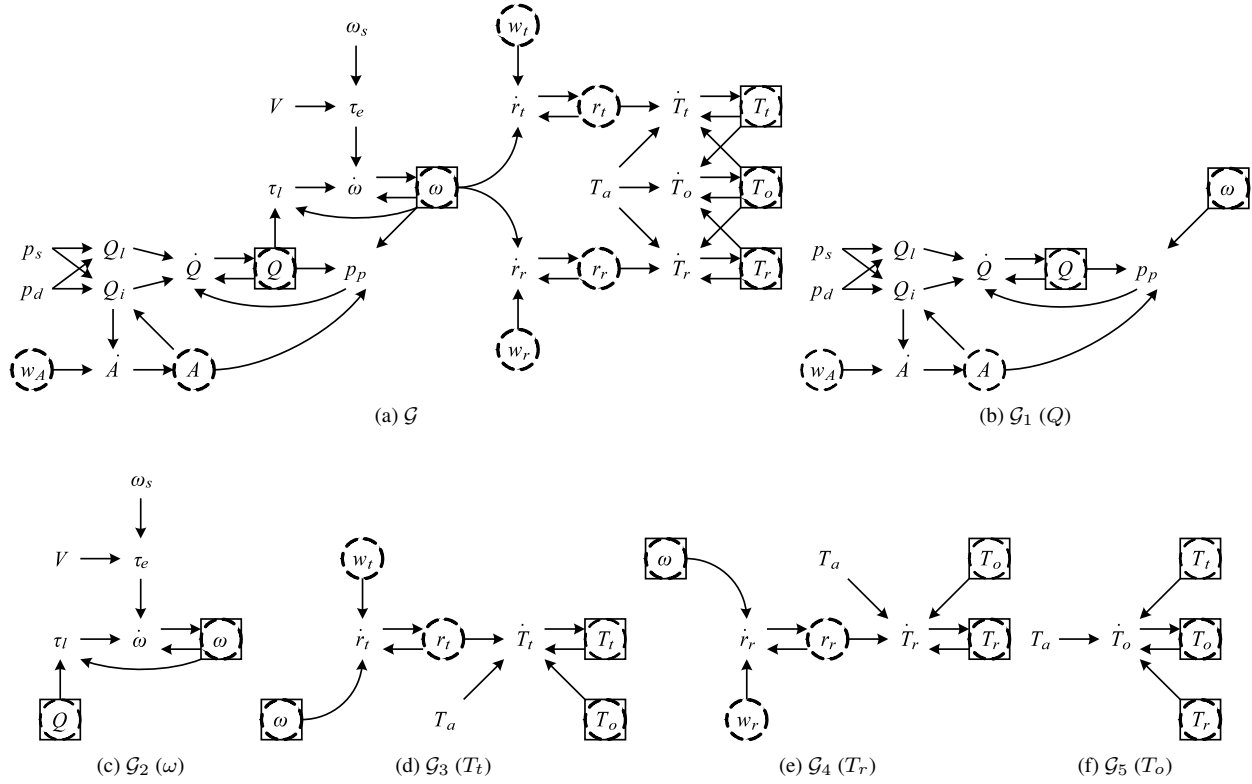


Figure 4. System graph and minimal subgraphs of the pump.

are exceeded over a typical usage cycle. So, $T_{EOL} = 1$ if $r_t(t) > r_t^+$ or $r_r(t) > r_r^+$.

So, the damage variables are given by

$$\mathbf{d}(t) = [A(t) \quad r_t(t) \quad r_r(t)]^T,$$

and the full state vector becomes

$$\mathbf{x}(t) = [\omega(t) \quad T_t(t) \quad T_r(t) \quad T_o(t) \quad A(t) \quad r_t(t) \quad r_r(t)]^T.$$

The wear parameters form the unknown parameter vector, i.e.,

$$\mathbf{w}(t) = \boldsymbol{\theta}(t) = [w_A \quad w_t \quad w_r]^T.$$

4. MODEL DECOMPOSITION

Especially for nonlinear systems, state and parameter estimation is a nontrivial problem for which no general closed-form solution exists. In general, these problems may be solved by numerical optimization methods, where the computational complexity is exponential in the size of the model, or by approximate Bayesian filtering methods, like particle filters, where the computational complexity is only linear in the size of the sample space, but, the number of sufficient samples grows with the model size.

Several approaches have been developed to decrease the computational complexity by model decomposition, in which the global model is decomposed into several independent submodels (Staroswiecki & Declerck, 1989; Williams & Millar,

1998). This results in a set of smaller, lower-dimensional estimation tasks. We adopt the PC approach for model decomposition. PCs are minimal subsets of equations with sufficient analytical redundancy to generate fault hypotheses from observed measurement deviations, and, in previous work, we used PCs to propose a more robust and computationally simpler parameter estimation approach for fault identification (Bregon, Pulido, & Biswas, 2009), where the parameter estimation task using the entire system model was replaced by a set of smaller estimation problems (one for each PC). In this approach, fault identification is fundamentally a joint state-parameter estimation problem. This is equivalent to the damage estimation problem in prognostics, only the models specifically include damage progression. So, in this paper, we adopt this paradigm for distributed damage estimation.

In order to compute the minimal set of submodels of a system, a structural representation of the system model is needed. In previous work, we computed submodels as PCs using hypergraphs (Pulido & Alonso-González, 2004) or Temporal Causal Graphs (TCGs) (Bregon, Pulido, & Biswas, 2009) as inputs. Representations that include computational causality, like TCGs, are favored because causality allows efficient derivation of PCs. In this work, we represent the system model with a directed hypergraph, and use an algorithm close to the TCG-based algorithm presented in (Bregon, Pulido, Biswas, & Koutsoukos, 2009).

The system model \mathcal{M} is represented using a set of functions \mathcal{F} over variables V , where a subset of the variables $X \subseteq V$ corresponds to the state variables \mathbf{x} , a subset $\Theta \subseteq V$ corresponds to the unknown parameters $\boldsymbol{\theta}$, a subset $U \subseteq V$ corresponds to the (known) inputs \mathbf{u} , and a subset $Y \subseteq V$ corresponds to the (measured) outputs \mathbf{y} . For the purposes of the model decomposition algorithm, we represent \mathcal{M} using an extended directed hypergraph $\mathcal{G} = (V, E, F)$, where V is the set of vertices corresponding directly to the variables in \mathcal{M} , E is the set of hyperedges of the form (V', v) with $V' \subseteq V$ being a set of vertices and $v \in V$ being a single vertex, and F is a map from an edge $(V', v) \in E$ to a function $f \in \mathcal{F}$ such that $v = f(v_1, v_2, \dots, v_n)$ where $V' = \{v_1, v_2, \dots, v_n\}$. \mathcal{M} and \mathcal{G} are equivalent data structures, where the direction of the edges in \mathcal{G} captures the computational causality of the functions in \mathcal{F} . The variable sets X , Θ , Y , and U are represented equivalently in \mathcal{G} as vertex sets, and we use $y_i \in Y$ to refer to the vertex/variable corresponding to the i th output in \mathbf{y} .

Fig. 4a shows the hypergraph \mathcal{G} for the pump model described in Section 3. Individual arrows pointing to the same vertex are to be interpreted as a single directed hyperedge. State variables are denoted using dashed circles, and measured variables are denoted with boxes.

Algorithm 1 computes subgraphs corresponding to PCs from a system graph \mathcal{G} . One PC will be computed for each output $y_i \in Y$. So, for the pump model, there will be five total submodels, one each for ω , Q , T_t , T_r , and T_o . For each vertex in Y , the algorithm propagates back to members of U and Y , which will be used as inputs to the submodel. Vertices which are not included in U or Y or have not yet been included in V_i are added to the set *vertices* for further backward propagation. For example, starting with T_t in Fig. 4a to form the subgraph shown in Fig. 4d, we propagate to T_o , a measured output, at which propagation terminates, and \hat{T}_t . From \hat{T}_t we propagate back further to the input T_a and the state r_t , from which we propagate to \hat{r}_t , from which we propagate to measured output ω and the unknown parameter w_t . All vertices and edges encountered are added to V_i and E_i , respectively, and the model equations corresponding to the added edges are included in the submodel as well. The algorithm forms from \mathcal{G} subgraphs $\mathcal{G}_i = (V_i, E_i, F_i)$, which may be easily translated to submodels \mathcal{M}_i . If each $v \in X \cup \Theta$ is causally linked to at least one output, then every variable $x \in X$ and $\theta \in \Theta$ will belong to at least one V_i over the set of \mathcal{G}_i computed, i.e., will belong to at least one submodel \mathcal{M}_i .

The algorithm decomposes the pump model into 5 submodels, with their corresponding subgraphs shown as Figs. 4b to 4f. For example, the T_t subgraph takes as input measurements of ω and T_o , and computes the expected value of T_t . Damage estimation for this submodel will compute estimates of T_t , r_t , and w_t . Note that, for the pump model, each state or parameter will be estimated by exactly one submodel, therefore, there will be no overlap in the local estimates.

Algorithm 1 $\{\mathcal{G}_i\}_{i=1}^{n_y} = \text{Decompose}(\mathcal{G})$

```

for  $i = 1$  to  $n_y$  do
   $V_i \leftarrow \{y_i\}$ 
   $E_i \leftarrow \emptyset$ 
  vertices  $\leftarrow \{y_i\}$ 
  while vertices  $\neq \emptyset$  do
     $v \leftarrow \text{vertices}\{1\}$ 
    vertices  $\leftarrow \text{vertices} \setminus \{v\}$ 
    edges  $\leftarrow \{(V', v) \in E : V' \subseteq V\}$ 
    for all  $(V', v) \in \text{edges}$  do
      for all  $v' \in V'$  do
        if  $v' \notin U$  and  $v' \notin Y$  and  $v' \notin V_i$  then
          vertices  $\leftarrow \text{vertices} \cup \{v'\}$ 
        end if
      end for
    end for
     $V_i \leftarrow V_i \cup V'$ 
     $E_i \leftarrow E_i \cup \{(V', v)\}$ 
     $F_i(V', v) \leftarrow F(V', v)$ 
  end while
   $\mathcal{G}_i \leftarrow (V_i, E_i, F_i)$ 
end for

```

5. DAMAGE ESTIMATION

In our local estimation scheme, the local estimator for each submodel \mathcal{M}_i produces a local estimate $p(\mathbf{x}_k^i, \boldsymbol{\theta}_k^i | \mathbf{y}_{0:k})$, where $\mathbf{x}_k^i \subseteq \mathbf{x}_k$ and $\boldsymbol{\theta}_k^i \subseteq \boldsymbol{\theta}_k$. The local estimates are combined into the global state estimate $p(\mathbf{x}_k, \boldsymbol{\theta}_k | \mathbf{y}_{0:k})$.

Due to the decoupling introduced by the decomposition scheme, we lose information about the covariance between states in separate submodels. If these covariances are nominally small, then this information loss is acceptable and the approximation of the global state estimate obtained by merging the local state estimates into a global state estimate will closely approximate the global estimate obtained through a global damage estimator. Although we lose information due to the decoupling, the advantage is that the local estimation tasks are naturally distributed, and therefore, unlike the global estimation approach, the distributed approach scales well as the size of the model increases. Further, the local estimation tasks become easier to solve and should require less computational resources without sacrificing estimation performance. This should be the case as long as the sensor measurements that are used as inputs to the submodels are reliable and do not exhibit extremely high noise.

A general solution to the problem of damage estimation is the *particle filter*, which may be directly applied to nonlinear systems with non-Gaussian noise terms (Arulampalam, Maskell, Gordon, & Clapp, 2002). The main disadvantage of the particle filter is the computational complexity, which is linear in the amount of samples, or *particles*, that are used to approximate the state distribution, as typically a large number of particles are needed, and the sufficient number of particles increases with the dimension of the state-parameter space. A

key advantage of the model decomposition algorithm is that it creates submodels that are simpler than the global model. Some of these submodels may be completely linear, and some may require only state estimation, not joint state-parameter estimation. If only state estimation is required, the Kalman filter or one of its nonlinear extensions may be used, which requires computation on the order of a particle filter using one particle (e.g., Kalman filter or extended Kalman filter) or a number of particles linear in the state dimension (e.g., unscented Kalman filter (Julier & Uhlmann, 1997)), resulting in a significant improvement in computational complexity. The remaining submodels that require joint state-parameter estimation represent several small, low-dimensional estimation problems that are easier to solve than the global one, therefore requiring less computation overall.

In particle filters, the state distribution is approximated by a set of discrete weighted samples, or particles:

$$\{(\mathbf{x}_k^{(i)}, \boldsymbol{\theta}_k^{(i)}, w_k^{(i)})\}_{i=1}^N,$$

where N denotes the number of particles, and for particle i , $\mathbf{x}_k^{(i)}$ denotes the state vector estimate, $\boldsymbol{\theta}_k^{(i)}$ denotes the parameter vector estimate, and $w_k^{(i)}$ denotes the weight. The posterior density is approximated by

$$p(\mathbf{x}_k, \boldsymbol{\theta}_k | \mathbf{y}_{0:k}) \approx \sum_{i=1}^N w_k^{(i)} \delta_{(\mathbf{x}_k^{(i)}, \boldsymbol{\theta}_k^{(i)})} (d\mathbf{x}_k d\boldsymbol{\theta}_k),$$

where $\delta_{(\mathbf{x}_k^{(i)}, \boldsymbol{\theta}_k^{(i)})} (d\mathbf{x}_k d\boldsymbol{\theta}_k)$ denotes the Dirac delta function located at $(\mathbf{x}_k^{(i)}, \boldsymbol{\theta}_k^{(i)})$.

We use the sampling importance resampling (SIR) particle filter, using systematic resampling. Each particle is propagated forward to time k by first sampling new parameter values, and then sampling new states using the model. The particle weight is assigned using \mathbf{y}_k . The weights are then normalized, followed by the resampling step. Pseudocode is provided in (Arulampalam et al., 2002; Daigle & Goebel, 2011).

The parameters $\boldsymbol{\theta}_k$ evolve by some unknown random process that is independent of the state \mathbf{x}_k . To perform parameter estimation within a particle filter framework, we assign a random walk evolution, i.e., $\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} + \boldsymbol{\xi}_{k-1}$, where $\boldsymbol{\xi}_{k-1}$ is a noise vector. During the sampling step, particles are generated with parameter values that will be different from the current values of the parameters. The particles with parameter values closest to the true values should match the outputs better, and therefore be assigned higher weight. Resampling will cause more particles to be generated with similar values, so the particle filter converges to the true values as the process is repeated over each step of the algorithm. In general, though, convergence is not always guaranteed.

The selected variance of the random walk noise determines both the rate of this convergence and the estimation performance after convergence. Therefore, this parameter should

be tuned to obtain the best possible performance, but the optimal value is dependent on the value of the hidden wear parameter, which is unknown. We use the variance control method presented in (Daigle & Goebel, 2011). In this approach, the variance of the hidden wear parameter estimate is controlled to a user-specified range by modifying the random walk noise variance. We assume that the $\boldsymbol{\xi}$ values are tuned initially based on the maximum expected wear rates. The algorithm uses relative median absolute deviation (RMAD) as the measure of spread. The adaptation scheme controls the error between the actual RMAD of a parameter $\boldsymbol{\theta}(j)$, denoted as v_j , and the desired RMAD value (e.g., 10%), denoted as v_j^* , using a proportional control strategy governed by a gain P (e.g., 1×10^{-3}). There are two different setpoints. The first allows for a convergence period, with setpoint v_{j0}^* (e.g., 50%). Once v_j reaches T (e.g., $1.2v_{j0}^*$), a new setpoint $v_{j\infty}^*$ (e.g., 10%) is established. The advantage of this methodology is that the random walk variance is automatically tuned to achieve the best performance for the requested relative spread for the actual value of the hidden parameter.

6. PREDICTION

Prediction is initiated at a given time k_P . In order to obtain a prediction that is valid for the global state-parameter vector, we must first combine the local estimates into a global estimate. To do this, we assume that the local and global state estimates may be sufficiently approximated by a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. So, for each local state-parameter distribution i , we obtain the mean $\boldsymbol{\mu}^i$ and covariance matrix $\boldsymbol{\Sigma}^i$. We then combine all of these into a global mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. If there is overlap in the state-parameter estimates, i.e., if two submodels both estimate the same state variable x or parameter θ , then we take the average value for common means and covariances (alternate strategies may also be used). The covariance information lost due to the decoupling will appear as zeros in the global covariance matrix.

We then sample from the global state-parameter distribution defined by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a number of times to obtain a set of samples that sufficiently approximates the distribution defined by those parameters, which may each be simulated to EOL. An alternate approach is to use the unscented transform to deterministically select the minimal number of samples from this distribution that capture the statistical moments as described in (Daigle & Goebel, 2010a). Although the latter method is computationally more efficient, we use the former method here in order to obtain a fair comparison to the results presented for the global estimation approach in (Daigle & Goebel, 2011).

Using the global joint state-parameter estimate at k_P , $p(\mathbf{x}_{k_P}, \boldsymbol{\theta}_{k_P} | \mathbf{y}_{0:k_P})$, which represents the current knowledge of the system at time k_P , the goal is to compute $p(EOL_{k_P} | \mathbf{y}_{0:k_P})$ and $p(RUL_{k_P} | \mathbf{y}_{0:k_P})$. As discussed in

Algorithm 2 EOL Prediction

Inputs: $\{\mathbf{x}_{k_P}^{(i)}, \boldsymbol{\theta}_{k_P}^{(i)}, w_{k_P}^{(i)}\}_{i=1}^N$
Outputs: $\{EOL_{k_P}^{(i)}, w_{k_P}^{(i)}\}_{i=1}^N$
for $i = 1$ **to** N **do**
 $k \leftarrow k_P$
 $\mathbf{x}_k^{(i)} \leftarrow \mathbf{x}_{k_P}^{(i)}$
 $\boldsymbol{\theta}_k^{(i)} \leftarrow \boldsymbol{\theta}_{k_P}^{(i)}$
while $T_{EOL}(\mathbf{x}_k^{(i)}, \boldsymbol{\theta}_k^{(i)}) = 0$ **do**
 Predict $\hat{\mathbf{u}}_k$
 $\boldsymbol{\theta}_{k+1}^{(i)} \sim p(\boldsymbol{\theta}_{k+1} | \boldsymbol{\theta}_k^{(i)})$
 $\mathbf{x}_{k+1}^{(i)} \sim p(\mathbf{x}_{k+1} | \mathbf{x}_k^{(i)}, \boldsymbol{\theta}_k^{(i)}, \hat{\mathbf{u}}_k)$
 $k \leftarrow k + 1$
 $\mathbf{x}_k^{(i)} \leftarrow \mathbf{x}_{k+1}^{(i)}$
 $\boldsymbol{\theta}_k^{(i)} \leftarrow \boldsymbol{\theta}_{k+1}^{(i)}$
end while
 $EOL_{k_P}^{(i)} \leftarrow k$
end for

Section 5, the particle filter computes

$$p(\mathbf{x}_{k_P}, \boldsymbol{\theta}_{k_P} | \mathbf{y}_{0:k_P}) \approx \sum_{i=1}^N w_{k_P}^{(i)} \delta_{(\mathbf{x}_{k_P}^{(i)}, \boldsymbol{\theta}_{k_P}^{(i)})} (d\mathbf{x}_{k_P} d\boldsymbol{\theta}_{k_P}).$$

We can approximate a prediction distribution n steps forward as (Doucet, Godsill, & Andrieu, 2000)

$$p(\mathbf{x}_{k_P+n}, \boldsymbol{\theta}_{k_P+n} | \mathbf{y}_{0:k_P}) \approx \sum_{i=1}^N w_{k_P}^{(i)} \delta_{(\mathbf{x}_{k_P+n}^{(i)}, \boldsymbol{\theta}_{k_P+n}^{(i)})} (d\mathbf{x}_{k_P+n} d\boldsymbol{\theta}_{k_P+n}).$$

So, for a particle i propagated n steps forward without new data, we may take its weight as $w_{k_P}^{(i)}$. Similarly, we can approximate the EOL as

$$p(EOL_{k_P} | \mathbf{y}_{0:k_P}) \approx \sum_{i=1}^N w_{k_P}^{(i)} \delta_{EOL_{k_P}^{(i)}} (dEOL_{k_P}).$$

To compute EOL, then, we simulate each particle forward to its own EOL and use that particle's weight at k_P for the weight of its EOL prediction. The pseudocode for the prediction procedure is given as Algorithm 2 (Daigle & Goebel, 2010b). Each particle i is propagated forward until $T_{EOL}(\mathbf{x}_k^{(i)}, \boldsymbol{\theta}_k^{(i)})$ evaluates to 1. The algorithm hypothesizes future inputs of the system, $\hat{\mathbf{u}}_k$. In this work, we consider the situation where a single future input trajectory is known.

7. RESULTS

We performed a number of simulation-based experiments to analyze the performance of the prognostics approach using local damage estimation. For the purposes of comparison, we include results from the global estimation approach. For the local estimation approach, we use the five submodels derived

in Section 4, where the three submodels associated with damage models use particle filters for joint state-parameter estimation, and the remaining two submodels, which require only state estimation, use extended Kalman filters. The global approach used $N = 500$, so when the three local particle filters each use $N = 167$, the total computational cost is equivalent to that of the global particle filter. We try also $N = 100$ and $N = 50$ to observe the changes in performance when less total computation is performed. Since the local estimators use measured values as inputs, performance will degrade as sensor noise is increased. We varied the sensor noise variance by factors of 1, 10, 100, and 1000, to explore this situation.

In a single experiment, combinations of wear parameter values were selected randomly within a range. We selected the true wear parameter values in $[1 \times 10^{-3}, 4 \times 10^{-3}]$ for w_A , and in $[1 \times 10^{-11}, 7 \times 10^{-11}]$ for w_t and w_r , such that the maximum wear rates corresponded to a minimum EOL of 20 hours. The local estimators had to estimate both the local states and the local unknown wear parameters. In all experiments, we used $T = 60\%$, $v_0^* = 50\%$, $v_\infty^* = 10\%$, and $P = 1 \times 10^{-4}$ for the variance control algorithm. We performed 20 experiments for each value of N and sensor noise level. We considered the case where the future input of the pump is known, and it is always operated at a constant RPM, in order to limit the uncertainty to only that involved in the noise terms and that introduced by the filtering algorithms.

The averaged estimation and prediction performance results are shown in Table 1. The part of the table with $|\mathcal{M}| = 1$ corresponds to results using the global model. The column labeled N lists the number of particles used per submodel, and the column labeled \mathbf{n} lists the sensor noise variance multipliers. Here, we use percent root mean square error (PRMSE) as a measure of estimation accuracy, relative accuracy (RA) (Saxena, Celaya, Saha, Saha, & Goebel, 2010) as a measure of prediction accuracy, and RMAD as a measure of spread. Each are averaged over multiple prediction points for a single scenario (see (Saxena et al., 2010; Daigle & Goebel, 2011) for the mathematical definitions of the metrics used here). Note that all metrics are expressed as percentages.

We can see that in the case where $N = 167$, for the same amount of computation, the local estimation approach obtains results very close to the global approach for damage estimation accuracy. In some cases, performance is slightly better, and in other cases, slightly worse. At the highest noise level, the local approach improves significantly for estimation of w_A . This is mostly due to the convergence properties of the global approach. It tends to converge with much more difficulty than the local approach for the same amount of noise. RMADs of the wear parameters are also quite evenly matched. Here again, in some cases the local approach is slightly better, and in others slightly worse. As the sensor noise increases, the variance is naturally larger and more difficult to control, resulting in the increase in RMAD as sen-

Table 1. Estimation and Prediction Performance

$ \mathcal{M} $	N	\mathbf{n}	PRMSE_{w_A}	PRMSE_{w_t}	PRMSE_{w_r}	$\overline{\text{RMAD}}_{w_A}$	$\overline{\text{RMAD}}_{w_t}$	$\overline{\text{RMAD}}_{w_r}$	RA	$\overline{\text{RMAD}}_{RUL}$
1	500	1	3.70	3.58	2.54	11.58	11.27	10.03	97.28	11.61
1	500	10	4.15	2.81	2.74	12.25	11.48	10.63	96.58	12.34
1	500	100	6.30	3.46	3.23	13.46	12.38	11.59	94.69	14.09
1	500	1000	12.93	6.25	5.29	13.92	12.99	12.64	79.37	15.32
5	167	1	3.19	2.61	2.88	12.26	10.85	10.76	96.61	12.01
5	167	10	3.66	2.90	3.56	12.69	11.09	11.85	95.28	13.32
5	167	100	4.44	3.39	3.78	13.05	11.78	12.56	93.17	14.57
5	167	1000	5.59	4.46	8.26	14.72	12.86	15.09	87.66	16.19
5	100	1	4.02	3.49	3.49	12.42	10.68	10.77	95.90	12.15
5	100	10	4.52	3.78	4.27	12.69	11.04	11.45	95.16	13.30
5	100	100	5.71	4.02	6.05	12.99	11.81	12.73	91.82	14.74
5	100	1000	9.06	4.76	6.91	13.83	12.15	13.92	79.90	16.40
5	50	1	5.66	4.98	5.19	12.33	10.41	10.39	94.59	12.39
5	50	10	6.12	4.92	6.29	12.41	10.71	11.21	93.44	12.99
5	50	100	7.43	6.08	8.24	12.94	11.16	11.91	90.05	14.19
5	50	1000	14.03	9.33	14.41	12.90	11.66	12.46	73.05	13.62

sensor noise increases for both approaches. As the number of particles used in the local approach is reduced, accuracy decreases, as expected, but not significantly. The RMADs actually generally decrease as the number of particles is reduced, corresponding to increased precision at the cost of decreased accuracy.

Prediction performance corresponds to the change in damage estimation performance. More accurate damage estimates correspond to higher RA, and increases in spread of the wear parameters leads to increases of the spread of the RUL. For $N = 167$, the performance is slightly worse, but still comparable to the global estimation approach, even though some of the covariance information in the global state estimate was lost due to the decoupling. At the highest noise level, the local approach has significantly better accuracy. This is due to the relatively poor convergence behavior of the global approach, leading to inaccurate early predictions that bring down the averaged RA. As N decreases, so that less total computation is being performed, accuracy reduces, but so does spread. For $N = 100$, less computation is performed with only a small change in performance. As N is reduced further to 50, performance begins to degrade. Moreover, as sensor noise increases, the local approach can lose its advantage over the global approach, and this occurs when N is reduced to 50. For $N = 50$ and the nominal amount of sensor noise, comparable prognostics performance is achieved to the global approach with less than a third of the computation. We expect the benefits of the local approach to be more pronounced as

the dimension of the state-parameter space increases.

8. CONCLUSIONS

In this paper, we developed a novel distributed damage estimation approach for model-based prognostics that is based on a formal framework for structural model decomposition. Using the concept of PCs, a system model is decomposed into a set of minimal submodels. A local damage estimation problem is defined for each submodel. Local state-parameter estimates obtained using an appropriate filter are merged into a global state-parameter estimate from which EOL predictions are computed. Results demonstrate that equivalent, or in some cases, better prognostics performance can be achieved using this methodology with less computation than a global approach. Further, the approach can be naturally distributed and therefore may serve as a fundamental aspect of a practical system-level prognostics approach.

The idea of using model decomposition to improve state and parameter estimation is not new. For example, subspace methods (Katayama, 2005) for system identification employ QR-factorization and singular-value decomposition (Overschee & Moor, 1996) for solving identification problems in large-dimension systems. These methods are numerically robust for linear systems. Recently, several extensions have been proposed that apply to nonlinear systems (e.g., (Westwick & Verhaegen, 1996)). However, methods to automatically derive the decomposition from the system model have not been addressed.

An approach for decomposing a system model into smaller hierarchically organized subsystems, called *dissents*, is described in (Williams & Millar, 1998). PCs are conceptually equivalent to dissents, and previous work applied PCs for model decomposition to generate a more robust and computationally simpler parameter estimation approach for fault identification (Bregon, Pulido, & Biswas, 2009). Simulation results in that case showed an improvement in estimation accuracy while having a faster convergence to true solutions. Similar work was proposed in (Roychoudhury et al., 2009) using a dynamic Bayesian network (DBN) modeling framework, in which an automatic approach for model decomposition into factors based on structural observability was developed for efficient state estimation and fault identification. This approach also obtained an improvement in state estimation efficiency without compromising estimation accuracy. The relation between both approaches has been established in (Alonso-Gonzalez, Moya, & Biswas, 2010), where DBNs are derived from PCs for the purposes of estimation.

In future work, we will investigate extensions to system-level and distributed prognostics. For one, the model decomposition algorithm suggests a sensor placement strategy to optimize the decomposition of a system-level model into independent component models. One need only place sensors at the inputs and outputs of components to ensure that component models may be decoupled. The submodels derived from the PC approach are minimal in that, for a given output, they contain only the subset of the model required to compute that output as a function of only inputs and other measured outputs. Nonminimal submodels may be formed by merging minimal submodels, and this may be desired in some cases, e.g., if it eliminates using some high-noise sensors as inputs. This forms part of a more generalized model decomposition framework under development. As described in the paper, distributed damage estimation is an essential part of a distributed model-based prognostics architecture. The computation associated with the prediction problem in our approach can be trivially distributed (via parallel EOL simulations), but in future work, we would like to develop a decomposition of the prediction problem into local prediction problems for a fully distributed prognostics architecture.

ACKNOWLEDGMENTS

Anibal Bregon's work has been partially supported by the Spanish MCI DPI2008-01996 and MCI TIN2009-11326 grants.

REFERENCES

Alonso-Gonzalez, C., Moya, N., & Biswas, G. (2010). Factoring dynamic Bayesian networks using possible conflicts. In *Proc. of the 21th International Workshop on Principles of Diagnosis* (p. 7-14). Portland, OR, USA.

- Arulampalam, M. S., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for on-line nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2), 174–188.
- Biswas, G., & Mahadevan, S. (2007, March). A Hierarchical model-based approach to systems health management. In *Proc. of the 2007 IEEE Aerospace Conference*.
- Bregon, A., Pulido, B., & Biswas, G. (2009, Sep). Efficient on-line parameter estimation in TRANSCEND for nonlinear systems. In *Proc. of the Annual Conference of the Prognostics and Health Management Society 2009*. San Diego, USA.
- Bregon, A., Pulido, B., Biswas, G., & Koutsoukos, X. (2009). Generating possible conflicts from bond graphs using temporal causal graphs. In *Proceedings of the 23rd European Conference on Modelling and Simulation* (p. 675-682). Madrid, Spain.
- Daigle, M., & Goebel, K. (2010a, October). Improving computational efficiency of prediction in model-based prognostics using the unscented transform. In *Proc. of the Annual Conference of the Prognostics and Health Management Society 2010*.
- Daigle, M., & Goebel, K. (2010b, March). Model-based prognostics under limited sensing. In *Proceedings of the 2010 IEEE Aerospace Conference*.
- Daigle, M., & Goebel, K. (2011, March). Multiple damage progression paths in model-based prognostics. In *Proceedings of the 2011 IEEE Aerospace Conference*.
- Doucet, A., Godsill, S., & Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10, 197–208.
- Hutchings, I. M. (1992). *Tribology: friction and wear of engineering materials*. CRC Press.
- Julier, S. J., & Uhlmann, J. K. (1997). A new extension of the Kalman filter to nonlinear systems. In *Proceedings of the 11th International Symposium on Aerospace/Defense Sensing, Simulation and Controls* (pp. 182–193).
- Kallesøe, C. (2005). *Fault detection and isolation in centrifugal pumps*. Unpublished doctoral dissertation, Aalborg University.
- Katayama, T. (2005). *Subspace Methods for System Identification*. Springer.
- Luo, J., Pattipati, K. R., Qiao, L., & Chigusa, S. (2008, September). Model-based prognostic techniques applied to a suspension system. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 38(5), 1156 -1168.
- Lyshevski, S. E. (1999). *Electromechanical Systems, Electric Machines, and Applied Mechatronics*. CRC.
- Overschee, P., & Moor, B. D. (1996). *Subspace Identification for Linear Systems*. Boston, MA, USA: Kluwer Academic Publishers.
- Pulido, B., & Alonso-González, C. (2004, October). Possi-

- ble Conflicts: a compilation technique for consistency-based diagnosis. *IEEE Trans. on Systems, Man, and Cybernetics. Part B: Cybernetics*, 34(5), 2192-2206.
- Roychoudhury, I., Biswas, G., & Koutsoukos, X. (2009, December). Factoring dynamic Bayesian networks based on structural observability. In *Proc. of the 48th IEEE Conference on Decision and Control* (p. 244-250).
- Saha, B., & Goebel, K. (2009, September). Modeling Li-ion battery capacity depletion in a particle filtering framework. In *Proceedings of the Annual Conference of the Prognostics and Health Management Society 2009*.
- Saha, B., Saha, S., & Goebel, K. (2009). A distributed prognostic health management architecture. In *Proceedings of the 2009 Conference of the Society for Machinery Failure Prevention Technology*.
- Saxena, A., Celaya, J., Saha, B., Saha, S., & Goebel, K. (2010). Metrics for offline evaluation of prognostic performance. *International Journal of Prognostics and Health Management*.
- Staroswiecki, M., & Declerck, P. (1989, July). Analytical redundancy in nonlinear interconnected systems by means of structural analysis. In *IFAC Symp. on Advanced Information Processing in Automatic Control*.
- Westwick, D., & Verhaegen, M. (1996). Identifying MIMO Wiener systems using subspace model identification methods. *Signal Processing*, 52(2), 235 - 258.
- Williams, B., & Millar, B. (1998). Decompositional model-based learning and its analogy to diagnosis. In *Proc. of the Fifteenth National Conference on Artificial Intelligence* (p. 197-204).

Experimental Polymer Bearing Health Estimation and Test Stand Benchmarking for Wave Energy Converters

Michael T. Koopmans¹, Stephen Meicke¹, Irem Y. Tumer¹, Robert Paasch¹

¹ Oregon State University, Corvallis, OR, 97331, USA

koopmans@engr.orst.edu

meickes@onid.orst.edu

irem.tumer@oregonstate.edu

paasch@engr.oregonstate.edu

ABSTRACT

Ocean waves can provide a renewable and secure energy supply to coastal residents around the world. Yet, to safely harness and convert the available energy, issues such as bearing reliability and maintainability need to be resolved. This paper presents the application of a Prognostics and Health Management (PHM) based research methodology to derive empirical models for estimating the wear of polymer bearings installed on wave energy converters. Forming the foundation of the approach is an applicable wave model, sample data set, and experimental test stand to impose loading conditions similar to that expected in real seas. The resulting wear rates were found to be linear and stable, enabling coarse health estimations of the bearing surface.

1. INTRODUCTION

Aggressive development of new energy resources for an ever growing human population is currently underway, and ocean waves have shown promise as a viable source of renewable energy. The interest in offshore power production is due in no small part to the proximity of consumers: over the next 15 years, 75% of the world's population is projected to live within 200 km of the coast (Hinrichsen, 1999), while the worldwide resource has been conservatively estimated to contain 200 - 500 GW of economically extractable energy (Cruz, 2008). Yet, designing, installing, operating, and maintaining systems to harness this renewable energy is an extremely complex problem from multiple standpoints. From an engineer's perspective, the most immediate and challenging problems revolve around device reliability and survivability within the marine environment.

Located in extremely energetic wave climates, a wave energy converter (WEC) is subjected to an array of loads and millions of oscillatory cycles per year. Depending on the device, certain components will deteriorate more rapidly than others, particularly the bearing surfaces that many WEC designs rely upon. Here, prognostic and health management (PHM) techniques can help create a strategy to cultivate information for predicting bearing degradation. These tech-

niques are important because often times the quality of the bearing surface directly affects the total cost of the device in terms not limited to 1) power take-off efficiency, 2) scheduled and/or non-scheduled maintenance, and 3) device survivability. Hence, the success of research efforts to assess and manage WEC reliability remains a critical step to the growth of the ocean renewable energy market.

Therefore, to help contextualize the problem and aid in WEC component-level experiments, system health research methods within the PHM community (Vachtsevanos, Lewis, Roemer, Hess, & Wu, 2006) were sought. A WEC's complexity, although not as involved as other complex systems such as aircraft, automobiles, or a submarine, is intensified with its naturally corrosive, brutal, and immense spectrum of marine operating conditions. Consequently, extensive and efficient use of laboratory experiments is needed to build the marine renewable community's database of seawater-based component life models. To populate this database, an accepted and scalable methodology is needed. This paper explores a proposed PHM research methodology (Uckun, Goebel, & Lucas, 2008) to lay the foundation for an experimental approach to measure bearing wear. More specifically, this study aims to assess the wear characteristics of polymer-based bearings immersed in seawater that are subject to loads and oscillations similar to those experienced by a point absorber WEC in real seas. Our investigation has three goals:

1. Verify and benchmark test stand design and operation for bearing wear measurements
2. Conduct wave energy research following a proposed PHM methodology
3. Present an initial study of polymer bearing health estimation utilizing wear models derived from a set of generalized representative sea states

1.1 Main Contributions of the Paper

The work presented here is the beginning of a larger research effort to assess and manage WEC reliability, maintainability, and overall system health using PHM based techniques. Beginning with the bearing design and operating effects, accurate material wear models become critical in determining the efficiency of the device power output. The contributions of this study are itemized as follows:

Michael T. Koopmans et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

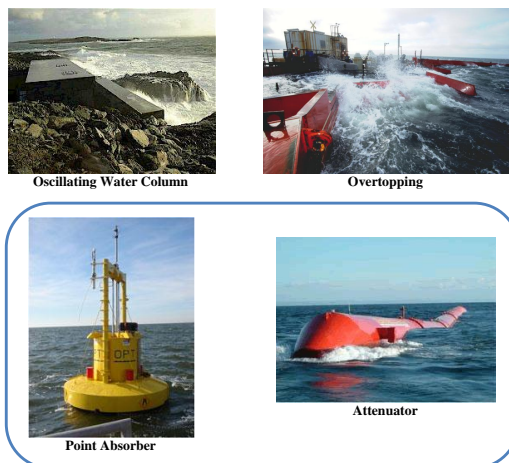


Figure 1: Oscillating wave energy converter devices.

- A PHM based methodology was used to determine polymer bearing wear models with respect to their pressure and velocity parameters in seawater;
- Wave climate load classification was detailed for a point absorber WEC in generalized real seas;
- Cumulative wear of proposed bearing material was estimated for a given month;
- Relevant information was provided to ocean renewable developers and partners to help assess the applicability of the materials and improve the technology;
- An experimental test stand's performance was benchmarked and recommendations were offered for future bearing tests.

1.2 Roadmap

The paper will begin with a brief background section, including an introduction to the point absorber WEC, application assumptions, and an overview of the PHM research method. Next, the wave climate and the process used to determine experimental wave cases are discussed, followed by a description of the experimental setup. Results of the bearing wear tests, their implications, and future studies are also presented.

2. BACKGROUND

This section provides a brief description of the chosen wave energy converter (WEC), test stand effects, and modeling considerations. To begin, there are generally four main groups of WEC designs: oscillating water columns, overtopping devices, point absorbers, and attenuators (Fig. 1) (Ocean Power Technologies, 2011; Wave Dragon, 2011; Wavegen, 2011). Each device relies on bearings to either support a turbine shaft (water columns and overtopping) or provide a sliding surface on which two large masses can move relative to each other (Yemm, 2003). Specifically, the point absorber and attenuator WECs are designed to harvest the heave motion of a passing wave through their power take-off (linearly or rotationally), where the relative motion of two or more large masses is exploited to generate electricity. Other examples of seawater exposed bearing applications include wind plus wave energy harvesters (Floating Power Plant, 2011) and sea floor based rotational power take-offs (Aquamarine Power, 2011).

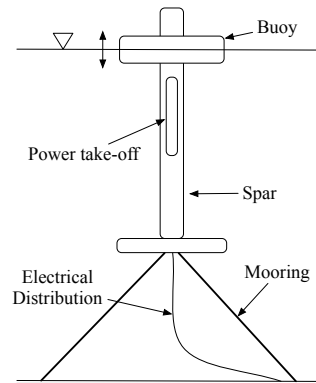


Figure 2: A generic linear power take-off point absorber WEC architecture layout, where relative motion between the buoy and spar provide energy conversion opportunities.

2.1 The Point Absorber

Focusing on the point absorber design, the system contains a few core subsystems: power take-off, mooring, structures, control, and distribution (Fig. 2). The device is capable of translating in three degrees: heave (up/down), surge (forward/back), sway (starboard/port) and rotating three degrees about its axis: pitch, yaw, and roll. This investigation will only consider the structures subsystem of a point absorber WEC (buoy and spar) and its heave dynamics with respect to the sea floor. Power take-off, mooring, and control do play very important roles in the loading conditions of the bearing surface, albeit require much more knowledge about the WEC system itself and is not covered in this paper. Essentially, this study assumes one degree of freedom (heave) and a float that is a perfect wave follower. In other words, when solving for the heave dynamics, it will be assumed that as each wave passes, the buoy will travel up and down with the water surface. This relative velocity between buoy and spar is the assumed velocity the bearing surface will experience during operation (i.e., power generation). In storms however, the WEC is most likely not converting energy and may switch to a survivability mode; one possible technique locks the buoy in place to impede system damage.

The bearing subsystem is integrated into the structure of the WEC and provides a surface on which the buoy and spar may move relative to each other. To avoid installing a bearing material sleeve along the entire inner diameter of the buoy, one possible solution lays two to four equally-spaced counterface extrusions around the spar, where they are mated with bearing blocks impregnated within the buoy. Here, the bearing requirements for many WEC technologies demand the surface to be inexpensive, corrosion-resistant, low maintenance, and near-zero friction in a large variety of loading conditions. One proposed solution utilizes a polymer-based approach, similar to those found in current naval designs (Cowper, Kolomojcev, Danahy, & Happe, 2006) and hydropower applications (McCarthy & Glavatskih, 2009; Ren & Muschta, 2010).

This simple polymer-based approach has proven to be beneficial in such applications for its ability to self-lubricate and deposit a transfer film on the counterface, filling in surface asperities, linearizing the wear rate, and even reducing friction in some cases (Wang, Yan, & Xue, 2009). However, water's tendency of inhibiting or wholly preventing transfer film formation is a research topic itself and will only be indi-

rectly addressed in this work. Research regarding wear characterization of polymer journal bearings has been published at various pressures, velocities, and environmental conditions (Ginzburg, Tochil'nikov, Bakhareva, & Kireenko, 2006; Rymuza, 1990); yet, few studies have been shared with the wave energy community presenting the results of seawater immersion (W.D. Craig, 1964; Tsuyoshi, Kunihiro, Noriyuki, Shozo, & Keisuke, 2005), let alone under pressures and velocities expected to be experienced by WECs (Caraher, Chick, & Mueller, 2008). So, with an immature technology being relied upon by a large complex system, an experimental test stand has been designed and used to procure knowledge about the bearing material's performance characteristics under representative loading conditions.

2.2 PHM Based Techniques

As previously mentioned, the research methodology born in the PHM community provides a good platform on which test stand research can be integrated into a larger, more comprehensive effort to assess system health. A general outline is shown in Fig. 3, where the path to implementing and relying upon a prognostic solution begins first with high-level system requirements (health predictions for subsystems and/or the system itself) that define the subsequent metric, fault, and sensor selection process. Next, the third step determines the most appropriate approach in terms of desired performance, available resources, and acceptable uncertainty to satisfy the component-level predictions. Here the proper number of samples to sacrifice for an accurate inference is also set. The fourth step ascertains the test scenarios, design of experiments, and data collection, while the fifth step is dedicated to building models and remaining useful life algorithms for nominal and faulted conditions. The last two steps encompass the health estimation and actual usage comparisons, in addition to the verification and validation sequence. A good application of the entire PHM research methodology was presented in estimating battery capacitance over time using high quality test chambers (Goebel, Saha, Saxena, Celaya, & Christophersen, 2008). For this work however, only a few steps of the methodology are addressed for estimating WEC bearing wear. Knowing that it would be useful to predict bearing wear in extreme marine conditions, the initial strategy to determine adequate experimental conditions and data collection procedures is described in addition to how the test stand itself contributes to the main goals of this investigation.

2.3 Test Stand Considerations

The test stand design and operation are critical to the validity of the empirical bearing wear models. Many interested researchers have built test stands to measure the degradation of particular components, including batteries (Saha, Goebel, Poll, & Christophersen, 2009), actuators (Balaban et al., 2010; Bodden, Clements, Schley, & Jenney, 2007), and polymer bearings (Gawarkiewicz & Wasilczuk, 2007). The particular test stand employed for the experiments presented in this paper is a modified version of American Society for Testing and Materials' (ASTM) standard test for ranking plastic resistance (ASTM, 2009), where the major changes to the standard include an oscillatory velocity, varying loads, and immersing the sample in seawater. Being a relatively new field of research, a lack of verification and validation of the modified test stand contributes to the uncertainty of the results. A goal of this work is to verify and benchmark test stand design and operation, ensuring the bearing wear measured repeatedly and accurately reflects imposed loading conditions.

2.4 Modeling Considerations

When investigating and modeling polymer bearing wear, it is important to note that multiple factors contribute to the wear rate. A polymer bearing / counterface tribosystem failure modes and effects analysis may contain only a few failure causes, where a primary failure would be the direct result of the physical amount of bearing material removed, and secondary failures may be attributed to biofouling or sediment-rich seawater. This study only covers the primary failure (wear) and does not address secondary failures. Also, a wear estimation is considered synonymous with a bearing health estimation because the bearing's ability to perform as designed is assumed to be directly attributed to the physical amount of material remaining in place.

One must also consider the naturally stochastic ocean waves. Their modeling effort has been well documented (Tucker & Pitt, 2001; Holthuijsen, 2007; Young, 1999) and the trade-off between the relevance of a higher fidelity numerical model and a closed-form solution must be done. For this work, the mapping of sea state to bearing pressure and velocity will be solved analytically with several conservative assumptions (e.g., linear waves, buoy / spar dynamics) that serve well as an initial attempt to assess the applicability of this research.

3. THE WAVE CLIMATE

Within the fourth step of the PHM methodology, expected sea states are sought to derive the pressures and velocities experienced by the bearing surface. In order to choose experimental cases representative of WEC oscillations and loads, a wave climate comparable to permitted sites was chosen (FERC, 2011). A wave climate is defined here as the aggregation of all the reported wave measurements taken at a specific location. The most accessible sources for past wave climate information include the Coastal Data Information Program (CDIP, 2011) and the National Data Buoy Center (NDBC, 2011) who manages a worldwide buoy network. A buoy of particular interest for its similarities to a potential WEC installation (proximity to coast / large population areas, consistent and predictable wave energy) is located 15.5 nautical miles northwest of Winchester Bay, Oregon (NDBC station ID: 46229), where the water depth is 186 meters and the buoy is assumed to be a perfect wave follower.

3.1 Wave Data

Wave information is often reported in the frequency domain as a wave spectrum, where, for each frequency and respective bandwidth, the energy or wave energy density is registered (Tucker, 1991). Other parameters included in the report can denote the wave direction, depending on the buoy. Much more wave data is also available apart from the spectral information, including the raw time series values, which is used for much higher fidelity WEC modeling. For the purpose of this study however, only two parameters were used in defining the wave climate: significant wave height (H_s) and dominant wave period (T_D). The significant wave height (in meters) is the average of the highest one-third of all the wave heights encountered during the 20 minute sampling period. The dominant wave period (in seconds) is the period with maximum wave energy as taken from the wave spectrum over the sampling period (Steele & Mettlach, 1993).

3.2 The Sample Data Set

Significant wave heights and dominant wave periods were taken for years 2005 - 2010 (NDBC, 2011). Reporting data

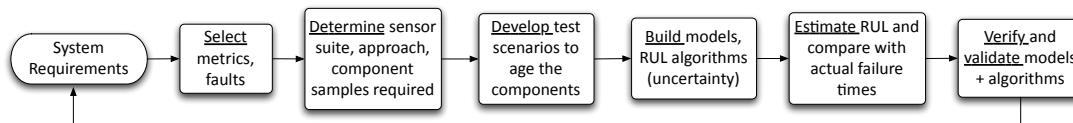


Figure 3: A universal PHM research methodology.

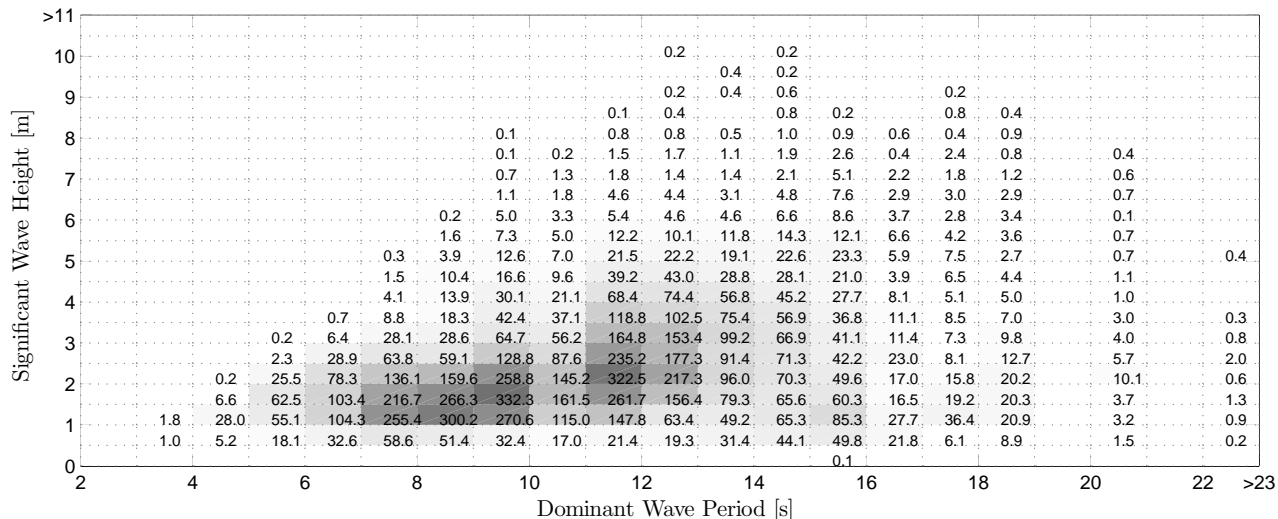


Figure 4: The total wave climate, where each bin contains the average number of hours for each sea state for an average year during the past six years (2005 - 2010).

every hour until the end of January 2008, the sampling rate was increased to every half hour. The entire data set is not complete, as some data points are erroneous (e.g., $H_s = T_D = 99$) or absent altogether. To include some of these reports in the sample data set, the erroneous points were replaced with the average of their nearest neighbors, whereas the absent points were left out of the averaging process. No weighting was installed to unbias the months with more hours reported over the months with lesser hours reported. There were four major gaps in the data, where no reports were given for the following dates: 1/1/05 - 4/1/05, 2/25/06 - 5/11/06, 5/29/06 - 7/13/06, and 3/16/09 - 4/1/09. Three of the four gaps occur in the spring and summer, while the largest consecutive gap occurs in the winter. This may be due to a more energetic sea state during these months causing system failures. Overall, the six years of coverage yielded only 5.06 years of data. This fact affects the total wave climate picture in terms of number of hours per particular sea state, but for the purpose of choosing test wave parameters, it is not foreseen to affect the results of this study. Therefore, the data set from which the experimental cases were determined can be seen in Fig. 4, where each bin covers one second wave periods and half meter wave heights with the average number of hours reported for that bin over the measured time period displayed in the plot. The most common sea state was an 9 - 10 second period and 1.5 - 2.0 meter wave height, accounting for approximately 3.8% of the yearly total.

3.3 Choosing Experimental Cases

In order to effectively achieve a spread of experimental cases, the wave period distribution was analyzed as shown in Fig. 5 while the wave heights were taken at each period interval. An

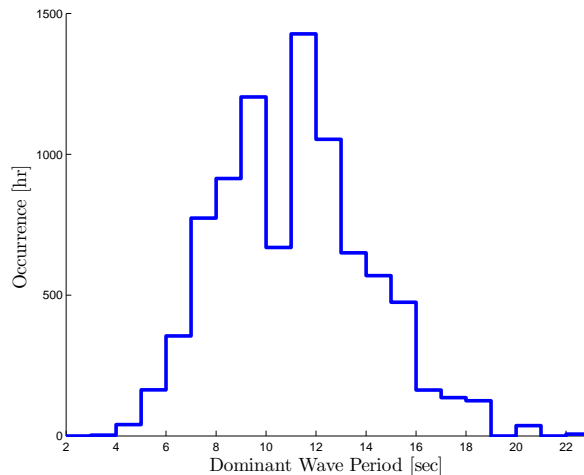


Figure 5: Wave period distribution over the entire climate data set, with an average of 10.89 sec and a standard deviation of 2.95 sec.

interval is defined here as a particular one second period bin determined by the average and standard deviation of the cumulative wave period distribution where the column of wave heights is then sampled to find the exact experimental case (i.e., H and T). For the test period of 10.89 sec, the 10 - 11 sec period bin was analyzed (Fig. 6), as were the other three test period bins (7 - 8 sec, 13 - 14 sec, and 16 - 17 sec) to achieve all four experimental cases (Tbl. 1).

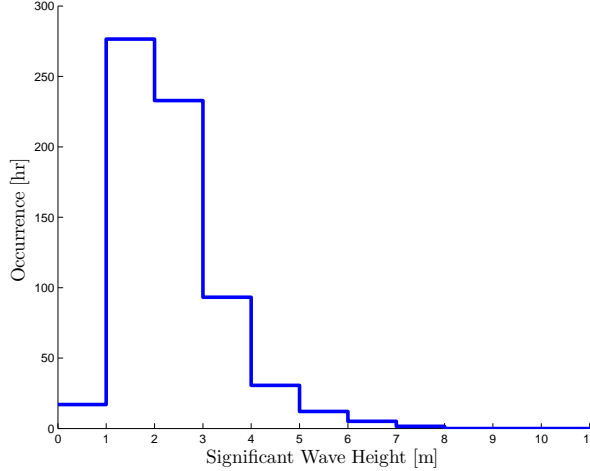


Figure 6: Distributions of significant wave heights for the 10 -11 sec period bin with an average of 2.3 m and a standard deviation of 1.0 m.

Exp. Case	T (s)	H (m)
1	10.89	2.31
2	13.84	5.51
3	16.79	2.92
4	7.95	1.74

Table 1: Chosen test wave heights and periods.

4. EXPERIMENTAL DESIGN

This section explains the design decisions and limitations behind the bearing wear experiments and their corresponding parameters, including the bearing health estimation algorithm (addressing step three and parts of step five of the PHM methodology). Knowing the experimental wave parameters, the calculation of pressures and velocities at the surface of interest is described. First, a description of the procedure to compute the loading condition input for each bearing wear experiment is presented, followed by a table containing each experimental case parameter. Many assumptions support the closed-form procedure taken in this paper and will be discussed as they are applied.

4.1 Wave Modeling and Force Calculation

First, the wave experienced by the WEC is classified using four main parameters: water depth h , wave height H , wave length L , and wave period T (Fig. 7), where η describes the wave surface elevation in terms of x and t while having a value of z meters. The wave itself is assumed to be harmonic and linear (or regular); other wave classifications include irregular, ocean, and stochastic ocean waves (Ochi, 1998). Generalizing the sea state under linear wave theory is the most basic approach to modeling the ocean surface and is deemed appropriate for this initial study.

The generalization assumes the fluid to be incompressible and inviscid (irrotational), enabling the local water particle velocities to be solved explicitly and facilitating the use of Morison's equation (Dean & Dalrymple, 1991). In a typical design, a software program is tasked with computing the

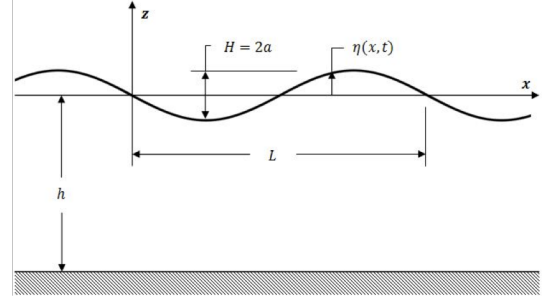


Figure 7: A regular two dimensional wave with relevant parameters and coordinate system shown.

structural loading (e.g., AQWA, WAMIT). However, in our case, the Morison equation will be shown as an initial approach to calculate bearing pressure.

Next, assuming an intermediate water depth, the wave length is solved numerically using Eq. 1, where g is the acceleration due to gravity. A water depth of 91.4 meters was used in this study to mimic Oregon sites where WEC developers currently hold permits (FERC, 2011).

$$L = \frac{g}{2\pi} T^2 \tanh \frac{2\pi h}{L} \quad (1)$$

The wave length can be verified for use in an intermediate water depth by checking the inequality (Eq. 2), where the wave number is $k = \frac{2\pi}{L}$. When calculating a kh scalar towards the lower or upper extremes, a shallow or deep water assumption, respectively, would instead prove more accurate.

$$\frac{\pi}{10} < kh < \pi \quad (2)$$

Next, the water surface displacement, η , is given in Eq. 3, where $\sigma = \frac{2\pi}{T}$ and its correlated velocity potential, ϕ , is given in Eq. 4.

$$\eta(x, t) = \frac{H}{2} \cos(kx - \sigma t) \quad (3)$$

$$\phi = -\frac{gH}{2\sigma} \frac{\cosh k(h+z)}{\cosh kh} \sin(kx - \sigma t) \quad (4)$$

The closed-form velocity potential allows for the calculation of horizontal ($-\frac{\partial\phi}{\partial x}$) and vertical ($-\frac{\partial\phi}{\partial z}$) water particle velocities, which can be seen in Eq. 5 and Eq. 6, respectively.

$$u = -\frac{\partial\phi}{\partial x} = \frac{gHk}{2\sigma} \frac{\cosh k(h+z)}{\cosh kh} \cos(kx - \sigma t) \quad (5)$$

$$v = -\frac{\partial\phi}{\partial z} = \frac{H\sigma}{2} \frac{\sinh k(h+z)}{\sinh kh} \sin(kx - \sigma t) \quad (6)$$

The local horizontal acceleration is shown in Eq. 7.

$$\frac{\partial u}{\partial t} = \frac{H\sigma^2}{2} \frac{\cosh k(h+z)}{\sinh kh} \sin(kx - \sigma t) \quad (7)$$

Using these equations, an estimation of the horizontal force imposed on the buoy by a passing wave can be computed.

Typically used to design and estimate loads on columns embedded in the sea floor, Morison's equation (Eq. 8) can be employed during conceptual WEC design for computing the horizontal wave force imparted on the device by a passing regular wave (Morison, O'Brien, Johnson, & Schaaf, 1950).

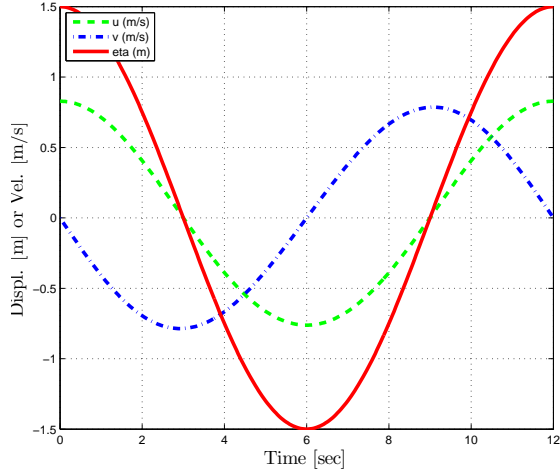


Figure 8: Example surface displacement and corresponding water particle velocities for a $H = 3$ m, $T = 12$ sec wave.

The equation is composed of two elements, the first captures the drag forces and the second captures the inertial forces,

$$F(z) = \frac{1}{2}C_D\rho Du|u| + C_M\rho V \frac{Du}{Dt} \quad (8)$$

where C_D , C_M , ρ , D , and V represent the drag & inertial coefficients, seawater density ($1025 \frac{\text{kg}}{\text{m}^3}$), buoy diameter, and buoy volume. Ultimately integrated over a water depth with respect to z , total horizontal force is represented in Eq. 9,

$$F_x = \int_a^b F(z) dz \quad (9)$$

where b is usually the water displacement(η), and a is some value in the vertical length (z) of the geometry. For example, if $a = -h$, the force would integrate over a continuous column to the sea floor. The aggregation of Eqs. 3 - 9 can be viewed in Fig. 8 and Fig. 9, where the parameters of a $H = 3$ m, $T = 12$ sec wave are plotted implementing the zero crossing method.

4.2 Experimental Case Parameters

Incorporating the above wave model, chosen wave heights and periods, and force calculations, the experiment case parameters can now be set (Tbl. 2). To reiterate, the experimental cases represent a first attempt at a sample set of representative wave parameters to classify polymer bearing wear during WEC operation. The third column states the maximum velocity the counterface experiences during the oscillatory profile (i.e., Eq. 6). Next, geometric assumptions that enable a specific velocity and pressure to be applied during wear tests are held and explained as follows. A buoy diameter of 11 m was used in the Morison force calculation while the force was integrated over a depth of 1.5 m. This depth was chosen based off the assumed buoy height (1.5 m) and assuming the buoy was fully submerged throughout the length of the passing wave. Next, knowing linear wave theory was being utilized, the drag and inertial coefficients were taken as 1.3 and 2.0, respectively (Agerschou & Edens, 1965). The bearing pressure was computed using the wave force calculation and an assumed bearing area of 0.232 m^2 . This particular

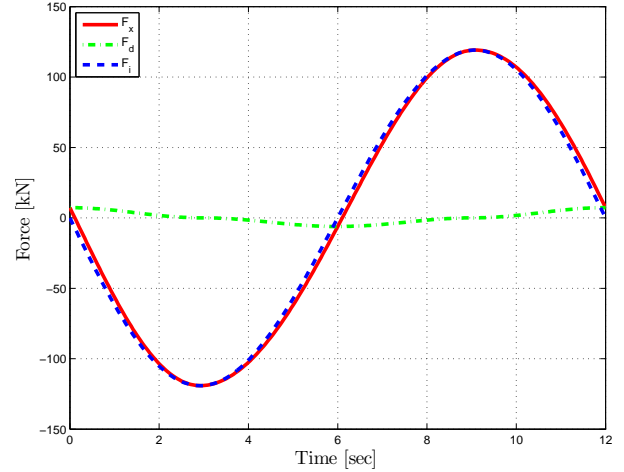


Figure 9: Example force oscillation imposed on the buoy by a passing $H = 3$ m, $T = 12$ sec wave, where F_d and F_i represent the individual components of F_x : the drag and inertial forces, respectively. The actual normal force applied to bearing sample was taken as the root mean squared value of the maximum F_x due to test stand limitations.

area was chosen as a conservative estimate of the total bearing area and set the active bearing pressure below the bearing manufacturer's recommendations.

The final parameter set for the wear testing experiments was the number of runs for each experiment case. Using the operating characteristic (OC) curve to minimize the type II error, Eq. 3 was implemented (Montgomery, 2009),

$$\Phi^2 = \frac{nD^2}{2a\sigma^2} \quad (10)$$

where Φ and β (probability of type II error) make up the OC curve x and y parameters. Further, n is the number of runs for each test climate, D is the difference between two treatment means desired to be detected (0.5), a is the number of experimental cases (4), and σ is the assumed maximum standard deviation of wear rate at any power level (0.1). These values were based on previous wear studies completed. Tbl. 3 shows the results of checking various sample sizes and it was decided due to the infancy of this research that a probability of 0.85 would be adequate for detecting a difference in wear means (D) for separate experiment cases. Consequently, three test runs were specified for each experimental case.

n	Φ^2	Φ	$a(n-1)$	β	Power ($1-\beta$)
2	6.3	2.5	4	0.5	0.5
3	9.3	3.0	8	0.15	0.85
4	12.5	3.5	12	0.01	0.99

Table 3: Determining each experimental case's sample size using the operational characteristic curves with $\alpha = 0.01$.

4.3 Bearing Health Estimation

Once the bearing wear experiments have concluded, the post-processing of the raw linear variable differential transformer

Exp. Case	T (s)	H (m)	ν_{\max} (m/s)	F_{rms} (kN)	P (kPa)	kh
1	10.89	2.31	0.66	78	334	3.1
2	13.84	5.51	1.25	120	500	2.0
3	16.79	2.92	0.55	47	202	1.4
4	7.95	1.74	0.69	108	445	5.8

Table 2: Experiment Case Parameters

(LVDT) measurements should ideally indicate a linear and stable wear rate. Under these circumstances, the wear models can be pieced together to create a cumulative data driven life model of the bearing surface. This inference allows ocean renewable developers the capability to predict the bearing's health after some length of time. For example, if the life model indicates the amount of bearing material departed is approaching a critical threshold, then operators and maintainers can make informed decisions. Given enough time, the repairs could be scheduled to minimize the cost associated with servicing the bearings. It is important to note that the prediction accuracy of the bearing health estimation is directly attributed to wear model quality and its associated experimental design.

In order to quantify the raw bearing wear in a format applicable to wear predictions, the recorded vertical wear from the LVDT is multiplied by the constant contact area to form the total volumetric wear for the sample seen in Eq. 11,

$$V = 2wrq \sin^{-1}\left(\frac{l}{2r}\right) \quad (11)$$

where w is the vertical wear, r is the counterface outer radius, l is the sample length, and q is the sample width (all variables in mm). To avoid biasing the wear estimate to focus on force or distance or time alone, a specific wear rate variable is used (Eq. 12),

$$V = eFs \quad (12)$$

where V is the total volumetric wear (mm^3), e is the specific wear rate ($\frac{\text{mm}^3}{\text{Nm}}$), F is the normal load (N), and s is the sliding distance (m). Solving for e using the stable portion of the wear plot, a set of specific wear rates are then available to the user for calculating volumetric wear of the bearing during different climates than those tested in the experiment. Assuming the worst case scenario for the specific wear rate model formulation, forces and sliding distances are derived for each particular hour of reported wave parameters. The cumulative volumetric bearing wear is tracked using Eq. 13,

$$\sum_{i=0}^m V_i c_i \quad (13)$$

where i is the bin index (wave height and period), m is the number of discrete sea states reported during the time interval, V is the volumetric wear associated with a particular bin and c is the total number of hours the WEC experienced seas classified to the particular bin. This purely data driven model would preferably be used in parallel with the wave climate in Fig. 4 and although relatively elementary, could be enormously useful in estimating the overall bearing health, while further informing WEC design, operation, and maintenance decisions.

5. EXPERIMENTAL SETUP

This section describes the bearing material and its mating counterface used during this study - addressing step four of the PHM methodology. The test stand is also shown and the procedure to measure bearing wear is described.

5.1 Bearing Material

Each bearing sample was machined out of disks (with an inner radius equal to the counterface) 6.40 mm in width into sections of 15.85 mm in length and approximately 10 mm in height. The Thordon SXL bearing material was used throughout the study (Thordon, 2011). Each bearing sample was cleaned with methanol prior to each test to guard against any incidental debris from contaminating the experiment.

5.2 Counterface

Two identical 316 stainless steel counterfaces were used during testing, each with a diameter of 63.5 mm (derived from the rpm limit of the motor so as to maximize the range of test surface velocities) as seen in Fig. 10. Before and after each test run, the surface roughness of the counterface was measured using a Mitutoyo surface roughness tester in an attempt to determine any transfer of material to the counterface. As per design recommendations from the manufacturer, the counterface surface roughness was made to be less than 0.8 μm Ra before each test. In an effort to allow for better mechanical bonding of the polymer, roughening was completed perpendicular to the direction of rotation (Marcus, Ball, & Allen, 1991). The roughness measurements were taken in parallel to the direction of rotation at three different points along the width of the counterface and six different section widths around the circle. Prior to each test, the counterface was also thoroughly cleaned with methanol.

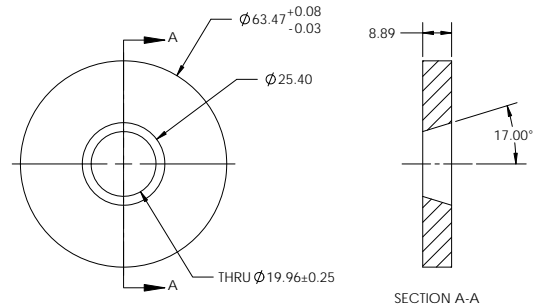


Figure 10: Counterface dimensions.

5.3 Test Stand

Implementing a testing method derived from the ASTM G176-03 standard (ASTM, 2009) for ranking wear resistance

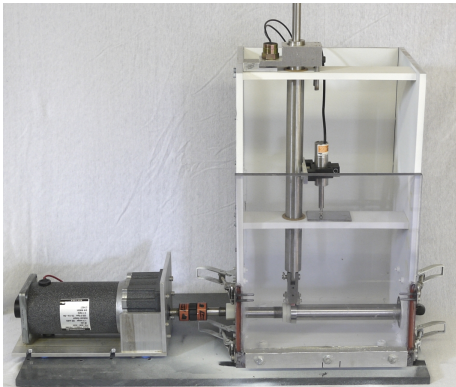


Figure 11: The bearing wear test stand.

of plastic bearings, the test stand can be seen in Fig. 11. Modifications to the setup have been made to allow for complete immersion of both the bearing sample and counterface in seawater. A procedure to run a bearing wear test follows:

1. Empty all seawater from reservoir and wash out with freshwater, lightly touching the counterface (remove salt, but not the transferred bearing material) and remove bearing sample.
2. Remove the counterface from drive shaft, air dry, and measure surface roughness.
3. Take the second, prepped counterface and couple to drive shaft, ensuring minimum change in deflection of the surface during rotation. The authors recommend using a dial indicator to measure this deflection.
4. Set the new, prepared bearing material in place, load mass on vertical shaft, latch front plate, fill reservoir, input test parameters to software, and begin test.

The removable counterface is held in place with two plastic nuts on a stainless steel drive shaft directly coupled to a DC brushed motor. A $0.5 \mu\text{m}$ resolution LVDT was tasked with measuring the vertical wear of the bearing sample while linked to the vertical shaft responsible for holding the mass load in place. The drive shaft and all connecting parts were cleaned with methanol prior to each test. The seawater used during testing is seawater filtered to $50 \mu\text{m}$, taken from Yaquina Bay in Newport, Oregon.

A National Instruments cRIO unit was programmed to control motor velocity using the LabVIEW interface and shaft encoder relaying speed information. The bearing samples were subjected to sinusoidal velocity profiles (ν) oscillating at their specified frequency ($\frac{1}{T}$) and each wear test was run for 20 hours with no intermittent stops. In order to determine the correct mass to load the sample, the test climate pressure (P) was multiplied by the bearing sample projected area and divided by the gravity constant, g .

6. RESULTS

This section presents the results of all twelve wear tests, grouped into their four respective experiment cases, followed by the specific wear rate model formulation, a month long bearing health estimation, and the corresponding before and after counterface surface roughness measurements. The raw LVDT readout was smoothed for graphing purposes. Each wear plot contains two x-axes: sliding distance (computed

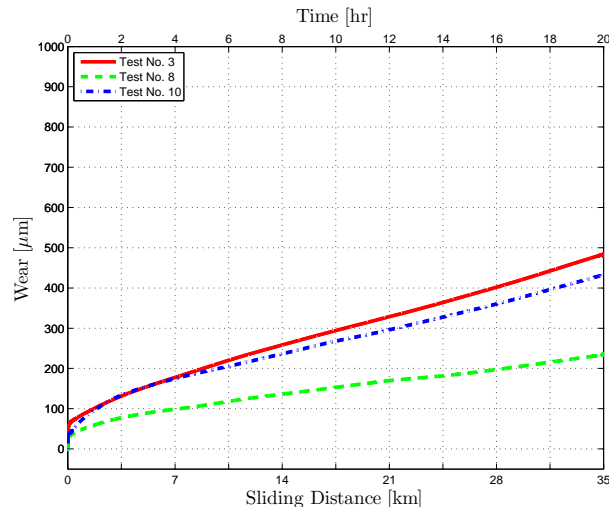


Figure 12: Experiment case one, pressure = 334 kPa, maximum surface velocity = 0.66 m/s, mass = 3.382 kg.

from oscillation frequency, amplitude, and counterface radius) and time (each wear test was 20 hours long). The first case is shown in Fig. 12, while the second, third, and fourth cases are shown in Figs. 13 - 15, respectively. The plots show the highest pressure resulted in the highest wear rate, while the lowest pressure resulted in the lowest wear rate, as expected. And for the majority of test runs, similar patterns exist within each experimental case. However, test run number twelve is an anomaly: around hour seven, the wear measurement diverges and increases 350% less than the previous two test runs. Another test run that is unlike its counterparts is number eight, where its wear measurements are offset 50 - 100% less than complementary tests three and ten.

Next, to ensure wear is linear with respect to time and distance, hour six to twenty was set as the stable portion of the wear plot for all test runs. Analyzing this segment, a vertical bearing wear measurement can be used to derive the total volumetric wear and specific wear rate for each test run. Here the results can be seen in Fig. 16, where the dotted line represents a worst case scenario specific wear rate model. For a month long wear estimation, the specific wear rate model was used, where the volumetric wear for each hour of reported wave data was calculated using 1) a specific wear rate, e , from the model, 2) a normal force, F , derived from Morison's equation, and 3) a sliding distance, s , derived from the particular climate's reported wave parameters. For the month of January 2011, a total of 4.5 mm was estimated to have been lost during the theoretical point absorber WEC operation (Fig. 17). Additional information was recorded before and after each test run that included the counterface surface roughness measurements (Tbl. 4).

7. DISCUSSION

Upon completing the experiments for this study, the wear plots show the bearing material's performance is dependent on a few external factors including, a direct correlation with the loading conditions and a peculiar association with counterface preparation. The test stand was shown to operate reliably throughout the investigation, however it too affects the wear rate indirectly through load application and velocity control attributes. Further exploring the findings, this section

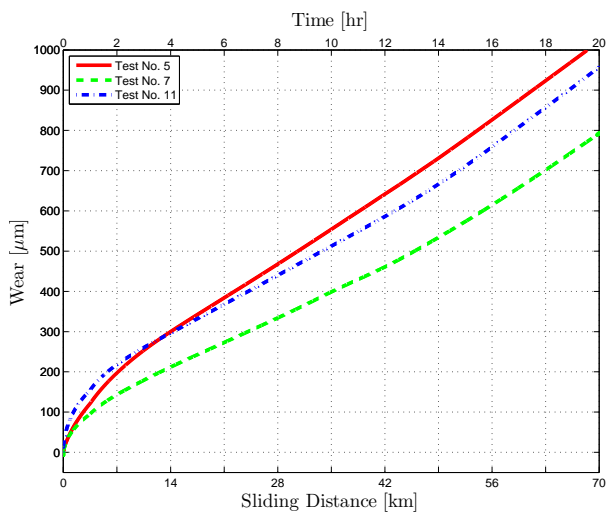


Figure 13: Experiment case two, pressure = 500 kPa, maximum surface velocity = 0.1.25 m/s, mass = 5.000 kg.

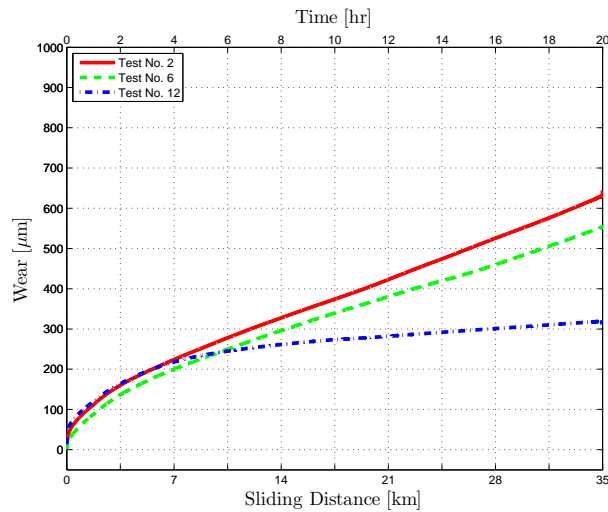


Figure 15: Experiment case four, pressure = 445 kPa, maximum surface velocity = 0.69 m/s, mass = 4.442 kg.

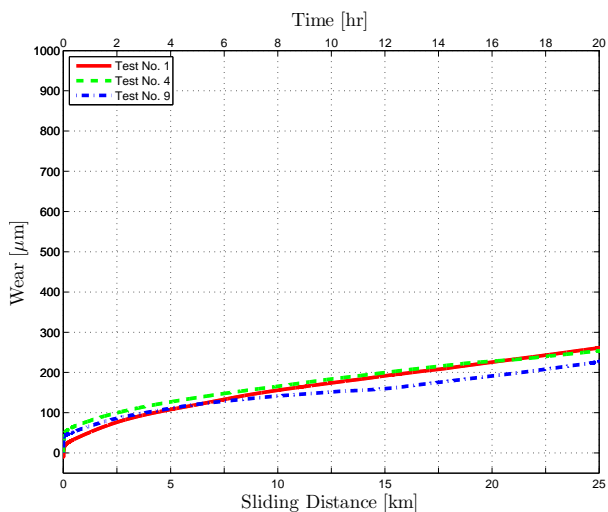


Figure 14: Experiment case three, pressure = 202 kPa, maximum surface velocity = 0.55 m/s, mass = 2.045 kg.

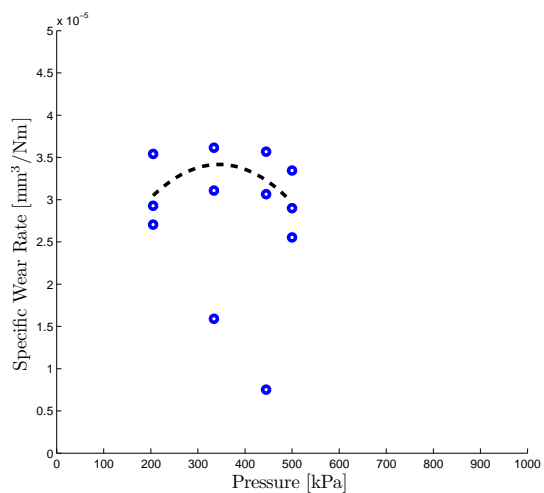


Figure 16: Specific wear rates plotted vs. applied bearing pressure for all twelve test runs with the conservative model overlay (dotted line).

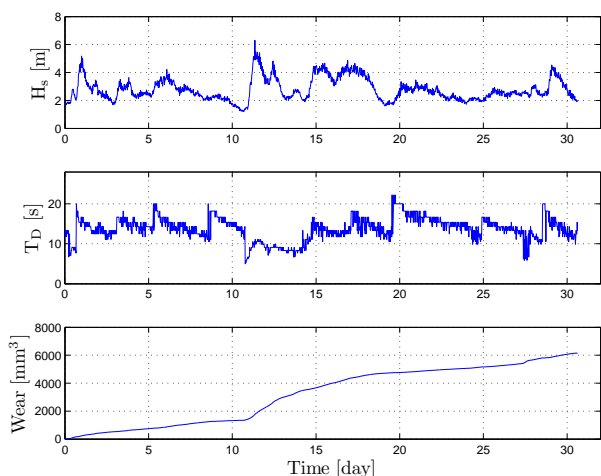


Figure 17: An example wear estimation for the month of January 2011.

Exp. Case	Test No.	Rate ($\frac{\mu\text{m}}{\text{hr}}$)	Before ($\mu\text{m Ra}$)	After ($\mu\text{m Ra}$)
1	3	18	.58 .69 .48	.71 .84 .56
	8	8	.61 .71 .48	.61 .79 .46
	10	16	.56 .69 .43	.53 .71 .38
2	5	45	.66 .81 .51	.63 .79 .41
	7	37	.61 .76 .53	.51 .64 .43
	11	42	.51 .58 .41	.48 .61 .33
3	1	9	.69 .76 .51	.58 .74 .43
	4	7	.69 .74 .58	.69 .79 .58
	9	8	.53 .69 .46	.53 .66 .41
4	2	25	.74 .79 .69	.64 .74 .58
	6	21	.66 .76 .58	.71 .94 .51
	12	5	.61 .71 .53	.56 .69 .43

Table 4: Stable wear rates for each test run and their corresponding before and after surface roughness measurements (average, maximum, minimum).

discusses several factors contributing to the uncertainty in the results. Topics affecting the accuracy of the prediction include the effect of counterface surface roughness, wave modeling, wear data quality, and test stand effects.

7.1 Effect of Counterface Roughness

To begin, the effect of surface roughness on the stable wear rate is plainly apparent and as one would expect, a higher roughness generally yields a higher wear rate. Observing experiment case four in particular, test twelve yielded a stable wear rate 4 - 5 times smaller with a pre-test surface roughness less than $0.06 \mu\text{m}$ smoother than test two or six. Perhaps this result is specific to the experiment (relatively high pressure and frequency oscillation) as the difference between pre-test roughness measurements for experiment cases two and three are similar, yet their subsequent stable wear rates are analogous. It should be noted that there are limits as to how smooth the initial counterface can be as one study showed a roughness of $0.03 \mu\text{m}$ increased the wear rate by 33 times (Marcus & Allen, 1994). Experiment case one and four both contain a test run dissimilar to the others while their pre-test roughness measurement differences are negligible, indicating that there may be some other factor affecting the results and warranting more experiments.

From previous experience, the bearing material studied exhibited an unusually higher wear rate for their respective loading conditions in the majority of test runs. Acknowledging the customization of the experimental design and operation, the obvious absence of a transfer film may indicate the need for a better application of pressure and velocity to the bearing sample itself via a different test stand design and/or operation.

7.2 Wave Modeling

Second, the method of wave modeling used in this investigation assumes a regular wave, which is not an accurate representation of real seas. Propagating linear waves and the assumption of the buoy being a perfect wave follower are likely the most influential assumptions within this study. The most rigorous of ocean wave modeling efforts solve the Navier Stokes non-linear differential equation for velocities and pressure fields, yet is only suggested for higher fidelity investigations. However, the success of applying the often used principle of superposition (as many frequency domain wave models do) to the wear rates remains to be seen given the limitations of linear wave theory (Young, 1999). Another, more promising strategy would be to utilize the WEC dynamics derived from previous modeling efforts (Ruehl, Brekken, Bosma, & Paasch, 2010).

Further, choosing NDBC 46229 as the source of ocean surface measurements was designed to allow researchers the freedom of employing either a time or frequency domain based approach. Also, for a more complete input to the wave climate, the authors suggest employing a method that explicitly presents representative wave spectra (Lenee-Bluhm, 2010).

7.3 Wear Data Quality and Health Estimations

Third, the health estimation, however unrefined, was possible because of quality wear data. The empirical models yielded few extraordinary anomalies and provided a good basis for regression and validation of the sample size suggestion. Applying the wear algorithm, approximately 6000 mm^3 , or 4.5 mm of bearing material was estimated to be lost during the month long WEC operation. This initial estimate is quite large and

could be attributed to several factors, including the material itself, the loading conditions chosen, the load application via the test stand's design and operation, and/or the counterface's surface roughness.

Also, a method for how to rectify the fact that wear rates do not exist for each bin within the wave climate has yet to be developed and would constitute a very interesting future work. Although the experiments do not change parameters during the 20 hour tests, future work would require the programming of varying parameters, resulting in more accurate loading conditions. Also, some of the next steps in this research would apply more advanced aspects of PHM by incorporating uncertainty assessment (Tang, Kacprzynski, Goebel, & Vachtsevanos, 2008) and prognostic evaluation (Saxena et al., 2008).

7.4 Test Stand Effects

Fourth, the effect of the test stand on the bearing experiments is inherent in the wear data, so only by modifying the test stand and running the same experiments would the effect be measurable. During testing, the motor was observed to jerk near the crest and trough of the sinusoid velocity profile, indicating poor torque control. This phenomenon occurred with greater intensity during experimental cases with higher pressures. To solve this problem, a torque sensor and high torque motor would be ideal additions to accurately and smoothly follow the desired velocity profile. Other test stand modifications to produce more accurate results would be the integration of a varying pressure function and time domain velocity profile. Currently, the test stand is limited to constant force application and only after running these initial experiments has it become readily obvious that the test stand is not capable of accurately recreating loading conditions that a bearing sample would see in the field - a much smoother control of the counterface velocity profile is required.

8. CONCLUSION

Twelve bearing wear experiments were conducted using a simplified wave model coupled with an average sea climate to derive representative loading conditions for polymer bearings installed on a point-absorbing WEC. Following a PHM based research method, a stable and linear wear rate was established for each experiment, leading to the use of empirical methods for estimating bearing wear. Not only was essential information gained regarding the limits of the experiments, but the actual research methodology as well. Much work remains, albeit progress was made towards careful benchmarking of the test stand and successful employment of PHM research tenets.

As a note, PHM is often an afterthought in complex system design because of many unanswered questions regarding prognostic requirements and their resulting validation sequence (Ferrell, 2010). This research focused on one component of the WEC and illuminated experimental attributes critical to its life predictions, even as developers work to install production-level devices where bearing health estimation may be the lowest of priorities. Promoting a scalable and technically sound approach to classifying WEC bearing performance early in the industrial development is significant, as benefits can quickly materialize for all parties.

ACKNOWLEDGMENTS

The authors wish to thank Thordon Bearings for supplying the bearing samples used in this study.

NOMENCLATURE

PHM	Prognostics and Health Management
WEC	Wave Energy Converter
T_D	dominant wave period
H_s	significant wave height
η	water surface displacement, a function of x and t
k	wave number
ϕ	water particle velocity potential
F_x	horizontal force imposed on buoy by passing wave
e	specific wear rate
i	bin index (wave height and wave period)
V_i	volumetric wear
c_i	total bin index hours

REFERENCES

- Agerschou, H., & Edens, J. (1965). Fifth and First Order Wave - Force Coefficients for Cylindrical Piles. In *ASCE Coastal Engr. Speciality Conf. Ch. 10. pp. 239*.
- Aquamarine Power. (2011). (<http://www.aquamarinepower.com/>)
- ASTM. (2009). G176 - 03(2009) Standard Test Method for Ranking Resistance of Plastics to Sliding Wear using Block-on-Ring Wear Test, Cumulative Wear Method. (<http://www.astm.org/Standards/G176.htm>)
- Balaban, E., Saxena, A., Narasimhan, S., Roychoudhury, I., Goebel, K. F., & Koopmans, M. T. (2010). Airborne Electro-Mechanical Actuator Test Stand for Development of Prognostic Health Management Systems. In *Annual Conference of the Prognostics and Health Management Society*.
- Bodden, D. S., Clements, N. S., Schley, B., & Jenney, G. D. (2007). Seeded Failure Testing and Analysis of an Electro-mechanical Actuator. In *IEEE*.
- Caraher, S., Chick, J., & Mueller, M. (2008). Investigation into Contact and Hydrostatic Bearings for use in Direct Drive Linear Generators in Submerged Wave Energy Converters. In *International Conference on Ocean Energy*. Brest, France.
- CDIP. (2011). *Coastal Data Information Program*. On the WWW. (<http://cdip.ucsd.edu/>)
- Cowper, D., Kolomojcev, A., Danahy, K., & Happe, J. (2006). USCG Polar Class Aft Sterntube Bearing Design Modifications. In *International Conference on Performance of Ships and Structures in Ice*. Banff, Canada.
- Cruz, J. (2008). *Ocean wave energy: current status and future perspectives*. Springer.
- Dean, R., & Dalrymple, R. (1991). *Water Wave Mechanics for Engineers and Scientists*. World Scientific.
- FERC. (2011). *Hydrokinetic Projects*. On the WWW. (<http://www.ferc.gov/industries/hydropower/indus-act/hydrokinetics.asp>)
- Ferrell, B. (2010). *Joint Strike Fighter Program Fielded Systems Presentation*. 2010 Annual Conference of the PHM Society.
- Floating Power Plant. (2011). (<http://floatingpowerplant.com/>)
- Gawarkiewicz, R., & Wasilczuk, M. (2007). Wear measurements of self-lubricating bearing materials in small oscillatory movement. *Wear*, 263, 458-462.

- Ginzburg, B., Tochil'nikov, D., Bakhareva, A., & Kireenko, O. (2006). Polymeric materials for water-lubricated plain bearings. *Russian Journal of Applied Chemistry*, 79, 695-706.
- Goebel, K., Saha, B., Saxena, A., Celaya, J., & Christophersen, J. (2008). Prognostics in Battery Health Management. In *Proc. IEEE*.
- Hinrichsen, D. (1999). *Coastal Waters of the World: Trends, Threats, and Strategies*. Island Press.
- Holthuijsen, L. (2007). *Waves in oceanic and coastal waters*. New York, Cambridge University Press.
- Lenee-Bluhm, P. (2010). *The Wave Energy Resource of the US Pacific Northwest*. MS Thesis, Oregon State University, Corvallis, OR.
- Marcus, K., & Allen, C. (1994). The sliding wear of ultrahigh molecular weight polyethylene in an aqueous environment. *Wear*, 178(2), 17-28.
- Marcus, K., Ball, A., & Allen, C. (1991). The effect of grinding direction on the nature of the transfer film formed during the sliding wear of ultrahigh molecular weight polyethylene against stainless steel. *Wear*, 151(2), 323-336.
- McCarthy, D. M. C., & Glavatskih, S. B. (2009). Assessment of polymer composites for hydrodynamic journal-bearing applications. *Lubrication Science*, 21(8), 331-341.
- Montgomery, D. (2009). *Design and Analysis of Experiments 7th ed.* John Wiley & Sons.
- Morison, J., O'Brien, M., Johnson, J., & Schaaf, S. (1950). The Force Exerted by Surface Waves on Piles. *Petroleum Transactions, AIME*, 189, 149-154.
- NDBC. (2011). *National Data Buoy Center*. On the WWW. (<http://www.ndbc.noaa.gov/>)
- Ocean Power Technologies. (2011). (<http://www.oceanpowertechnologies.com/>)
- Ochi, M. (1998). *Ocean waves*. Cambridge, Cambridge University Press.
- Ren, G., & Muschta, I. (2010). Challenging edge loading: A case for homogeneous polymer bearings for guidevanes. *International Journal on Hydropower and Dams*, 17(6), 121-125.
- Ruehl, K., Brekken, T., Bosma, B., & Paasch, R. (2010). Large-Scale Ocean Wave Energy Plant Modeling. In *IEEE CITERES*. Boston, MA.
- Rymuza, Z. (1990). Predicting wear in miniature steel-polymer journal bearings. *Wear*, 137(2), 211-249.
- Saha, B., Goebel, K., Poll, S., & Christophersen, J. (2009). Prognostics Methods for Batter Health Monitoring Using a Bayesian Framework. *IEEE Transactions on Instrumentation and Measurement*, 58(2), 291-296.
- Saxena, A., Celaya, J., Balaban, E., Goebel, K., Saha, B., Saha, S., et al. (2008). Metrics for Evaluating Performance Prognostic Techniques. In *International Conference on Prognostics and Health Management, Denver, CO*.
- Steele, K., & Mettlach, T. (1993). NDBC wave data - current and planned. In *Ocean Wave Measurement and Analysis - Proceedings of the Second International Symposium* (pp. ASCE, 198-207).
- Tang, L., Kacprzynski, G., Goebel, K., & Vachtsevanos, G. (2008). Methodologies for Uncertainty Management in Prognostics. In *IEEE Aerospace Conference*.
- Thordon. (2011). *SXL*. On the WWW. (<http://www.thordonbearings.com/clean-power-generation/tidalcurrentpower/design>)
- Tsuyoshi, K., Kunihiro, I., Noriyuki, H., Shozo, M., & Keisuke, M. (2005). Wear Characteristics of Oscillatory Sliding Bearing Materials in Seawater. *Journal of the Japan Institution of Marine Engineering*, 40(3), 402-407.
- Tucker, M. (1991). *Waves in Ocean Engineering: Measurement, Analysis, and Interpretation*. Ellis Horwood, LTD.
- Tucker, M., & Pitt, E. (2001). *Waves in ocean engineering*. Oxford, Elsevier.
- Uckun, S., Goebel, K., & Lucas, P. J. F. (2008). Standardizing Research Methods for Prognostics. In *International Conference on Prognostics and Health Management, Denver, CO*.
- Vachtsevanos, G., Lewis, F. L., Roemer, M., Hess, A., & Wu, B. (2006). *Intelligent Fault Diagnosis and Prognosis for Engineering Systems*. John Wiley and Sons.
- Wang, J., Yan, F., & Xue, Q. (2009). Tribological behavior of PTFE sliding against steel in sea water. *Wear*, 267, 1634-1641.
- Wave Dragon. (2011). (<http://www.wavedragon.net/>)
- Wavegen. (2011). (http://www.wavegen.co.uk/what_we_offer_limpet.htm)
- W.D. Craig, J. (1964). Operation of PTFE Bearings in Sea water. *Lubrication Engineering*, 20, 456-462.
- Yemm, R. (2003). *Pelamis WEC, Full Scale Joint System Test (Summary Report)*. Ocean Power Delivery Ltd. (http://www2.env.uea.ac.uk/gmmc/energy/energy-pdfs/pelamis-joint_system_test.pdf)
- Young, I. (1999). *Wind generated ocean waves*. Netherlands, Elsevier.

Michael T. Koopmans is a master's student and graduate research assistant at Oregon State University, where he is employed within the Complex Engineered Systems Design Laboratory. He earned a B.S. in Mechanical Engineering from California Polytechnic State University San Luis Obispo in 2009. He has held multiple summer internships at Aperio Technologies, Solar Turbines, NASA Ames Research Center, and most recently in the embedded reasoning area at Palo Alto Research Center. His research interests include actuator prognostics, integration of system health algorithms, and applying PHM techniques to wave energy devices. Mr. Koopmans is an ASME student member and CA certified Engineer in Training.

Stephen Meicke is a masters student in Mechanical Engineering under Dr. Bob Paasch, researching ocean wave energy. He completed his bachelors degree in Mechanical Engineering at Virginia Tech in Blacksburg, VA in May 2009. Steves research interests lie in stress analysis, fluid-structure interaction, and finite element analysis of floating structures. His research during his first year at OSU focused on the testing of marine journal bearings for use in wave energy converters (WECs). More recently, he intends to compare strain and hydrodynamic response data taken on Columbia Power Technologies 1:7 scale Sea Ray WEC deployed in the Puget Sound, near Seattle, Washington, to response estimates obtained from simulations using LS-DYNA, a multi-physics nonlinear finite element code .

Dr. Irem Y. Tumer is an Associate Professor at Oregon State University, where she leads the Complex Engineered System Design Laboratory. Her research focuses on the overall problem of designing highly complex and integrated engineering systems with reduced risk of failures, and developing formal methodologies and approaches for complex system design and analysis. Her expertise touches on topics such as risk-based design, systems engineering, function-based design, failure analysis, and model-based design. Since moving to Oregon State University in 2006, her funding has largely been through NSF, AFOSR, DARPA, and NASA. Prior to accepting a faculty position at OSU, Dr. Tumer led the Complex Systems Design and Engineering group in the Intelligent Systems Division at NASA Ames Research Center, where she worked from 1998 through 2006 as Research Scientist, Group Lead, and Program Manager. She received her Ph.D. in Mechanical Engineering from The University of Texas at Austin in 1998.

Dr. Robert Paasch is an Associate Professor in Mechanical Engineering, and the Boeing Professor of Mechanical Engineering Design. His current research interests include the design of mechanical systems for reliability and maintainability; knowledge-based monitoring and diagnosis of mechanical systems; applications of artificial intelligence for ecological systems monitoring; and design of marine renewable energy systems. He is a member of ASME, SAE, and received his PhD from the University of California, Berkeley in 1990. Dr. Paasch's research is supported by NSF, BPA, US Navy, and USDOE.

Experiments with Neural Networks as Prognostics Engines for Patient Physiological System Health Management

Peter K. Ghavami¹, Kailash C. Kapur²

^{1,2}Department of Industrial & Systems Engineering, University of Washington, Seattle, WA 98195, USA

pghavami@uw.edu
kkapur@uw.edu

ABSTRACT

Prognostics and prediction of patients' short term physiological health status is of critical importance in medicine because it affords medical interventions that prevent escalating medical complications. Accurate prediction of the patients' health status offers many benefits including faster recovery, lower medical costs and better clinical outcomes. This study proposes a prognostics engine to predict patient physiological status. The prognostics engine builds models from historical clinical data using neural network as its computational kernel. This study compared accuracy of various neural network models. Given the diversity of clinical data and disease conditions, no single model is ideal for all medical cases. Certain algorithms are more accurate than others depending on the type, amount and diversity of possible outcomes. Thus multiple neural network algorithms are necessary to build a generalizable prognostics engine. The study proposes using an oracle, an overseer program to select the most accurate predictive model that is most suited for a particular medical prediction among several neural network options.

1. INTRODUCTION

Prognostics and Health Management (PHM) is an engineering discipline that links studies of failure mechanisms to system lifecycle management (Uckun, Goebel, & Lucas, 2008). Other definitions of PHM describe it as a method that permits the assessment of the reliability of a system under its actual application conditions, to determine the advent of failure, and mitigate system risks (Pecht, 2008). A system can be broadly defined as an integrated set of elements that accomplish a defined objective (International Council on Systems Engineering Systems Engineering Handbook, 2000). The human body is a biological system that functions as a collection of interrelated systems. The question that we wish to answer is Peter Ghavami et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

how PHM can be applied to human biological systems as a methodology to predict and prevent adverse medical conditions in patients.

The term "diagnostics" pertains to the detection and isolation of faults or failures. "Prognostics" is the process of predicting a future state (of reliability) based on current and historic conditions (Vichare & Pecht, 2008).

The emphasis of this study is on prognostics (prediction) of the individual's short term future health condition and a rule-based prognostics engine that makes such predictions possible. Short term is defined as a time frame that spans from a few seconds to several days from any given moment. The prognostics engine is a computational component that can analyze vast amounts of historical and current physiological data and predict future health of an individual. Predictions are continuous over time as new, real time data are gathered from multiple physiological systems including warnings, alerts, events and precautions.

Admittedly developing mathematical models that make accurate predictions in biology and medicine is challenging but researchers suggest that soon such mathematical models will become a useful adjunct to laboratory experiment (and even clinical trials), and the provision of 'in silico' models will become routine (Smye & Clayton, 2002).

Advances in vital-signs monitoring software/hardware, sensor technology, miniaturization, wireless technology and storage allow recording and analysis of large physiological data in a timely fashion (Yu, Liu, McKenna, Reisner, Reifman, 2006). This provides both a challenge and an opportunity. The challenge is that the medical decision maker must sift through vast amount of data to make the appropriate treatment plan decisions. The opportunity is to analyze this large amount of data in real time to provide forecasts about the near term health state of the patient and assist with clinical decisions.

The application of PHM to human systems promises to deliver several benefits such as:

- Continuously assess the physiological and biological predictions to provide advance warning of clinical complications
- Minimize the frequency of reactive and emergency medical response by predicting and applying preventative medical interventions
- Enhance the quality of life and improve remaining useful life (RUL) of patients
- Manage the patient healthcare life cycle more effectively to improve patient care outcome and reduce medical care costs

Continuous and periodic monitoring of the individual's physiological systems involve collecting data from more than ten distinct human physiological sub-systems ranging from circulatory to respiratory to immune system. The collected data is used by the prognostics engine to predict future health of the individual.

The input data may consist of a wide array of clinically relevant information, medical history, allergies, medications, clinical procedures, genetic disposition and current physiological monitored data.

Medical science is grounded in scientific evidence, prior research, experiments and studies that have produced a body of medical knowledge based on generalizations and meta-analysis of research data. Such generalizations explain the causal relationships between risk factors, diseases and diagnosis. There are however gray areas in medical prognostics where many health treatment and screening decisions have no single 'best' choice or because there is scientific uncertainty about causes of certain diseases, or the clinical evidence is insufficient (O'Connor, Bennett, Stacey, Barry, Col, Eden, Entwistle & Fiset, 2009).

In many areas of medical science, the causal relationships are still incompletely understood and controversial. There are environmental, situational, cultural and unique factors that provide unique clinical data about a disease or groups of patients. Although this data is inadequate for making scientific generalizations and clinical evidence, it can provide valuable information to make assessments of individual's health status. The research hypothesis is that such data can be employed to make early predictions about the future health status of individuals and allow doctors apply medical interventions that prevent diseases or adverse medical events.

2. MARKERS AND PREDICTORS: THE CANARY AND THE DOG

The use of canaries for predicting failures has been discussed in literature as an illustrative example of prognostics (Vichare & Pecht, 2006). The analogy comes from an old mining practice. Canaries are more susceptible

to dangerous gases than humans. Since gases are not easily detected by humans, miners carried canaries to mines as an early indicator of dangerous gases. When the canary died, it signaled presence of dangerous gases and miners got out. The canary is an example of what in medicine is referred to as a marker (Souter, 2011).

For years, dogs have been trained to assist patients for medical purposes. Studies have shown that medical response dogs can be trained to predict and alert their owners of seizures before they occur (Brown & Strong, 1999, 2001, and Kirton, Winter, Wirrell & Snead, 2008). Other anecdotal studies claim that certain response dogs have been able to detect presence of melanoma cancer (Williams & Pembroke, 1989). A dog's ability to alert its owner of a pending seizure is an example of a prediction in medicine.

A prediction is a form of speculation about a state in the future. A *prediction* is foretelling a medical event or disease when the ingredients for that medical event are in place but have not combined to affect their significance in form of a disease yet. *Predictors* are variables that offer predictions about a disease. A *marker* is the recognition that the ingredients for a medical event are in place and have indeed combined to result in form of a disease but in lower and milder yet measurable doses (Souter, 2011).

The precursor to a disease is known as risk factors in medicine. Thus, a timeline of medical predictions starts with risk factors, leading to predictors (pre-disease state), and then on to markers (disease is in place but in low, mild state) and finally to the occurrence (onset) of the disease or medical event itself. Figure 1 illustrates the chronology of events and progression of the individual's health status from risk factors leading to the final disease or medical event manifestation. The distance between time ticks are arbitrary and vary between individuals.

Traditional medical prediction models rely on risk factors to make crude speculations about the patient's future health status. This research attempts to predict medical health problems in a more accurate and timely manner using real time physiological measurements collected from patients.

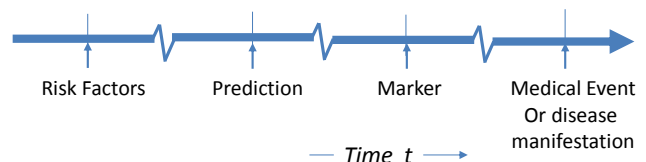


Figure 1. The progression of patient health condition

3. THE PROPOSED MODEL

The model proposed by this research considers medical treatment plan as input to the patient's physiological system.

Represented by $u(t)$, medical treatment plan involves some set of medications, procedures and care protocols prescribed by the physician. The patient's physiology is the process that produces a clinical outcome at time t , represented by $y(t)$. The patient's clinical outcome is the output or the response variable. The outcome is a vector of a single or multiple states of health for that patient. The input and response variable can be shown as:

$$U(t) = (u_1(t), u_2(t), \dots, u_q(t)) \quad (1)$$

$$Y(t) = (y_1(t), y_2(t), \dots, y_m(t)) \quad (2)$$

The internal physiological measurements consisting of clinical and vital sign data such as lab results and monitored data are represented by $x(t)$:

$$X(t) = (x_1(t), x_2(t), \dots, x_k(t)) \quad (3)$$

The model includes a prognostics engine that consists of prediction rules R , and uses specific model M to predict specific outcome for time $(t + t_1)$. The prognostics engine collects vital clinical data from the patient's physiological system and makes a prediction for t_1 minutes in advance, for time $(t + t_1)$. The prognostics engine delivers a prediction that can be used to modify the medical treatment plan $u(t)$. The prediction rules are based on prior evidence and formed from retrospective collection of past patient data. The set of rules can be defined by:

$$R: = (r_1, r_2, \dots, r_p) \quad (4)$$

The model is shown in Figure 2.

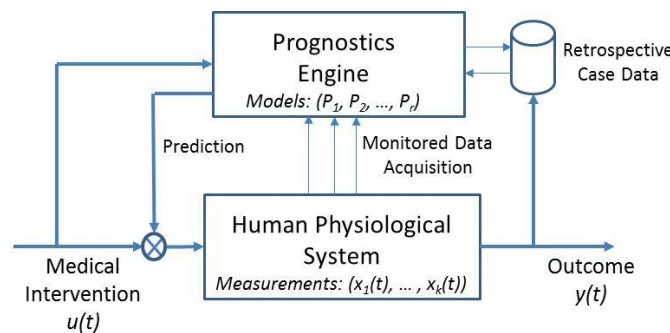


Figure 2. The Medical Prognostic Model

The prognostics engine works continuously by monitoring real time patient data and applying mathematical algorithms that can discern data patterns and make predictions about propensity of certain disease or adverse events occurring in the near future.

The medical intervention, retrospective case information and monitored data can be mathematically described as sets of variables. We can write prediction as a function of multiple variables including the input clinical data and medical intervention. Prediction is a mapping between new input data and an outcome from a set of retrospective cases.

$$Prediction(t+t_1) = F(X, U, R) \quad (5)$$

where physiological data set collected from the patient is represented by vector X ; medical treatment plans are selected from a set of treatment plans shown as U ; and retrospective cases are denoted by R as the set of prior relationships established between physiological data and outcome

The goal of this research is to identify the appropriate mathematical model $F(X, U, R)$ that selects the appropriate prediction from a set of possible outcomes. The model is a mapping function developed based on historical data patterns that maps the input data to a specific outcome.

3.1. Mathematical Model

A large volume of literature concerning mathematical models to predict biological and medical conditions has been published. But only a few of such works in predictive mathematical tools have found their way into mainstream clinical applications and medical practice. Several reasons are cited for the low adoption of predictive tools: either important biological processes were unrecognized or crucial parameters were not known, or that the mathematical intricacies of predictive models were not understood (Swierniak, Kimmel & Smieja, 2009).

The properties of an appropriate mathematical model for predicting medical health condition include: accuracy, prediction, economy, well-posedness and utility (Smye & Clayton, 2002). Among constructs used in prior research, several distinct mathematical models can be found, such as: multivariate regression analysis, Markov chains and stochastic processes, Bayesian networks, fuzzy logic, control theory, discrete event simulation, dynamic programming and Neural Networks.

There are three evolving philosophies pertaining to biological and medical prediction: one is grounded in control theory. Decay of human physiology and adverse medical conditions such as Intra-cranial Pressure (ICP), or carcinogenesis can be viewed as a result of loss of body's control over its critical mechanisms. For example, loss of control over blood flow regulation leads to irregular intracranial pressure; or loss of control over cell cycle causes altered function of a certain cell population that leads to cancer. Medical intervention is viewed as a control action for which the human body is the system. This approach requires a deep understanding of the internal causal models between control mechanisms and human physiology.

The second approach follows the Markov chain model as it considers the disease cycle as a sequence of phases traversed by each physiological subsystem from birth to expiration. For example, a patient that develops pneumonia starts from a healthy normal state and then deteriorates through four stages of Congestion, Red hepatization, Gray hepatization, Resolution (recovery).

The third approach considers the human body as a black box. Since we don't have perfect knowledge about each individual's physiology, environmental, genetic and cultural information and in the areas of medicine where our knowledge of clinical evidence is uncertain, we can only rely on predictive models that take data from physiological sensors and laboratory results and apply certain models and rules developed through retrospective studies to make predictions.

These models have considered both deterministic and probabilistic approaches. Other mathematics constructs consider the asynchronous nature of biology and thus their approach uses simulation models. For example, one study applied simulation and statistical process control to estimate occurrence of hospital-acquired infections and to identify medical interventions to prevent transmission of such infections (Limaye, Mastrangelo, Zerr & Jeffries, 2008).

Other predictive models in cancer therapy have used stochastic process to predict drug resistance of cancer cells and variability in cell lifetimes (Kimmel, Axelrod, 2002). The most successful predictive methods in literature are model-free approaches using neural networks and fuzzy sets (Kodell, Pearce, Baek, Moon & Ahn, 2009) (Arthi & Tamilarasi 2008).

A vast majority of mathematical models in medicine are developed for diagnosis. A survey of literature from 1970's to present, reveals that more attention has been given to decision support and diagnoses models than to prediction. The development of prognostics models to predict short term medical health condition of individuals has been under explored.

Given the non-linear aspect of relationships between physiological measurements, medical outcome and medical treatment plan, a mathematical method that best models non-linear relationships is needed for the prognostics engine. It has been established that neural networks are among the most effective methods to discern patterns and non-linear relationships between data.

3.2. Neural Networks

Neural networks have been successfully applied to classify patterns based on learning from prior examples. Different neural network models use different learning rules, but in general they determine pattern statistics from a set of training examples and then classify new data according to the trained rules. Stated differently, a trained neural network model classifies (or maps) a set of input data to a specific disease from a set of diseases.

To illustrate the classification model for this case study, a simple example is described below and in Figure 3. We can classify input data about patients into two categories of predictions: Disease-True and Disease-False, by looking at prior patient data. The objective of the single-layer

perceptron is to determine a linear boundary that classifies the patients on either side of the linear boundary. As shown in Figure 3, we wish to classify patients into two categories separating by a boundary called a decision boundary line. A linear set of equations define this boundary. The region where the linear equation is >0 is one class (Disease-True), and the region where the linear equation is <0 is the other class (Disease-False). The line is defined as:

$$W_0X_0 + W_1X_1 + W_2 = 0 \quad (6)$$

We can apply a threshold function to classify patients based on the following threshold function:

$$f X_0, X_1 = \begin{cases} 1 & \text{if } W_0X_0 + W_1X_1 + W_2 \geq 0 \\ -1 & \text{if } W_0X_0 + W_1X_1 + W_2 < 0 \end{cases} \quad (7)$$

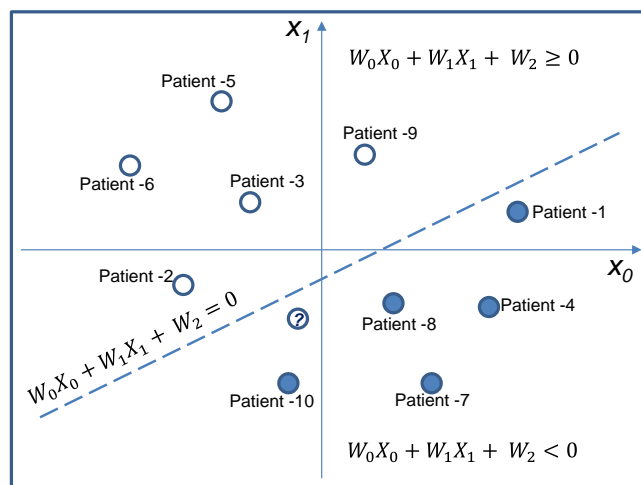


Figure 3. Classification using single-layer perceptron

Suppose we're considering classifying patients by only four input variables, Glucose (G), Body mass (M), Systolic Blood pressure (S) and White blood cell count (B). The threshold function would be computed as follows:

$$f X_0, X_1, X_2, X_3 = \begin{cases} 1 & \text{if } W_4 + \sum_{i=0}^3 W_i X_i \geq 0 \\ -1 & \text{if } W_4 + \sum_{i=0}^3 W_i X_i < 0 \end{cases} \quad (8)$$

The appropriate predictive mathematical model must offer accuracy and simplicity to learn from prior cases and easily be extensible to apply new data to make predictions about a patient's health condition. It has been established that the most accurate neural network models for prediction are as follows:

1) PNN - Probabilistic Neural Networks are four layer networks. They classify data in a non-parametric method and are less sensitive to outlier data. It's been demonstrated that probabilistic neural networks using only four layers of input, pattern, summation and output perceptron can provide accurate and relatively faster classifications than the back-propagation neural networks.

2) SVM – Support Vector Machine networks. SVM performs classification by constructing a two-layer network that defines a hyperplane that separates data into multiple classifications. The SVM is a non-probabilistic binary linear classifier. It takes a set of input data and determines which of possible classes the input is a member of.

3) Generalized Feedforward Multi-Layer Perceptron (MLP) trained with LM – A feedforward neural network consists of one or more layers of nodes where the information flows in only one direction, forward from the input nodes and there are no cycles or loops in the network. In the multi-layer model, each node has direct connection to the nodes in the subsequent layer. The sum of products of the weights and the inputs are calculated in each node (Haykin 1999).

4) MLP trained with LM – Multi-layer perceptron, a method similar to gradient descent approach with variable step modification. Several variations of this model have been proposed, including the Levenberg-Marquardt model (Wilamowski & Chen, 1999) which is known to be among the most efficient algorithms.

This study applied and compared prediction results from all four neural network models. These models were compared based on their accuracy.

4. CLINICAL CASE STUDY

The clinical case study consisted of 468 patient cases who were admitted to a hospital for various treatments. The patient data consisted of 21 independent variables and one dependent variable. The input data included various relevant physical and vital sign data ranging from blood pressure to heart rate and blood lab test results. The input variables consisted of both continuous and dichotomous variables. The dependent variable was a dichotomous variable that represented the clinical outcome, the occurrence or absence of a disease. In this study, the output was defined by a marker called Deep Vein Thrombosis (DVT). Of the patient population in this study, 89 were positively diagnosed with DVT.

DVT is the formation of blood clots in deep veins, typically in leg veins. Blood clots can dislodge and flow to lungs causing a more critical condition called Pulmonary Embolism (PE). DVT/PE is a serious medical condition that can cause serious pain and even death. In the US alone approximately 350,000 to 600,000 patients suffer from DVT and at least 100,000 deaths per year are attributed to DVT/PE (The Surgeon General's Call to Action to Prevent Deep Vein Thrombosis and Pulmonary Embolism, 2008).

Neural networks have been successfully applied to classify patterns based on learning from prior examples. Different neural network models use different learning rules, but in general they determine pattern statistics from a set of training examples and then classify new data according to the trained rules. Stated differently, a trained neural

network model classifies (or maps) a set of input data to a specific disease from a set of diseases.

Four models were trained and tested in two stages: in the first stage, we used genetic neural network algorithm to identify the input variables with most predictive power. We narrowed the list of input variables from 21 down to 14 variables. In the second stage, we trained and tested all four models on the 14 input variables from stage 1. The list of the most predictive variables is given in Table 1.

Input Variable	Data Type	Definition
ADMITTED OVER 48 HRS	Dichotomous	In hospital over 48 hours?
INPATIENT	Dichotomous	Is patient admitted as inpatient?
MAX GLUCOSE	Continuous	Maximum Glucose level during patients' stay.
MAX WEIGHT	Continuous	Maximum weight during stay in Kg.
MIN PLATELET	Continuous	Minimum no. of blood platelets, tiny cells that assist in blood clotting
MIN INR	Continuous	Minimum INR (International Normalized Ratio). The standard for a healthy person is 1.
MAX INR	Continuous	Maximum INR (International Normalized Ratio).
MAX RBC	Continuous	Maximum no. of red blood cells
MIN RBC	Continuous	Minimum no. of red blood cells
MAX HEMOGLOBIN	Continuous	Maximum no. of hemoglobin, a red protein that carries oxygen in the blood.
MIN HEMOGLOBIN	Continuous	Minimum no. of hemoglobin. a red protein that carries oxygen in the blood.
MAX HCT	Continuous	Maximum hematocrit: the proportion, by volume, of red blood cells
MIN HCT	Continuous	Minimum hematocrit: the proportion, by volume, of red blood cells
MIN RDW CV	Continuous	Minimum red blood cell distribution width.
MIN RDW CV3	Continuous	Minimum red blood cell distribution width Coefficient Variation-3.
MIN RDW CV4	Continuous	Minimum red blood cell distribution width Coefficient Variation-4.

Table 1. Input variables description

4.1. Computational Method

In this study, we computed and optimized four different prediction and classification algorithms on 21 data input variables and 468 patient cases. There were 89 true positive cases in the retrospective study. We used NeuroSolutions V6.0 (NeuroDimension, Inc. 2011) tools to build and test the models.

In each computation, we trained the network using one of the four neural network methods. For all four methods, we selected the “Leave-N-out” technique. This technique is a cross training and validation method used to minimize bias due to random data selection. This approach trains the network multiple times, each time omitting a different subset of the data and using that subset for testing. The outputs from each tested subset are combined into one testing report and the model is trained one final time using all of the data.

The test results of all four models can be compared using classification measures such as number of false positives (FP), false negatives (FN), true positives (TP) and true negatives (TN), as shown in Table-2:

Model	TP	FP	TN	FN	Total
Probabilistic Neural Network	5	7	372	84	468
Support Vector Machines	30	83	296	59	468
Multi-layer Perceptron with LM	24	78	301	65	468
Generalized Feed forward with LM	19	68	311	70	468

Table 2. Model test results

4.2. Accuracy and Validation

External validity of medical prediction models is an extremely challenging task. Clinical validation is challenging not just because it involves prospective patient studies, double-blind studies and careful administration of research protocols, but for two other reasons: first, if a patient gets the treatment, could that patient have exhibited the predicted disease? Second, if a patient is not treated and the disease occurs, would the outcome been the validation of the model’s prognosis had the patient been treated? We’ll focus on accuracy in this paper and consign clinical validation to a future research project.

Several measurements have been proposed as methods for internal validation. Some of the measurements that are commonly used to compare accuracy of classification models include: Accuracy, Sensitivity, Specificity, Area Under Receiver Operating Curve (AUROC) and Likelihood Ratio (LR). Sensitivity measures the fraction of positive cases that are classified correctly as positive. Specificity is the fraction of negative cases that are classified correctly as negative. AUROC is a good overall measure of predictive

accuracy of a model. It represents a plot of sensitivity versus (1 - specificity). An AUROC close to 1.0 is a considered an excellent discrimination, but a value near 0.50 suggests no discrimination (similar to a coin flip). Positive LR is the ratio of sensitivity to (1 - specificity). (Delen 2009). The accuracy measures may be defined as:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

$$sensitivity = \frac{TP}{TP+FN} \quad (10)$$

$$specificity = \frac{TN}{TN+FP} \quad (11)$$

$$LR+ = \frac{sensitivity}{1-specificity} = \frac{Pr T+ D+}{Pr T+ D-} \quad (12)$$

where (T+ / D+) denotes a case with positive test of a disease when the disease actually exists, and (T+ | D-) denotes a case with positive test of a disease but the patient does not present with the disease.

When a model uses continuous data measurements, then different thresholds may be applied in order to decide which value is the cut-off to distinguish between patients with disease. The best model has the highest values for sensitivity and specificity. In certain situations, both may not be equally important. For example, a false-negative (FN) prediction might be more critical than a false-positive (FP) prediction. If we apply no preference to either measurement then, Youden’s index (J) may be used to choose an appropriate cut-off, computed by (Bewick, Cheek, Ball 2004):

$$J = sensitivity + specificity - 1 \quad (13)$$

The maximum value that J can take is 1, when the test is perfect.

4.3. Comparison of results

All four models were optimized for classification of cases into a dichotomous dependent variable: the presence or absence of DVT.

The results showed that the SVM algorithm was most accurate followed by the MLP model and the General feed forward neural network model. All four methods are compared using the accuracy measurements in Table 3.

Measurement	Probabilistic Neural Network	Support Vector Machine	Multi-Layer Perceptron - LM	Generalized Feed forward - LM
Accuracy	0.8056	0.6966	0.6944	.7051
Sensitivity	0.0562	0.3371	0.2697	0.2135
Specificity	0.9815	0.7810	0.7942	0.8206
LR+	3.0417	1.5392	1.3103	1.1899
Youden’s J	0.0377	0.1181	0.0639	0.0341

Table 3. Accuracy Measures of Neural network models

All four models exhibited low sensitivity measures indicating their poor ability to detect true positives. This is due to the lower number of positive DVT cases in this study (only 89 out of 468 cases had positive DVT cases).

4.4. Use of an Oracle to Select the Best Model

Since their introduction in 1960’s, various neural network algorithms have been proposed and successfully implemented to classify and predict future state of output variables. Certain models are more suitable to specific class of problems based on the type and number inputs and output classifications. Typically, no single neural network model is best for all types of problem.

Given that there are many neural networks to select from, the goal is to select the most accurate model for prediction of each disease condition. Therefore, we propose that the prognostics engine utilizes several different algorithms to determine accuracy of each method and then use an oracle, an overseer program that selects the most accurate model. An oracle is defined as the medium which selects the best answer amongst a set of options. An oracle can be defined to select the best algorithm or a combined prediction from an ensemble of algorithms based on desired accuracy or characteristic of the prediction problem at hand.

Given that one model performs better in predicting true positives and another better at predicting the true negatives, we propose the oracle program to combine the predictions from models in a way that the model with higher accuracy is assigned a higher weight and the worst model still contributes to the prediction but at a smaller weight. This way, the oracle can improve the classification accuracy, sensitivity and specificity by combining the best classification characteristics from different models.

An approach that uses an ensemble of prognostic algorithms is shown to be effective in providing more accurate prediction (Hu, Youn & Wang 2010).

We produced two Oracle methods to compute the combined predictions of the ensemble. The first Oracle used conditional logic to maximize the number of TP and minimize the number of FP predictions.

The results of the oracles are shown in Table 4 and accuracy comparison in Table 5.

Model	TP	FP	TN	FN	Total
Oracle #1 – Ensemble of PNN & SVM models	35	107	272	54	468
Oracle#2 – Ensemble of all four models	46	141	238	43	468

Table 4. Results of the two oracle program

Ensemble1 essentially took the best traits from the PNN and SVM models to produce a more accurate prediction. The second Oracle combined weighted sum of predictions from

each model in the ensemble. The weights were determined to maximize the number of FP predictions.

Measurement	Oracle #1	Oracle #2
Accuracy	0.6560	0.6068
Sensitivity	0.3933	0.5169
Specificity	0.7177	0.6280
LR+	1.3929	1.3893
Youden’s J	0.1109	0.1448

Table 5. Comparison of Oracles’ accuracy

One method to compare all four models and the two oracle programs is to use the Receiver Operating Curve (ROC) plot. The ROC curve is a plot of sensitivity versus (1 – specificity), and generally is considered a good accuracy measure of binary classifiers (Bourdes, Ferrieres, Amar, Amelineau, et al, 2011). Figure 4 shows a scatter plot of ROC for all models. The best prediction method would result in a point in the upper left corner of the diagram. The diagonal line depicts a random guess or prediction by a flip of coin.

The diagram illustrates two observations: The prediction results are not as accurate as one would like. This is attributed to the fact there were too few positive cases in the entire population to help train a more accurate predictive model. Furthermore, several of input variables were highly correlated such that the predictive contribution of some variables was less significant for making a more accurate prediction.

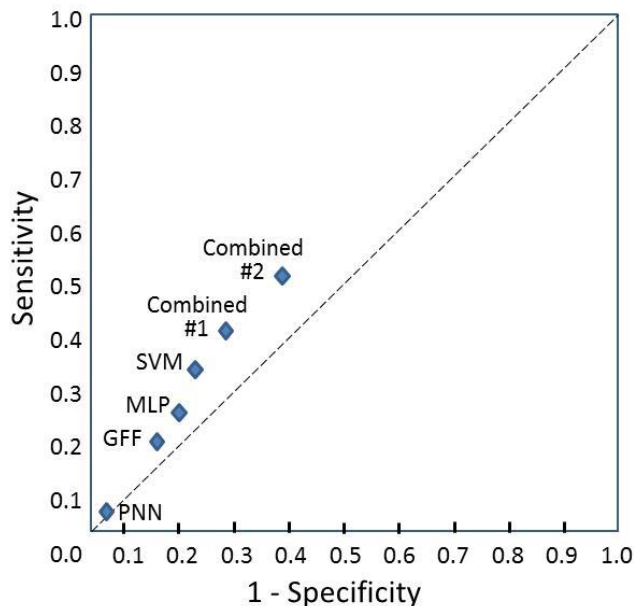


Figure 4. ROC curve for results of all models

The second observation is that the combined ensemble methods #1 and #2 were more accurate than each neural network model alone.

5. CONCLUSION

Various neural network methods can be used to identify the most accurate model to make short term predictions about patient health condition. The performance of these models varies depending on the type and volume of input and output variables. The conclusion of our study is not to say which model or method is better, but, to recognize that each model has strengths and weaknesses. By combining multiple models we can improve classification accuracy. Since no single model can be the best fit for all medical prediction problems, an oracle program is proposed to select the best weighted combination of multiple neural network models.

ACKNOWLEDGEMENT

Authors would like to thank and acknowledge insights and feedback provided by Doctors John Bramhall, M.D, PhD. and Michael Souter, M.B, Ch.B. to this research.

REFERENCES

- Arthi, K., Tamilarasi, A. (2008). Prediction of autistic disorder using neuro fuzzy systems by applying ANN technique, *International Journal of Developmental Neuroscience*, 26 (2008) 699-704.
- Bewick, V., Cheek, L., Ball, J., Statistics review 13: Receiver operating characteristic curves, *Critical Care*, Vol. 8, no. 6, December 2004
- Bourdes, V., Ferrieres, J., Amar, J., Amelineau, E., Bonnevey, S., Berlion, M., Danchin, N., Prediction of persistence of combined evidence-based cardiovascular medications in patients with acute coronary syndrome after hospital discharge using neural networks, *Medical Biological Engineering Computing*, 49:947-955, 2011
- Brown, S.W., Strong, V. (2001), The use of seizure-alert dogs, *Seizure*, 2001, 10:39-41.
- Daley, M., Narayanan, N., Leffler, C. W. (2010), Model-derived assessment of cerebrovascular resistance and cerebral blood flow following traumatic brain injury, *Experimental Biology and Medicine*, Vol 235, April 2010
- (Delen 2009) D. Delen, "Analysis of cancer data: a data mining approach", *Expert Systems*, February 2009, Vol. 26, No. 1
- Gao, E., Young, W., Ornstein, E., Pile-Spellman, J., Qiyuan, M. (1997), A theoretical model of cerebral hemodynamics: Application to the study of Arteriovenous Malformations, *Journal of Cerebral Blood Flow and Metabolism*, 1997, 17, 905-918
- Hahnfeldt, P., Panigraphy, D, Folkman, J., Hlatkey, L., Tumor development under angiogenic signaling: a dynamic theory of tumor growth, treatment response and postvascular dormancy, *Cancer Research* 59, 4770-4778
- Haykin, S. (1998), *Neural Networks, A Comprehensive Foundation*, 2nd Edition, Prentice Hall, 1999
- Hu, C., Youn, B.D., Wang, P., Ensemble of data-driven prognostics algorithms with Weight Optimizatio and K-Fold Cross Validation, Annual Conference of the Prognostics and Health Management (PHM) Society, Oct 10-16 2010, Portland, OR.
- INCOSE (International Council on Systems Engineering Council) *Systems Engineering Handbook* (2000), *What is a system?*, Version 2.0, July 2000.
- Kimmel, M., Axelrod, D. E. (2002), *Branching processes in biology*, Springer Verlag, New York, NY, 2002
- Kirton, A., Winter, A., Wirrell, E., Snead, O.C. (2008), Seizure response dogs: Evaluation of a formal training program, *Epilepsy & Behavior*, 13 (2008) 499-504.
- Kodell, R. L. , Pearce, B. A. , Baek, S., Moon, H., Ahn, H., et al. (2009), A model-free ensemble method for class prediction with application to biomedical decision making, *Artificial Intelligence in Medicine*, 46, 267-276
- Limaye, S. S., Mastrangelo, C. M., Zerr, D. M., Jeffries, H. (2008), A statistical approach to reduce hospital-associated infections, *Quality Engineering*, 20:414-425, 2008
- NeuroDimension, Inc. (2011), Gainesville, Florida, NeuroSolutions software, Version 6.0
- Niu, G., Yang, B., Pecht, M., Development of an optimized condition-based maintenance system by data fusion and reliability-centered maintenance, *Reliability Engineering and System Safety*, V95, n7, p786-796, 2010
- O'Connor, A. M., Bennett, C. L., Stacey, D., Barry, M., Col, N. F., Eden, K. B. , Entwistle, V. A., Fiset, V., et. al. (2009), Decision aids for people facing health treatment or screening decisions (Review), *The Cochrane Collaboration*, Wiley 2009
- Pecht, M. (2008), *Prognostics and Health Management of Electronics*, Wiley 2008
- Smye, S. W., Clayton, R. H. (2002), Mathematical modeling for the new millennium: medicine by numbers, *Medical Engineering & Physics*, 24 (2002), 565-574
- Souter, M. (2011), Conversations about diagnostic markers and predictors, April 7, 2011.
- Strong, V., Brown, S. W., Walker, R. (1999), Seizure-alert dogs-fact or fiction?, *Seizure*. 1999; 8:26-65.
- Swierniak, A., Kimmel, M., Smieja, J. (2009), Mathematical modeling as a tool for planning anticancer therapy, *European Journal of Pharmacology*, 625 (2009) 108-121
- Uckun, S., Goebel, K., Lucas, P. J. F. (2008), Standardizing Research Methods for Prognostics, *2008 International Conference on Prognostics and Health Management*
- Vichare, N. M., Pecht, M. (2006), Prognostics and Health Management of Electronics, *IEEE Transactions on Components and Packaging Technologies*, Vol 29, No. 1, March 2006.
- Virchow, R. (1856), Virchow's triad was first formulated by the German physician Rudolf Virchow in 1856.

Williams, H., Pembroke, A. (1989), Sniffer dogs in the melanoma clinic?, *Lancet*, 1989, 1(8640);734

Wilamowski, B. M., Chen, Y. (1999), Efficient algorithm for Training Neural Networks with one Hidden Layer, IEEE International Joint Conference on Neural Networks, 1999

Yu, C., Liu, Z., McKenna, T., Reisner, A. T., Reifman, J. (2006), A Method for Automatic Identification of Reliable Heart Rates Calculated from ECG and PPG Waveforms, *Journal of the American Medical Informatics Association*, vol. 13, No. 3, May/June 2006.

Peter K. Ghavami Received his M.S. in Engineering Management from Portland State University and B.S. from Oregon State University. He is currently Director of Imaging Informatics at Harborview Medical Center and a graduate student at University of Washington, Dept. of

Industrial and Systems Engineering. He has authored papers on software process improvement, vector processing, distributed network architectures, and software quality. He authored the book *Lean, Agile and Six Sigma Information Technology Management* in 2008. He is a member of IEEE Reliability Society and IEEE Life Sciences Initiative.

Kailash C. Kapur is a Professor of Industrial & Systems Engineering in the College of Engineering at the University of Washington. He was the Director of Industrial Engineering at the University of Washington from January 1993 to September 1999. He has co-authored the book *Reliability in Engineering Design*, John Wiley & Sons, 1977. He is a *Fellow* of American Society for Quality and a fellow of the Institute of Industrial Engineers.

Exploring the Model Design Space for Battery Health Management

Bhaskar Saha¹, Patrick Quach², and Kai Goebel³

¹*Mission Critical Technologies, Inc. (NASA ARC), El Segundo, CA 90245, USA*
bhaskar.saha@nasa.gov

²*NASA Langley Research Center, Hampton, VA 23681, USA*
cuong.c.quach@nasa.gov

³*NASA Ames Research Center, Moffett Field, CA 94035, USA*
kai.goebel@nasa.gov

ABSTRACT

Battery Health Management (BHM) is a core enabling technology for the success and widespread adoption of the emerging electric vehicles of today. Although battery chemistries have been studied in detail in literature, an accurate run-time battery life prediction algorithm has eluded us. Current reliability-based techniques are insufficient to manage the use of such batteries when they are an active power source with frequently varying loads in uncertain environments. The amount of usable charge of a battery for a given discharge profile is not only dependent on the starting state-of-charge (SOC), but also other factors like battery health and the discharge or load profile imposed. This paper presents a Particle Filter (PF) based BHM framework with plug-and-play modules for battery models and uncertainty management. The batteries are modeled at three different levels of granularity with associated uncertainty distributions, encoding the basic electrochemical processes of a Lithium-polymer battery. The effects of different choices in the model design space are explored in the context of prediction performance in an electric unmanned aerial vehicle (UAV) application with emulated flight profiles.

1. INTRODUCTION

Battery-powered devices have become ubiquitous in the modern world, from tiny headsets to cameras, cell phones and laptops to hybrid and electric vehicles. Yet the battery is not a new invention. Battery artifacts date back to the early centuries A.D. (the Baghdad battery) and electric cars were favored over their gasoline counterparts in the late nineteenth century because of higher reliability. However,

the uncertainty in determining battery life plagued electric vehicles then as it does now. A recent report by the Consumer Electronics Association, “Electric Vehicles: The Future of Driving”, indicates that although these vehicles are increasing in popularity, running out of battery power on the road is the top concern for consumers (71% of adults surveyed). Consequences of battery exhaustion may range from reduced performance to operational impairment and even to catastrophic failures, thus motivating the study of Battery Health Management (BHM).

One of the most critical applications of BHM technologies is in the field of electric vehicles (EVs). Usually combustion based powertrains run within narrow bands of RPMs (revolutions per minute) with metered fuel delivery. This combined with a known volume fuel tank allows reasonably accurate predictions of remaining use-time or travel distance. Batteries on the other hand, decrease in capacity with time and usage. Various factors like ambient storage temperatures and the state-of-charge (SOC) at which the battery was stored affects capacity fade. Additionally, the amount of usable charge of a battery for a given discharge profile is not only dependent on the starting SOC, but also other factors like battery health and the discharge or load profile imposed.

In this paper, the BHM problem is approached from the model-based point of view. The following sections will address the salient battery characteristics that need to be modeled, the BHM framework, explorations of the model design space, an electric unmanned aerial vehicle (UAV) application example, battery end-of-discharge (EOD) prediction results, and relevant conclusions.

2. BATTERY CHARACTERISTICS

Batteries are essentially energy storage devices that facilitate the conversion, or *transduction*, of chemical energy into electrical energy, and vice versa (Huggins,

Bhaskar Saha et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

2008). They consist of a pair of *electrodes* (*anode* and *cathode*) immersed in an *electrolyte* and sometimes separated by a *separator*. The chemical driving force across the cell is due to the difference in the chemical potentials of its two electrodes, which is determined by the difference between the *standard Gibbs free energies* the products of the reaction and the reactants. The theoretical *open circuit voltage*, E^o , of a battery is measured when all reactants are at 25°C and at 1M concentration or 1 atm pressure. However, this voltage is not available during use. This is due to the various passive components inside like the electrolyte, the separator, terminal leads, etc. The voltage drop due to these factors can be mainly categorized as follows.

Ohmic Drop

This refers to the diffusion process through which Li-ions migrate to the cathode via the electrolytic medium. The internal resistance to this ionic diffusion process is also referred to elsewhere as the IR drop. For a given load current this drop usually decreases with time due to the increase in internal temperature that results in increased ion mobility, and is henceforth referred to as ΔE_{IR} .

Activation Polarization

Self-discharge is caused by the residual ionic and electronic flow through a cell even when there is no external current being drawn. The resulting drop in voltage has been modeled to represent the activation polarization of the battery, referred to from now on as ΔE_{AP} . All chemical reactions have a certain activation barrier that must be overcome in order to proceed and the energy needed to overcome this barrier leads to the activation polarization voltage drop. The dynamics of this process is described by the Butler–Volmer equation. This process was represented by an exponential function in Saha and Goebel (2009). However, a log function is a more accurate representation, as abstracted from the Butler–Volmer equation.

Concentration Polarization

This process represents the voltage loss due to spatial variations in reactant concentration at the electrodes. This is mainly caused when the reactants are consumed by the electrochemical reaction faster than they can diffuse into the porous electrode, as well as due to variations in bulk flow composition. The consumption of Li-ions causes a drop in their concentration along the cell, between the electrodes, which causes a drop in the local potential near the cathode. This voltage loss is also referred to as concentration polarization, represented in this paper by the term ΔE_{CP} . The value of this factor is low during the initial part of the discharge cycle and grows rapidly towards the end of the discharge or when the load current increases.

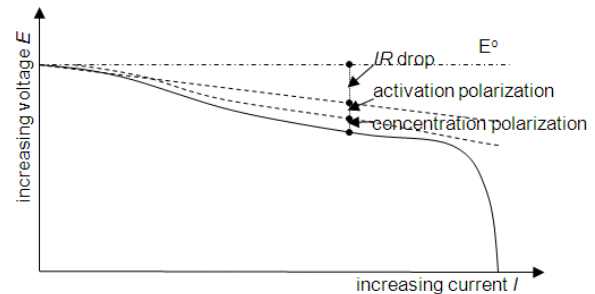


Figure 1. Typical polarization curve of a battery

Figure 1 depicts the typical polarization curve of a battery with the contributions of all three of the above factors shown as a function of the current drawn from the cell. Since, these factors are current-dependent, i.e., they come into play only when some current is drawn from the battery, the voltage drop caused by them usually increases with increasing output current.

Since the output current plays such a big role in determining the losses inside a battery, it is an important parameter to consider when comparing battery performance. The term most often used to indicate the rate at which a battery is discharged is the *C-Rate* (Huggins, 2008). The discharge rate of a battery is expressed as C/r , where r is the number of hours required to completely discharge its nominal capacity. So, a 2 Ah battery discharging at a rate of $C/10$ or 0.2 A would last for 10 hours. The terminal voltage of a battery, as well as the charge delivered, can vary appreciably with changes in the C-Rate. Furthermore, the amount of energy supplied, related to the area under the discharge curve, is also strongly C-Rate dependent. Figure 2 shows the typical discharge of a battery and its variation with C-Rate. Each curve corresponds to a different C-Rate or C/r value (the lower the r the higher the current) and assumes constant temperature conditions.

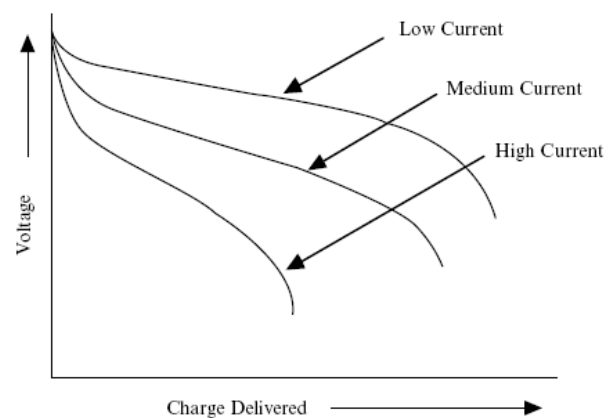


Figure 2. Schematic drawing showing the influence of the current density upon the discharge curve (Reproduced from Figure 1.14 in (Huggins, 2008))

3. HEALTH MANAGEMENT FRAMEWORK

Before investigating the issues with modeling the battery, this section takes a look at how the BHM framework is implemented using Particle Filters. The framework has been described before (Saha *et al.*, 2009), however, some basic elements are reproduced below in order to set the context.

3.1 Particle Filter

The Particle Filter (PF) framework (Gordon *et al.*, 1993) assumes that the state equations can be modeled as a first order Markov process with additive noise and conditionally independent outputs. Under these assumptions the state equations can be represented as:

$$\mathbf{x}_k = \mathbf{f}_{k-1}(\mathbf{x}_{k-1}) + \omega_{k-1} \quad (1)$$

$$\mathbf{z}_k = \mathbf{h}_k(\mathbf{x}_k) + \mathbf{v}_k \quad (2)$$

The filter approximates the posterior probability distribution denoted as $p(\mathbf{x}_k | \mathbf{Z}_k)$, where $\mathbf{Z}_k = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_k]$ is the set of all measurements until t_k , by a set of N weighted particles $\{\langle x_p^i, w_p^i \rangle; i=1, \dots, N\}$, such that $\sum_i w_k^i = 1$, and the posterior distribution can be approximated as:

$$p(\mathbf{x}_k | \mathbf{Z}_k) \approx \sum_{i=1}^N w_k^i \delta(\mathbf{x}_k - \mathbf{x}_k^i). \quad (3)$$

Using the model in Eq. (1) the prior distribution going from t_{k-1} to t_k becomes:

$$p(\mathbf{x}_k | \mathbf{Z}_{k-1}) \approx \sum_{i=1}^N w_{k-1}^i \mathbf{f}_{k-1}(\mathbf{x}_{k-1}^i). \quad (4)$$

The weights are updated according to the relation:

$$\bar{w}_k^i = w_{k-1}^i \frac{p(\mathbf{z}_k | \mathbf{x}_k^i) p(\mathbf{x}_k^i | \mathbf{x}_{k-1}^i)}{q(\mathbf{x}_k^i | \mathbf{x}_{k-1}^i, \mathbf{z}_k)}, \quad (5)$$

$$w_k^i = \frac{\bar{w}_k^i}{\sum_{j=1}^N \bar{w}_k^j}. \quad (6)$$

Resampling is used to avoid the problem of degeneracy of the PF algorithm, i.e., avoiding the situation that all but a few of the importance weights are close to zero. If the weights degenerate, we not only have a very poor representation of the system state, but we also spend valuable computing resources on unimportant calculations. More details on this are provided in Saha *et al.* (2009). The basic logical flowchart is shown in Figure 3.

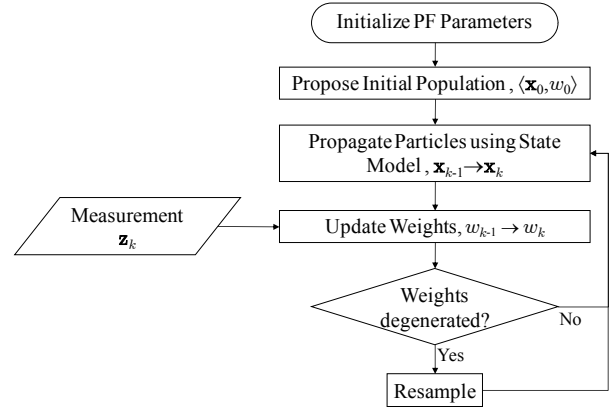


Figure 3. Particle filtering flowchart

During prognosis this tracking routine is run until a long-term prediction is required, say at time t_p , at which point Eq. (4) will be used to propagate the posterior pdf (probability density function) given by $\{\langle x_p^i, w_p^i \rangle; i=1, \dots, N\}$ until \mathbf{x}^i fails to meet the system specifications at time t_{EOL}^i . The remaining useful life (RUL) pdf, i.e., the distribution of $p(t_{EOL}^i - t_p)$, is given by the distribution of w_p^i . Figure 4 shows the flow diagram of the prediction process.

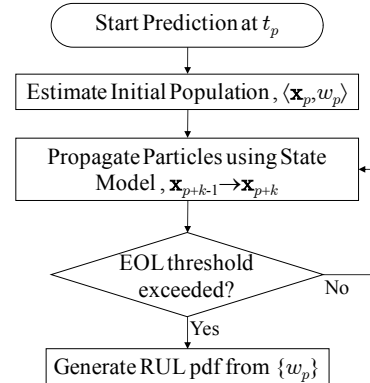


Figure 4. Prediction flowchart

3.2 Model Adaptation

One of the key motivating factors for using Particle Filters for prognostics is the ability to include model parameters as part of the state vector to be estimated. This performs model adaptation in conjunction with state tracking, and thus, produces a tuned model that can be used for long term predictions.

Assume that the system health state is 1-dimensional, given by x_k , and the state evolution model \mathbf{f} and the measurement model \mathbf{h} are stationary in nature with known noise distributions ω and \mathbf{v} respectively. Additionally, we also assume that the parameter values of \mathbf{h} are known. This assumption can be relaxed in a more generic approach.

Indeed, considering a non-stationary measurement model can be used to account for progressive degradation in sensors caused by corrosion, fatigue, wear, etc. The parameters of \mathbf{f} , denoted by $\alpha_k = \{\alpha_{j,k}; j = 1, \dots, n_f\}$, $n_f \in \mathbb{N}$, are combined with x_k to give the state vector $\mathbf{x}_k = [x_k \ \alpha_k]^T$, where T represents the transpose of a vector or matrix. Equations (1) and (2) can then be rewritten as:

$$x_k = \mathbf{f}(x_{k-1}, \alpha_{k-1}) + \omega_{k-1} \quad (7)$$

$$z_k = \mathbf{h}(x_k) + v_k. \quad (8)$$

The issue now is to formulate the state equations for α_k . One easy solution is to pick a *Gaussian random walk* such that:

$$\alpha_{j,k} = \alpha_{j,k-1} + \omega_{j,k-1} \quad (9)$$

where $\omega_{j,k-1}$ is drawn from a normal distribution, $\mathcal{N}(0, \sigma_j^2)$, with zero mean and variance σ_j^2 . Given a suitable starting point $\alpha_{j,0}$, and variance σ_j^2 , the PF estimate will converge to the actual parameter value α_j , according to the *law of large numbers*.

It is not necessary to include all model parameters as part of the state to be estimated. In fact, the smaller the subset of parameters to be estimated, the faster the convergence since the state dimensionality is lower (Daum, 2003). However, this leads to the notion that the higher the model fidelity with respect to the real system, the lesser the number of parameters that need to be identified at run-time leading to better convergence properties.

4. MODEL DESIGN SPACE

The issue of modeling is paramount in any model-based algorithm like the PF. There can be many approaches to modeling, and for well studied systems like batteries the model design space is very large. There are several models that exist in literature at various levels of granularity and abstraction, like Gao, Liu, and Dougal (2002), Hartmann II (2008), Santhanagopalan, Zhang, Kumaresan, and White (2008), etc. Building these models require significant expenses in time and expertise. However, there are still issues with applicability in the field, since complex models need identification of several parameters, which might be impractical. Sometimes the models may be too complex to be run in real time.

For the purposes of the electric UAV BHM, we explore the model design space at a high level of abstraction of the underlying physics. It is desired to model the SOC of the battery in order to predict the EOD event as discussed below. In the results section the prediction performance for the different model choices are presented.

4.1 Model 1

For the empirical charge depletion model considered here, we express the output voltage $E(t_k)$ of the cell in terms of the effects of the changes in the internal parameters, as shown below:

$$E(t_k) = E^\circ - \Delta E_{IR}(t_k) - \Delta E_{AP}(t_k) - \Delta E_{CP}(t_k) \quad (10)$$

where E° is the Gibb's free energy of the cell, ΔE_{IR} is the Ohmic drop, ΔE_{AP} is the drop due to activation polarization and ΔE_{CP} denotes the voltage drop due to concentration polarization. These individual effects are modeled as:

$$\Delta E_{IR}(t_k) = \Delta I_k R - \alpha_{1,k} t_k, \quad (11)$$

$$\Delta E_{AP}(t_k) = \alpha_{2,k} \exp(-\alpha_{3,k} / t_k), \quad (12)$$

$$\Delta E_{CP}(t_k) = \alpha_{4,k} \exp(\alpha_{5,k} t_k). \quad (13)$$

where ΔI_k is the change in current that flows through the internal resistance R of the cell, and $\alpha_k = \{\alpha_{j,k}; j = 1, \dots, 5\}$ represents the set of model parameters to be estimated.

The problem is to predict the time instant t_{EOD} when the state x denoting the cell voltage E reaches the threshold level of 2.7 V. The PF representation of this problem is given by:

$$\begin{aligned} x_k &= x_{k-1} - \\ &\left\{ -\alpha_{1,k-1} + \alpha_{2,k-1} \alpha_{3,k-1} \exp(-\alpha_{3,k-1} / t_{k-1}) t_{k-1} \right. \\ &\left. - \alpha_{4,k-1} \alpha_{5,k-1} \exp(\alpha_{5,k-1} t_{k-1}) \right\} (t_k - t_{k-1}) \\ &- \Delta I_k R + \omega_{k-1}, \end{aligned} \quad (14)$$

$$\alpha_{1,k} = \alpha_{1,k-1} + \omega_{1,k-1}$$

$$\alpha_{2,k} = \alpha_{2,k-1} + \omega_{2,k-1}$$

$$\alpha_{3,k} = \alpha_{3,k-1} + \omega_{3,k-1}$$

$$\alpha_{4,k} = \alpha_{4,k-1} + \omega_{4,k-1}$$

$$\alpha_{5,k} = \alpha_{5,k-1} + \omega_{5,k-1}$$

$$z_k = x_k + v_k. \quad (15)$$

This is a 6 dimensional state vector with 1 dimension being the system health indicator (cell voltage) and the other dimensions coming from the model parameters. The term ΔI_k is the change in the load current at the time instant t_k .

4.2 Model 2

The model represented by Eqs. (14) – (15) does not represent the activation polarization process well. This is because the structure of the Butler Volmer equation is better

approximated by a log function rather than a negative exponential. Hence for Model 2, we change Eq. (12) to the following:

$$\Delta E_{AP}(t_k) = \alpha_{2,k} \ln(1 + \alpha_{3,k} t_k). \quad (16)$$

Correspondingly, Eq. (14) changes to:

$$\begin{aligned} x_k &= x_{k-1} - \\ &\left\{ -\alpha_{1,k-1} + \alpha_{2,k-1} \alpha_{3,k-1} / (1 + \alpha_{3,k-1} t_{k-1}) \right. \\ &\left. - \alpha_{4,k-1} \alpha_{5,k-1} \exp(\alpha_{5,k-1} t_{k-1}) \right\} (t_k - t_{k-1}) \\ &- \Delta I_k R + \omega_{k-1}, \\ \alpha_{1,k} &= \alpha_{1,k-1} + \omega_{1,k-1} \\ \alpha_{2,k} &= \alpha_{2,k-1} + \omega_{2,k-1} \\ \alpha_{3,k} &= \alpha_{3,k-1} + \omega_{3,k-1} \\ \alpha_{4,k} &= \alpha_{4,k-1} + \omega_{4,k-1} \\ \alpha_{5,k} &= \alpha_{5,k-1} + \omega_{5,k-1} \end{aligned} \quad (17)$$

The state vector is similar here as in Model 1. The level of granularity, indicating the different physical processes modeled, is the same although the abstraction of one of the processes has changed.

4.3 Model 3

It should be noted that for most batteries, the voltage as well as the charge delivered varies considerably with changes in I . This can be better represented by making two changes to the battery model described so far. Firstly, the parameters of the model must be load dependent. We model this by making α_3 and α_5 proportional to the load current I . Secondly, when we have step changes in the load, a higher load level followed by a lower one presents a period of relaxation for the battery. During this period the voltage does not immediately jump up but gradually rises which can be modeled by an exponential function. A similar effect can also be observed for a step increase in current level. These effects can be reconciled by considering the battery impedance as an RC equivalent circuit (Zhang, 2010). We can thus replace Eq. (11) by:

$$\Delta E_{IRC}(t_k) = \Delta I_k \alpha_6 (1 - \exp(-\alpha_7 (t_k - t_{\Delta I_k}))) - \alpha_1 t_k \quad (18)$$

where ΔI_k is the step change in current at time $t_{\Delta I_k}$. The other processes are represented as:

$$\Delta E_{AP}(t_k) = \alpha_{2,k} \ln(1 + \alpha_{3,k} I_k t_k), \quad (21)$$

$$\Delta E_{CP}(t_k) = \alpha_{4,k} \exp(\alpha_{5,k} I_k t_k). \quad (22)$$

The filter equations can be derived out as before and are shown in Saha *et al.* (2011). Model 3 represents a higher level of granularity in the model design space since some additional battery behavior to changes in load is being taken into effect. This leads to higher accuracy in the model output as well as a corresponding increase in the number of parameters. To maintain a tolerable rate of convergence, all but the parameters α_3 and α_5 are learnt from training data, while α_3 and α_5 are estimated by the PF online.

5. APPLICATION EXAMPLE

The test UAV platform for this research is a COTS 33% scale model of the Zivko Edge 540T. Details of this platform have been presented in Saha *et al.* (2011), but are also repeated here for the sake of readability. The UAV is powered by dual tandem mounted electric out-runner motors capable of moving the aircraft up to 85 knots using a 26 inch propeller. The gas engine in the original kit specification was replaced by two electric out runner motors which are mounted in tandem to power a single drive shaft. The motors are powered by a set of 4 Li-Poly rechargeable batteries. The batteries are each rated at 6000 mAh. The tandem motors are each controlled by separate motor controllers.

Testing on the Edge 540 UAV platform was carried out with the airframe restrained on the ground. The propeller was run through various RPM (revolutions per minute) regimes indicative of the intended flight profile (takeoff, climb, multiple cruise, turn and glide segments, descent and landing). Figure 5 shows the voltages during a typical flight. It is desired to predict when the battery will run out of charge, i.e., when the EOD event indicated by the end of the voltage plots after landing will occur.

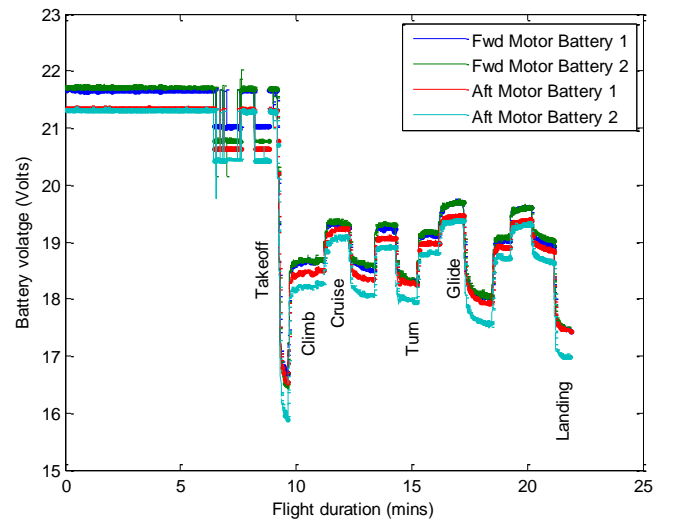


Figure 5. Battery voltages during a typical flight

6. RESULTS

In order to evaluate the prognostic algorithm we make 7 predictions spaced 1 minute apart starting from 800 secs into the flight. It is not desired to make predictions till the end of the flight since there needs to be some time for the UAV pilot to land the aircraft with some safety margin on the remaining battery life. Figures 6 – 8 show sample predictions generated by the Models 1 – 3 respectively. The time instants when the predictions are made are shown in green vertical dashed lines, with lighter shades indicating earlier predictions. The corresponding EOD pdfs are shown in green patches on the 17.4 V EOD threshold voltage line (dashed gray). The pdfs themselves are given by the distribution $\{ \langle t_{EOD}^i - t_p, w_p^i \rangle; i = 1, \dots, N \}$, where i is the particle index and t_{EOD}^i is the predicted time where the i th particle trajectory crosses the EOD threshold. The real voltages are shown as red x's, while the PF estimates are shown as blue dots. The large spread of the blue dots is caused by the bias errors and noise in the Hall effect current sensors used. Since this uncertainty has not been expressly modeled, the actual EOD can sometimes lie outside the predicted pdf as shown in Figures 6 – 8.

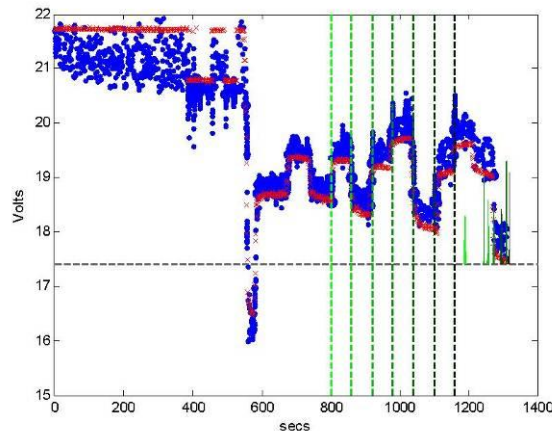


Figure 8. Sample prediction using Model 3

For statistical validation, we ran each model 100 times over the same data to generate the α - λ metric plots as defined in Saxena *et al.* (2008). This prognostic metric measures whether RUL predictions continue improve in accuracy with time as more run-time data is made available, where $t_{RUL}^i = t_{EOD}^i - t_p$. It also enforces the notion that the prediction error needs to reduce as the prediction time instant approaches the end of life (EOD in this case) since there is less time to take corrective action. For these experiments, the α value is chosen to be 0.1 and λ is chosen to be 0.5 (i.e. it is desired that the prediction trajectories be within 90% accuracy with 50% battery life left). Figures 9 – 11 show the α - λ plots for Models 1 – 3 respectively for $t_p = [800, 860, 920, 980]$, Model 1 shows the worst performance, while Model 3 is the best as was expected from the model choices. The worsening performance of both Models 1 and 2 toward the end predictions is most likely due to the inability of these models to adapt to the low load glide modes as shown in Figure 5.

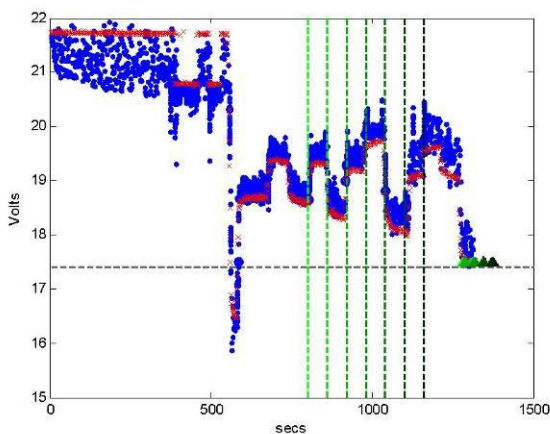


Figure 6. Sample prediction using Model 1

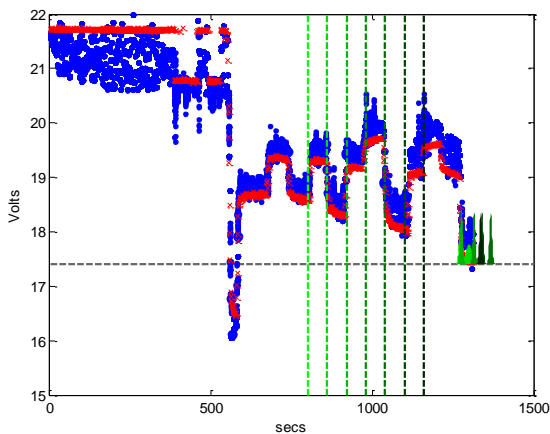


Figure 7. Sample prediction using Model 2

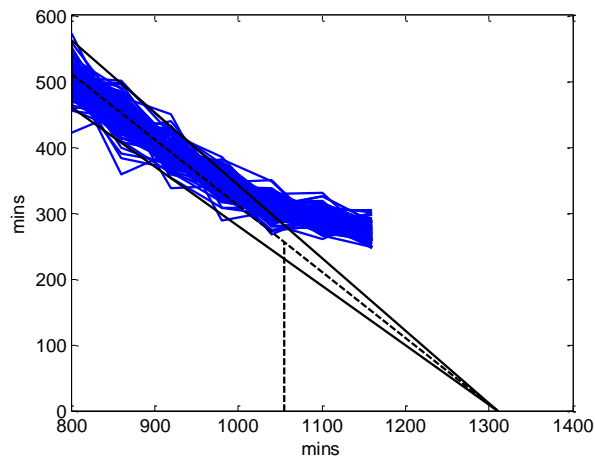
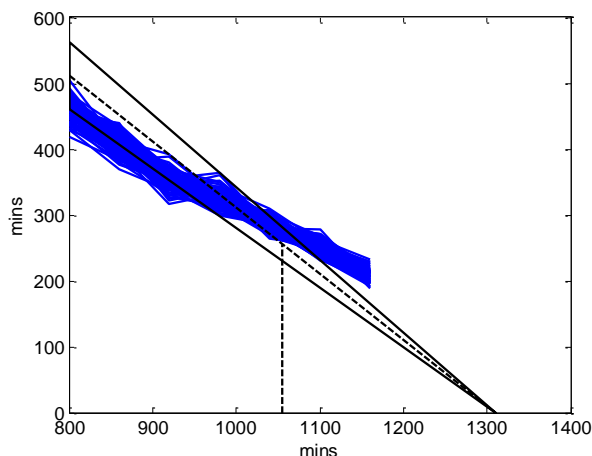
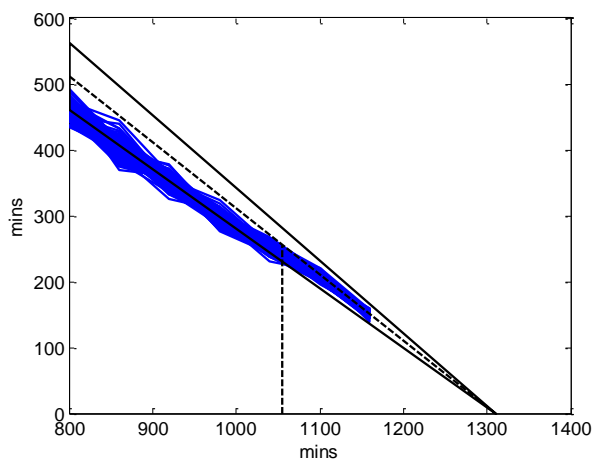


Figure 9. α - λ metric for Model 1

Figure 10. α - λ metric for Model 2Figure 11. α - λ metric for Model 3

7. CONCLUSION

In summary, this paper investigates the battery life prediction performance that result from different choice points in the model design space. This is meant as a first step in formalizing the effect of model choices with the goal of ultimately parametrizing the model design space to analyze the tradeoffs involved. Higher granularity and lower levels of abstraction might generally give more accurate models, but that also results in larger parameter sets which may not have good convergence properties if included in the state vector. To manage such models, we would need to estimate most of the parameters from training data and choose only a few for online adaptation. This predicates a higher model development cost and computational complexity. A more formal analysis of these concepts will be presented in future publications.

ACKNOWLEDGEMENT

This work was performed as a cross-center collaboration between NASA Ames and Langley Research Centers (ARC and LaRC) and Dryden Flight Research Center (DFRC). The authors would like to especially thank Edwin Koshimoto at DFRC, and Sixto L. Vazquez, Edward F. Hogge, Thomas H. Strom and Boyd L. Hill at LaRC for their contributions. The funding for this work was provided by the NASA Integrated Vehicle Health Management (IVHM) project under the Aviation Safety Program of the Aeronautics Research Mission Directorate (ARMD).

NOMENCLATURE

E	= battery voltage
ΔE	= voltage drop
E^o	= theoretical output voltage
x	= state variable
y	= measurement
t	= time
Δt	= time delay between consecutive time steps
ΔI	= change in load between consecutive time steps
α	= model parameter

REFERENCES

- Daum, F. E. & Huang, J. (2003). Curse of Dimensionality and Particle Filters. *Proceedings of IEEE Conference on Aerospace*, Big Sky, MT, 2003.
- Gao, L., Liu, S., & Dougal, R. A. (2002). Dynamic Lithium-Ion Battery Model for System Simulation. *IEEE Transactions on Components and Packaging Technologies*, vol. 25, no. 3, pp. 495-505, 2002.
- Gordon, N. J., Salmond, D. J., & Smith, A. F. M. (1993). Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation. *Radar and Signal Processing, IEE Proceedings F*, vol. 140, no. 2, pp. 107-113, 1993.
- Hartmann II, R. L. (2008). *An Aging Model for Lithium-Ion Cells*. Doctoral dissertation. University of Akron.
- Huggins, R. (2008). *Advanced Batteries: Materials Science Aspects*. 1st ed., Springer.
- Saha, B. & Goebel, K. (2009). Modeling Li-ion Battery Capacity Depletion in a Particle Filtering Framework. *Proceedings of the Annual Conference of the Prognostics and Health Management Society*, 2009, San Diego, CA.
- Saha, B., Goebel, K., Poll, S., & Christophersen, J. (2009). Prognostics Methods for Battery Health Monitoring Using a Bayesian Framework. *IEEE Transactions on Instrumentation and Measurement*, vol.58, no.2, pp. 291-296, 2009.
- Saha, B., Koshimoto, E., Quach, C., Hogge, E., Strom, T., Hill, B., & Goebel, K. (2011). Predicting Battery Life for Electric UAVs. *Proceedings of Aerospace@Infotech, AIAA*, 2011.

Santhanagopalan, S., Zhang, Q., Kumaresan, K., & White, R. E. (2008). Parameter Estimation and Life Modeling of Lithium-Ion Cells. *Journal of The Electrochemical Society*, vol. 155, no. 4, pp. A345-A353, 2008.

Saxena, A., Celaya, J., Balaban, E., Goebel, K., Saha, B., Saha, S., & Schwabacher, M. (2008). Metrics for Evaluating Performance of Prognostic Techniques. *Proceedings of Intl. Conf. on Prognostics and Health Management*, Denver, CO, Oct 2008.

Zhang, H. & Chow, M.-Y. (2010). Comprehensive Dynamic Battery Modeling for PHEV Applications. *Power and Energy Society General Meeting, IEEE*, July 2010.

Bhaskar Saha received his Ph.D. from the School of Electrical and Computer Engineering at Georgia Institute of Technology, Atlanta, GA, USA in 2008. He received his M.S. also from the same school and his B. Tech. (Bachelor of Technology) degree from the Department of Electrical Engineering, Indian Institute of Technology, Kharagpur, India. He is currently a Research Scientist with Mission Critical Technologies at the Prognostics Center of Excellence, NASA Ames Research Center. His research is focused on applying various classification, regression and state estimation techniques for predicting remaining useful life of systems and their components, as well as developing hardware-in-the-loop testbeds and prognostic metrics to evaluate their performance. He has been an IEEE member since 2008 and has published several papers on these topics.

Cuong C. Quach got his M.S. from the School of Physics and Computer Sciences at Christopher Newport University in 1997. He is a staff researcher in the Safety Critical Avionics Systems Branch at NASA Langley Research Center. His research areas include development and testing of software for airframe diagnosis and strategic flight path conflict detection.

Kai Goebel received the degree of Diplom-Ingenieur from the Technische Universität München, Germany in 1990. He received the M.S. and Ph.D. from the University of California at Berkeley in 1993 and 1996, respectively. Dr. Goebel is a senior scientist at NASA Ames Research Center where he leads the Diagnostics & Prognostics groups in the Intelligent Systems division. In addition, he directs the Prognostics Center of Excellence and he is the Associate Principal Investigator for Prognostics of NASA's Integrated Vehicle Health Management Program. He worked at General Electric's Corporate Research Center in Niskayuna, NY from 1997 to 2006 as a senior research scientist. He has carried out applied research in the areas of artificial intelligence, soft computing, and information fusion. His research interest lies in advancing these techniques for real time monitoring, diagnostics, and prognostics. He holds eleven patents and has published more than 100 papers in the area of systems health management.

Fault Diagnosis in Automotive Alternator System Utilizing Adaptive Threshold Method

Ali Hashemi¹, Pierluigi Pisu²

¹*Department of Mechanical Engineering, Clemson University, Clemson, SC 29634
ahashem@clemson.edu*

²*Department of Automotive Engineering and CU-ICAR, Greenville, SC 29607
pisup@clemson.edu*

ABSTRACT

In this paper, an observer-based adaptive threshold is developed as part of a fault diagnosis scheme to detect and isolate commonly occurring faults in a vehicle alternator system. Since the mathematical model of the alternator subsystem is quite involved and highly nonlinear; in order to simplify the diagnostic scheme, an equivalent linear time varying model based on the input-output behavior of the system is used for threshold equations derivation. A novel approach using Gaussian distribution to obtain the parameters of the system is investigated. The validity of the proposed diagnosis scheme is tested through simulation and the results are presented.

1. INTRODUCTION

Modern vehicles optimal performance is highly dependent of the reliable power generation and storage system (EPGS). Furthermore, most of the modern safety features such as X-by-wire system (Pisu, P., Rizzoni, G., et al. (2000)) are highly dependent on a smooth operation of EPGS systems. Thus, an effective diagnosis algorithm for EPGS system is necessary to maintain the optimal performance of the vehicle. Certain types of faults are commonly occurring in the alternator subsystem, namely belt slip fault, open diode fault, and voltage regulator fault. In this paper, the focus of diagnostic problem is on detecting and identifying these specific set of faults that may occur in the alternator in EPGS systems. In Scacchioli, Rizzoni, and Pisu, (2007) and Scacchioli, Rizzoni, and Pisu, (2006) model-based approaches are used to deal with the problem of fault detection and identification (FDI) for the EPGS system. In Scacchioli et al. (2006), a parity equation approach is used to compare the behaviour of the alternator with the behaviour of the equivalent model and the resulting residual

are used in the fault diagnosis algorithm design. In addition, the thresholds are derived statistically to minimize false alarms. Different methods to select the thresholds in fault detection and identification problems can be found in Ding, Guo and Frank, (1993), Ding and Guo, (1996), Emami-Naeini, Akhter, Rock, (1988), Frank, (1990), Hashemi and Pisu, (2011), Li et al., (2007), Pisu, Serrani, You and Jalics, (2006). In this paper, however, a novel method based on observer-based approach to design an adaptive threshold for a linear system with Gaussian distributed parameters is presented. Adaptive threshold changes according to the inputs to the system; thus, it has many advantages over the fixed threshold. In case of the fixed threshold, if the threshold is set too high, sensitivity to fault detection will decrease, whereas if the threshold is set too low, false alarm rate will increase. Adaptive threshold, however, does not have these problems. One downside of using adaptive threshold is its high order. Two approaches for deriving low order threshold approximations and analysis of the trade-off have been recently presented in Hashemi et al., (2011).

This paper is organized as follows. Section 2 describes the model of the system, while in section 3 the problem is formulated. Proposed fault diagnosis scheme is presented in section 4. Section 5 discussed the Gaussian distribution parameters approach. Simulation results are given in section 6. Section 7 presents the conclusion of the paper.

2. MODEL DESCRIPTION

An automotive electric-power generation storage system (EPGS) comprises two basic subsystems, the alternator and the battery, which together supply power to the vehicles electrical loads. The alternator, which is driven by the engine through a belt, provides power to the electrical loads and charges the battery. The battery, on the other hand, provides power when the engine is not running, or when the electrical power demand exceeds the alternator output. The

Ali Hashemi et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

typical alternator for an automotive electrical system comprises the following components:

- 1) AC synchronous generator
- 2) Three phase full bridge diode rectifier
- 3) Voltage regulator
- 4) Excitation field.

When the engine is running, the alternator AC voltage is rectified through the three phase bridge. The DC output voltage is regulated to be 14.4V. The role of the excitation field is to produce the field current necessary to excite the three-phase synchronous generator.

The details on mathematical model of the alternator can be found in Scacchioli et al. (2006) and details on battery mathematical model can be found in Li, Picciano, Rizzoni, Scacchioli, Pisu, and Salman, (2007).

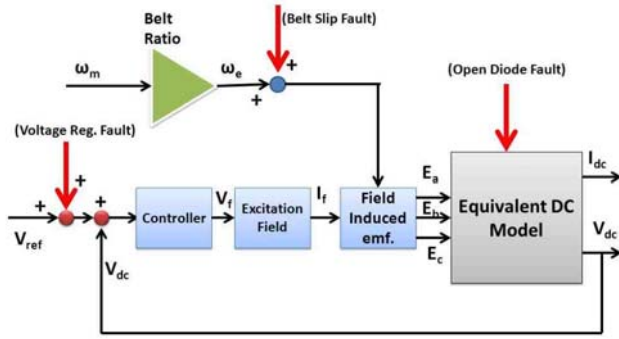


Figure 1. Functional block diagram of the automotive EPGs mathematical model with injected faults.

The mathematical model of the alternator & rectifier is highly nonlinear and complex. In order to obtain a robust diagnosis algorithm, an equivalent simpler model that still describes the behaviour of the original model in terms of input-output relations will be developed. A closer examination of the alternator subsystem shows that the behaviour of the system is functionally similar to that of a DC machine; hence, it can be modelled with an equivalent DC generator model (enclosed in the big rectangle) for the alternator and diode bridge rectifier as shown in Fig. 1.

The equations of the equivalent model are based on a DC generator, as in Eq. (1), and the equivalent excitation field, as in Eq. (2) and mentioned in Scacchioli, Li, Suozzo, Rizzoni, Pisu, Onori, Salman, and Zhang, (2010) with details.

$$\frac{dI_{dc}}{dt} = -\gamma I_{dc} + \gamma \omega_e + \kappa I_f - \lambda V_{dc} \quad (1)$$

$$\frac{dI_f}{dt} = -\alpha I_f + \beta V_f \quad (2)$$

where I_f is the alternator field current, V_f is the alternator field voltage, I_{dc} is the rectified output current, ω_e is the angular frequency of the alternator, and V_{dc} is the rectified output voltage. The parameters α , β , γ , κ , and λ are functions of ω_e . In order to obtain the variance and mean of these parameters, each parameter variation data with respect to

different speed cycles were collected. Afterwards, by fitting the proper Gaussian distribution, the variance and mean of each parameter were estimated. Note that vehicle speed and therefore ω_e can be classified into few different driving behaviours such as city driving, highway driving, cross country, etc., and parameter distributions can be pre-determined in each case. Then, in real-time, a pattern recognition algorithm can be used to identify in which class the current driving belongs therefore selecting the appropriate parameter distributions corresponding to that class.

Equations (1) and (2) in observable canonical form can be written as:

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & -\alpha\gamma \\ 1 & -(\alpha + \gamma) \end{bmatrix}}_{A_0} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \underbrace{\begin{bmatrix} \kappa\beta & \alpha\gamma & -\alpha\lambda \\ 0 & \gamma & -\lambda \end{bmatrix}}_{B_0} \begin{bmatrix} V_f \\ \omega_e \\ V_{dc} \end{bmatrix} \quad (3)$$

$$y = \underbrace{\begin{bmatrix} 0 & 1 \end{bmatrix}}_{C_0} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

where,

$$z_1 = \kappa I_f + \alpha I_{dc} \quad (4)$$

$$z_2 = I_{dc} \quad (5)$$

For the system under consideration in this paper, if for example, we denote $\alpha\gamma$ as b_{12} (corresponding element in the B matrix in Eq. (3)), the data distribution and its Gaussian fit are shown in Fig. 2.

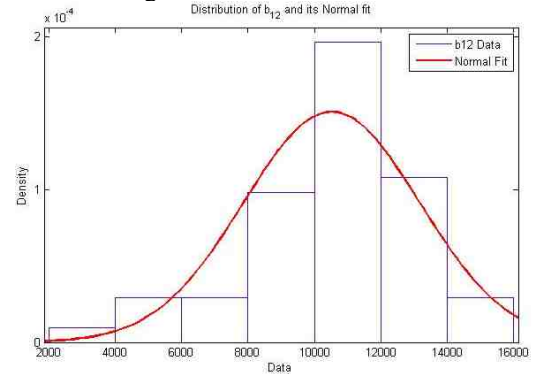


Figure 2. $b_{12}(\alpha\gamma)$ data distribution and its Normal distribution.

The formulation in Eq. (3) is later used to develop the adaptive threshold equations. In the proposed model, input signals are engine speed ω_e , the alternator voltage V_{dc} , and the excitation field voltage V_f and, alternator current I_{dc} is the output signal. This model is utilized in the design of the diagnosis scheme as described in the following sections.

3. PROBLEM FORMULATION

In this paper, the problem of detection and isolation of commonly occurring fault for the alternator in an EPGs is considered. To find a solution for this problem, a fault

diagnostic scheme part of which utilizes observer-based adaptive threshold is developed.

To this end, the following faults are considered in the system:

1) *Belt slip fault*: It is an input fault that occurs when the alternator belt does not have the proper tension to keep the alternator pulley rotating synchronously with the engine shaft. Its effect is a decrease in alternator output voltage, which the voltage regulator compensates by increasing the field voltage.

2) *Open diode rectifier fault*: This fault consists of a failure of one of the diodes in the three-phase bridge rectifier, causing unbalance in the bridge by loss of one phase. Characteristics of this type of fault are a large ripple in the output voltage and current.

3) *Voltage regulator fault*: This fault consists of a reduction in the reference voltage that produces a reduction in the alternator output current.

In the process of developing the fault diagnosis scheme, it is assumed that the faults occur separately. Moreover, to design the observer-based adaptive threshold, the measurable inputs and outputs of the system are defined. The inputs are V_{dc} , V_f , and ω_e , and the output is I_{dc} .

4. FAULT DIAGNOSIS SCHEME

The proposed diagnostic scheme combines observer design and adaptive thresholds in order to detect and isolate the three types of alternator faults (belt slip, open diode, and voltage regulator). Figure 3 shows the overall diagnosis scheme for FDI.

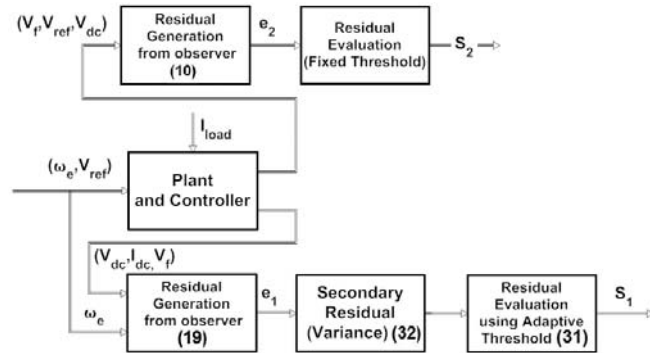


Figure 3. Fault diagnostic scheme.

The diagnostic scheme is comprised of three stages: a primary residual generation, a secondary residual generation, and a residual evaluation. The primary residual generation is constituted by the two observers generating two residuals e_1 and e_2 . A third residual is generated from e_1 by a moving standard deviation algorithm which constitutes the secondary residual generation stage. Finally, from the comparison of the residuals with thresholds two signatures S_1 , S_2 , are generated that represent the residual evaluation stage. The signature S_1 is obtained by comparing the

adaptive threshold with the variance of the residual e_1 from the first observer as described in the next section. Signature S_1 alone allows detecting all the previously described faults. For the purpose of isolation of the voltage regulator fault another signature must be introduced, namely signature S_2 . The following analysis demonstrates the method utilized to design an observer to isolate the voltage regulator fault. The alternator voltage regulator is implemented as a PI controller, with saturation on V_f that cannot be greater than V_{dc}

$$V_f = \text{sat} \left(K_p (V_{ref} - V_{dc}) + \text{sat} \left(K_I \int (V_{ref} - V_{dc}) dt \right) \right) \quad (6)$$

where K_I , and K_P are the integral and proportional controller gains. Saturation in this case is defined as:

$$\text{if } V_{dc} > V_{ref} \Rightarrow V_f = 0 \quad (7)$$

By defining $U = V_{dc} - V_{ref}$, and the state $x = K_I \int U(t) dt$, Eq. (8) away from the saturation of the integral can be represented by

$$\dot{x} = -K_I U \quad (8)$$

$$V_f = \text{sat}(x - K_P U) \quad (9)$$

Consider the observer:

$$\dot{\hat{x}} = L(V_f - \hat{V}_f) - K_I U \quad (10)$$

$$\hat{V}_f = \hat{x} - K_P U \quad (11)$$

$$e_2 = V_f - \hat{V}_f \quad (12)$$

By defining $e = \hat{x} - x$, the error dynamics in absence of faults and away from voltage saturation are

$$\dot{e} = L(V_f - \hat{V}_f) = Le_2 = -Le \quad (13)$$

In the presence of a voltage regulator fault, ΔU and no saturation conditions, we have

$$\dot{e} = Le_2 + K_I \Delta U = -Le - (LK_P - K_I) \Delta U \quad (14)$$

$$e_2 = -e - K_P \Delta U \quad (15)$$

which explicitly shows the dependence on the fault. When V_f saturates, nothing can be said about the presence of a fault.

Table 1 summarizes the fault isolation logic for the alternator fault diagnosis scheme. The main assumption in this fault diagnosis scheme is that faults are not occurring concurrently.

Fault type	S1	S2
No Fault	0	0
Belt Slip	1	0
Open Rectifier Diode Fault	1	0
Voltage Regulator Fault	1	1

Table 1. Error signature for the Alternator System

In Table 1, a “zero” means ‘residual does not cross the threshold’; while a “one” means ‘residual crosses the

threshold'. With the current scheme all faults are detectable but belt slip fault cannot be distinguished from diode fault.

5. ADAPTIVE THRESHOLDS IN THE CASE OF GAUSSIAN DISTRIBUTED PARAMETERS

To obtain the signature S_1 , an observer-based adaptive threshold is designed based on the state space representation of the equivalent DC generator Eqs. (1) and (2). Details of the derivation for a general case are shown below.

Consider a general state space presentation of a system with n states in observable canonical form:

$$\frac{dz}{dt} = \underbrace{\begin{bmatrix} 0 & 0 & \dots & 0 & -a_0 \\ 1 & 0 & \dots & 0 & -a_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & -a_{n-2} \\ 0 & 0 & \dots & 1 & -a_{n-1} \end{bmatrix}}_{A_0} z + \underbrace{\begin{bmatrix} b_{00} & b_{01} & \dots & b_{0,m-1} \\ b_{10} & b_{11} & \dots & b_{1,m-1} \\ \dots & \dots & \dots & \dots \\ b_{n-1,0} & b_{n-1,1} & \dots & b_{n-1,m-1} \end{bmatrix}}_{B_0} u \quad (16)$$

$$y = \underbrace{\begin{bmatrix} 0 & 0 & \dots & 1 \end{bmatrix}}_{C_0} z \quad (17)$$

where $z \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}$, $A_0 \in \mathbb{R}^{n \times n}$, $B_0 \in \mathbb{R}^{n \times m}$, and $C_0 \in \mathbb{R}^{1 \times n}$. Assuming parameters uncertainties, Eqs. (18) and (19) can be written as:

$$\dot{z} = (A_0 + \Delta A_0)z + (B_0 + \Delta B_0)u \quad (18)$$

$$y = C_0 z$$

Notice that $\Delta A_0 z = \Delta \underline{a} y$ with $\Delta \underline{a} = [\Delta a_0 \quad \Delta a_1 \quad \dots \quad \Delta a_{n-1}]^T$. An observer can be designed for Eq. (20) as below:

$$\begin{aligned} \dot{\hat{z}} &= A_0 \hat{z} + B_0 u + L(y - \hat{y}) \\ \hat{y} &= C_0 \hat{z} \end{aligned} \quad (19)$$

With $L = [l_1 \quad l_2 \quad \dots \quad l_n]^T$ to be defined so that the eigenvalues of $A_0 + LC_0$ are all negative and real.

By defining $e = z - \hat{z}$, the error dynamics can be written as:

$$\begin{aligned} \dot{e} &= (A_0 + LC_0)e - \Delta \underline{a} y + \Delta B_0 u \\ e_1 &= y - \hat{y} = C_0 e \end{aligned} \quad (20)$$

where Δa , and ΔB_0 are parameters uncertainties defined as normally distributed random variables with zero mean and known variance. Define p as,

$$p = [\Delta a_1, \Delta a_2, \dots, \Delta a_n, \Delta b_{0,0}, \Delta b_{0,1}, \dots, \Delta b_{0,m-1}, \Delta b_{1,0}, \dots, \Delta b_{1,m-1}, \dots, \Delta b_{n-1,m-1}] \in N(0, Q) \quad (21)$$

where Q is the covariance matrix defined as

$$Q = E\{P_j P_k\} \quad \forall j, k = (m+1) \times n \quad (22)$$

The solution of the differential equation for the error dynamics given by Eq. (20) is

$$\begin{aligned} e_1(t) &= C_0 e^{(A_0 + LC_0)t} e(0) + C_0 \int_0^t e^{(A_0 + LC_0)(t-\tau)} \sum_{i=1}^n E_i \Delta a_i y(\tau) d\tau \\ &+ C_0 \int_0^t e^{(A_0 + LC_0)(t-\tau)} \sum_{i=1}^n \sum_{j=0}^{m-1} E_i \Delta b_{i-1,j} u_j(\tau) d\tau \end{aligned} \quad (23)$$

where E_i have been introduced to write the solution in a compact form and are simply defined by

$$E_1 = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}^T, \quad E_i = \underbrace{\begin{bmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{bmatrix}^T}_{i^{th} \text{ position}}$$

for $i=2, \dots, n$.

By switching the summations with the integral, we have

$$\begin{aligned} e_1(t) &= C_0 e^{(A_0 + LC_0)t} e(0) + \sum_{i=1}^n C_0 \int_0^t e^{(A_0 + LC_0)(t-\tau)} E_i \Delta a_i y(\tau) d\tau \\ &+ \sum_{i=1}^n \sum_{j=0}^{m-1} C_0 \int_0^t e^{(A_0 + LC_0)(t-\tau)} E_i \Delta b_{i-1,j} u_j(\tau) d\tau \end{aligned} \quad (24)$$

Since the parameters have zero mean, the expected value of Eq. (24) can be easily calculated

$$E\{e_1(t)\} = C_0 e^{(A_0 + LC_0)t} e(0) = \varepsilon_0 \quad (25)$$

that can be made vanish at any desired rate by an appropriate selection of the matrix L .

Considering auxiliary filters, that need to be found, for the threshold implementation, the last state n of these filters can be defined as

$$\xi_{i,n} = \int_0^t C_0 e^{(A_0 + LC_0)(t-\tau)} E_i y(\tau) d\tau \quad (26)$$

$$\psi_{ij,n} = \int_0^t C_0 e^{(A_0 + LC_0)(t-\tau)} E_i u_j(\tau) d\tau \quad (27)$$

Utilizing Eq. (22), (26) and (27), the variance of Eq. (24) can be easily written as

$$\begin{aligned} Var\{e_1(t)\} &= E\{(e_1 - E(e_1))^2\} = \\ &= E\left\{\left(\sum_{i=1}^n \Delta a_{i-1} \xi_{i,n} + \sum_{i=1}^n \sum_{j=0}^{m-1} \Delta b_{i-1,j} \psi_{ij,n}\right)^2\right\} \end{aligned} \quad (28)$$

$$= [\xi_{1,n} \dots \xi_{n,n}, \psi_{10,n} \dots \psi_{nm-1,n}] Q [\xi_{1,n} \dots \xi_{n,n}, \psi_{10,n} \dots \psi_{nm-1,n}]^T$$

If we define $\Theta^T = [\xi_{1,n} \dots \xi_{n,n}, \psi_{10,n} \dots \psi_{nm-1,n}]$,

according to Rayleigh-Ritz theorem, an upperbound of the variance can be obtained as

$$Var\{e_1(t)\} = |\Theta^T(t) Q \Theta(t)| \leq \lambda_{\max} \|\Theta(t)\|_2^2 = z_{th}(t) \quad (29)$$

with $\lambda_{\max} = \max\{\text{eigenvalue}(Q)\}$. This upperbound constitutes the adaptive threshold dynamics.

The state space representation of the adaptive threshold z_{th} in Eq. (29) can be obtained by observing that Eq. (26) and (27) are the outputs of linear filters described by the triplet $(A_0 + LC_0, E_i, C_0)$. Therefore, z_{th} can be implemented as follows

$$\begin{aligned}\dot{\xi}_i &= (A_0 + LC_0)\xi_i + E_i y(t) \quad i=1,2,\dots,n \\ \xi_{i,n} &= C_0 \xi_i \\ \dot{\psi}_{ij} &= (A_0 + LC_0)\psi_{ij} + E_i u_j(t) \\ \psi_{ij,n} &= C_0 \psi_{ij}, \quad i=1,2,\dots,n \quad j=0,1,2,\dots,m-1 \\ z_{th}(t) &= \varepsilon_0 + \bar{\lambda}_{\max} \left(\sum_{i=1}^n \xi_{i,n}^2 + \sum_{j=0}^{m-1} \psi_{ij,n}^2 \right)\end{aligned}\quad (30)$$

where $\xi_i \in \mathbb{R}^n, \psi_{ij} \in \mathbb{R}^n, \forall i=1..n, j=0..m-1$ are the states of the auxiliary filters mentioned before with $\xi_{i,n}$ and $\psi_{ij,n}$ satisfying Eq. (26) and Eq. (27) respectively, $\xi_i(0) = 0, \psi_{ij}(0) = 0$, and $\bar{\lambda}_{\max}$ an upperbound of λ_{\max} .

In this case a fault is declared if $Var\{e_1(t)\} > z_{th}(t)$ which corresponds to signature $S_1=1$. The threshold just derived can be seen as $(m+1) \times n$ filters of order n . The high order of the threshold dynamics is the main drawback. The order can be further reduced to $m+1$ filters of order n by transforming the equations from observable form into controllable form, and combining the equations with the same input as shown in Eq. (31)

$$\begin{aligned}\dot{\gamma} &= (A_0 + LC_0)^T \gamma + C_0^T y(t) \\ \xi_{i,n} &= E_i^T \gamma \\ \dot{\psi}_j &= (A_0 + LC_0)^T \psi_j + C_0^T u_j(t) \\ \psi_{ij,n} &= E_i^T \psi_j \quad i=1..n; j=0..m-1 \\ Z_{th}(t) &= \varepsilon_0 + \bar{\lambda}_{\max} \left(\sum_{i=1}^n (\xi_{i,n})^2 + \sum_{j=0}^{m-1} (\psi_{ij,n})^2 \right)\end{aligned}\quad (31)$$

where $\gamma \in \mathbb{R}^n, \psi_j \in \mathbb{R}^n$.

As mentioned before, the signature S_1 is obtained by comparing the adaptive threshold with the variance of the residual e_1 . The variance of residual e_1 is here estimated by means of a recursive standard deviation algorithm described by

$$\begin{aligned}(STD_{k+1})^2 &= \sum_{i=k+2-N}^{k+1} \frac{(e_{1,i} - \mu_i)^2}{N-1} \\ &= (STD_k)^2 + \frac{(e_{1,k+1} - \mu_{k+1})^2}{N-1} - \frac{(e_{1,k+1-N} - \mu_{k+1-N})^2}{N-1}\end{aligned}\quad (32)$$

$$\mu_k = \sum_{i=k+1-N}^k \frac{e_{1,i}}{N} = \mu_{k-1} + \frac{e_{1,k}}{N} - \frac{e_{1,k-N}}{N}\quad (33)$$

where μ_k is the mean value of the residual and N is the moving window. Here a 1s moving window which contains 10,000 sampling points was considered for the implementation of the standard deviation (STD) algorithm.

Note that, as mentioned in section 2, parameter distributions and the corresponding covariance matrix can be pre-calculated for different classes of driving conditions (city, highway, etc.). An upperbound $\bar{\lambda}_{\max}$ can then be evaluated in each case and stored on board of the vehicle. A pattern recognition algorithm, like the one presented by Bo and Rizzoni (2006), can then be used to determine the current driving conditions and select the appropriate value of $\bar{\lambda}_{\max}$.

7. SIMULATION RESULTS

In order to test the effectiveness of the proposed diagnosis scheme, a system simulator was developed and the three different faults were injected into the system. The simulation time considered was 72s during which a portion of the Federal Urban Driving Schedule, Fig.4, was used to simulate the urban driving condition of the actual driving. Each fault is injected separately after 10s into the system. The belt slip and the voltage regulator faults are modeled as additive faults. The belt slip fault amount is 0.4 of the engine speed, and the voltage regulator is 0.3 of the nominal value of the voltage regulator. The residual along with thresholds plots are presented here. These plots show the effectiveness of the proposed fault diagnosis scheme in detecting and isolating the faults. This approach is capable in detecting the voltage regulator fault as it occurs whereas the belt slipping fault and open diode fault are detected at time 30s. That is when the input current takes effect combined with the change in speed. However, due to characteristics of the particular alternator chosen for this simulation, the movement of the threshold is limited. For S_2 signature, fixed thresholds at 13000, and -13000 are chosen as shown in Fig.8, Fig.10, and Fig.12. Figures 6, 7, 9 and 11 show the simulation results utilizing Gaussian distributed parameters threshold in order to obtain S_1 signature for the urban driving cycle. As it can be seen, this type of threshold is capable of detecting the fault when they occur specially in the case of voltage regulator fault. For the diode and belt slip fault, the detection occurs corresponding to a change of current load (Fig. 5).

One final note, this scheme can detect the belt slipping fault with fault amount as low as 30% with respect to the nominal value of the electrical frequency. Voltage regulator fault can be detected as low as 11% with respect to the nominal value of the voltage reference.

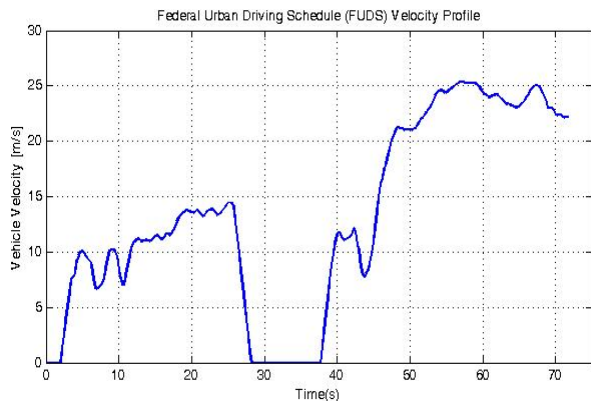


Figure 4. Federal Urban Driving Schedule.

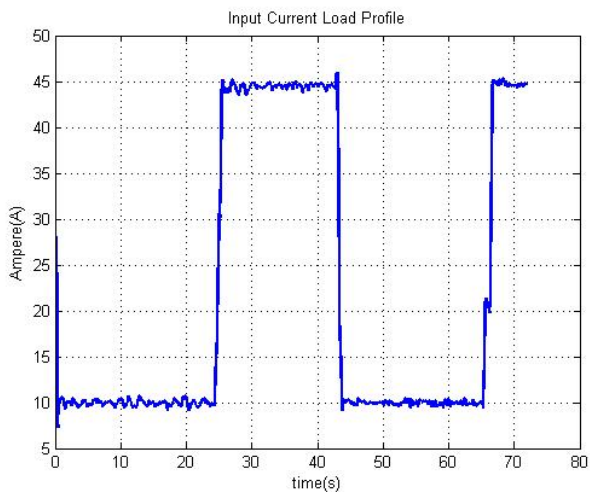


Figure 5. Current load profile.

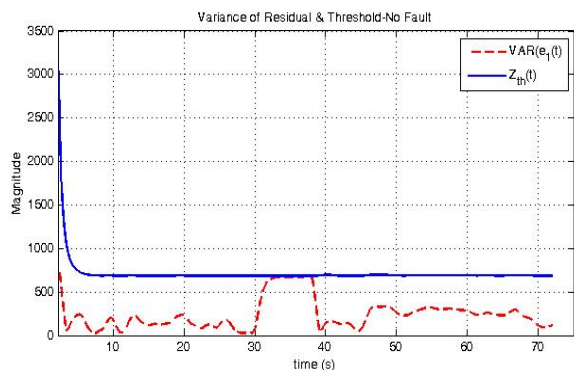


Figure 6. Residual of S_1 signal when no fault is injected.

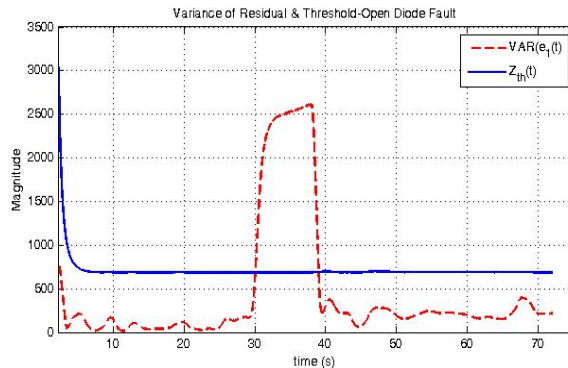


Figure 7. Residual of S_1 signal for open diode fault.

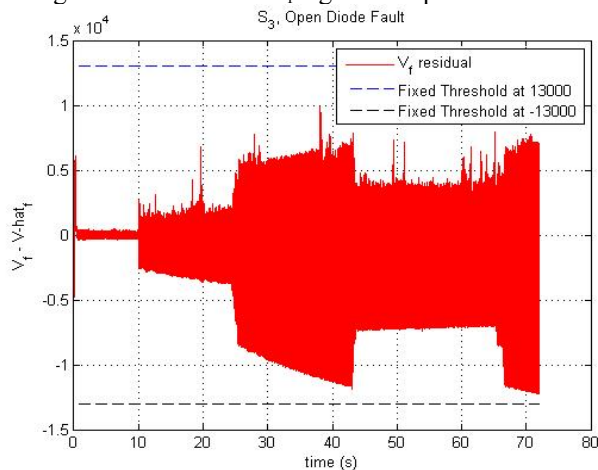


Figure 8. Residual of S_2 signal for open diode fault.

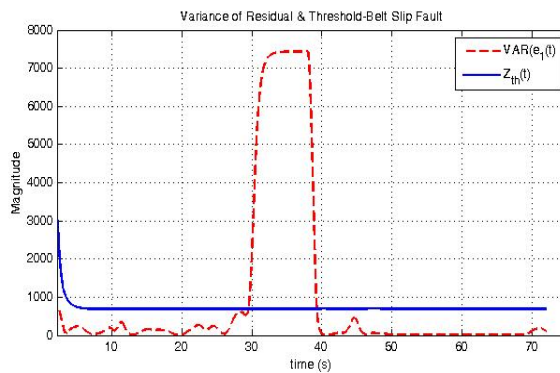


Figure 9. Residual of S_1 signal for belt slip fault.

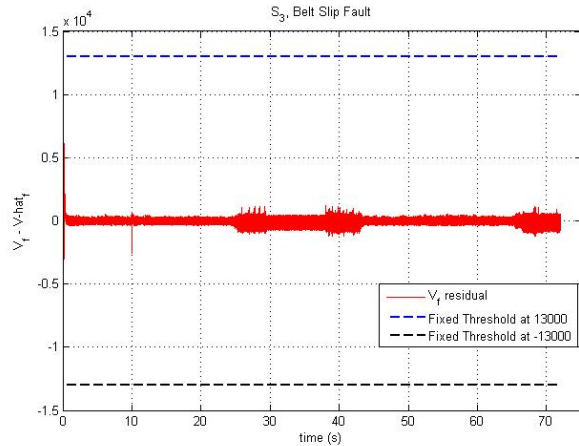


Figure 10. S_2 residual when belt slip fault is injected.

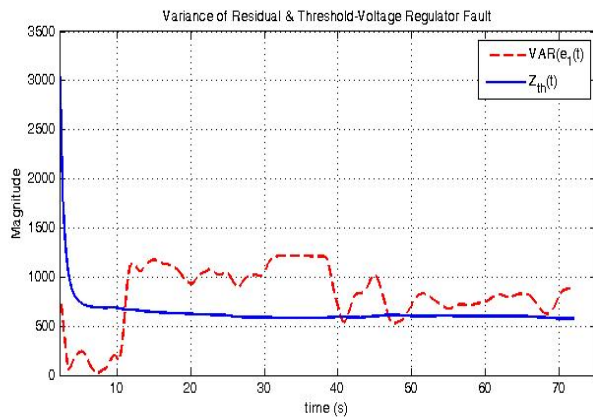


Figure 11. S_1 residual with voltage regulator fault.

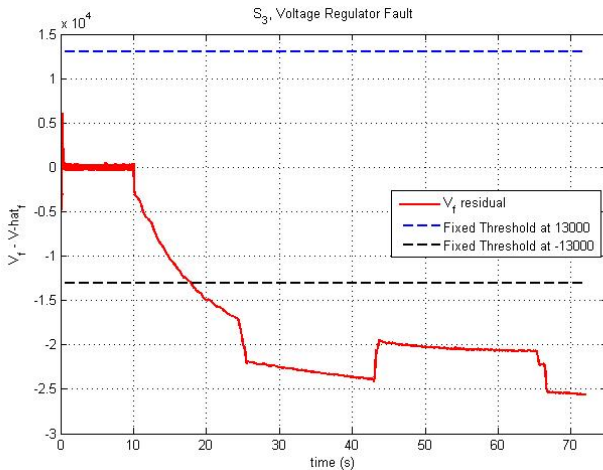


Figure 12. S_2 residual with voltage regulator fault.

8. CONCLUSION

This paper demonstrates the utilization of an adaptive threshold approach in designing a fault diagnosis scheme for the alternator subsystem in the EPGS system. An equivalent DC generator model was used in obtaining the

observer-based adaptive threshold for the fault diagnosis scheme. Simulation results show that the proposed fault diagnosis scheme is effective in detecting and identifying the faults occurring in the alternator. Furthermore, the Gaussian distributed parameters adaptive threshold shows its effectiveness in detecting the faults occurring in the system and obtaining S_1 error signature.

ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant No. 0825655.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

- Bo, G. and Rizzoni, G., (2006). An adaptive algorithm for hybrid electric vehicle energy management based on driving pattern recognition. In Proceedings of the IMECE 2006, Chicago, IL.
- Demerly, J. D., Toumi, K. Y., (2000). Non-Linear Analysis of Vehicle Dynamics (NAVDyn): A Reduced-Order Model for Vehicle Handling Analysis. SAE Paper 2000-01-1621.
- Ding, X., Guo, L., Frank, P.M., (1993). A frequency domain approach to fault detection of uncertain dynamic systems," In Proceedings of the 32nd Conference on Decision and Control, pages 1722-1727, San Antonio, TX.
- Ding, X., Guo, L. (1996). Observer based (optimal) fault detector. In Proceedings of the 13th IFAC World Congress, volume N, pages 187-192, San Francisco, CA.
- Emami-Naeini, A., Akhter, M.M. Rock, S.M., (1988). Effect of model uncertainty on failure detection - The threshold selector. IEEE-Trans. on Aut. Control AC-33 12 (1988), pp. 1106-1115.
- Frank, P.M., (1990). Fault Diagnosis in Dynamic Systems Using Analytical and Knowledge-Based Redundancy—A Survey and Some New Results. Am.J. Cardiol., 26, pp. 459-474.
- Hashemi, A. and Pisu, P., (2011). Adaptive threshold-Based fault detection and isolation for automotive electrical systems, accepted to WCICA2011.
- Li, W., Picciano, N., Rizzoni, G., Scacchioli, A., Pisu, P., Salman, M., (2007). Fault Diagnosis of Lead-Acid Battery for Automotive Electrical Systems. Conference Proceedings of SAE, Detroit, MI.
- Li, W., Suozzo, C., Rizzoni, G., (2008). Experimental calibration and validation of fault diagnosis of automotive electric power generation system. Proceedings of ASME 2008 Dynamic Systems and

Control Conference, Ann Arbor, Michigan, pp. 1317-1324.

- Pisu, P., Rizzoni, G., et al. (2000). Model Based Diagnostics for Vehicle Systems. Proceedings of the 2000 International Mechanical Engineering Congress & Exposition, Orlando, FL.
- Pisu, P., Serrani, A., You, S., Jalics, L., (2006). Adaptive Threshold Based Diagnostics for Steer-By-Wire Systems. ASME Trans. on Dynamics Systems, Measurement and Control, 128(2): pp. 428-435.
- Scacchioli, A., Rizzoni, G., Pisu, P. (2006). Model-Based Fault Diagnosis for an Electrical Automotive System. Conference Proceedings of ASME 2006, Chicago, IL.
- Scacchioli, A., Rizzoni, G., Pisu, P., (2007). Hierarchical Model-Based Fault Diagnosis for an Electrical Power Generation Storage Automotive System. Proceedings of the 26th IEEE American Control Conference, pp. 2991-2996, New York City, NY.
- Scacchioli, A., Li, W., Suozzo, C., Rizzoni, G., Pisu, P., Onori, S., Salman, M., Zhang, X., (2010). Experimental Implementation of a Model-Based Diagnosis for an electric Power Generation and Storage Automotive System, Submitted ASME Tran. DSMC.

Ali Hashemi was born in Tehran, Iran in 1984. He is currently a masters student in Mechanical Engineering at Clemson University.



Pierluigi Pisu was born in Genoa, Italy in 1971. He received his Ph.D. in Electrical Engineering from the Ohio State University (Columbus, Ohio, USA) in 2002. In 2004, he was granted two US patents in area of model-based fault detection and isolation. He is currently Assistant Professor at the Department of Automotive Engineering at Clemson University. His research interests are in the area of fault diagnosis with application to vehicle systems, and energy management control of hybrid electric vehicles; he also worked in the area of sliding mode control and robust control. He is member of the ASME and SAE, and a recipient of the 2000 Outstanding Ph.D. Student Award by the Ohio State University Chapter of the Honor Society of Phi Kappa Phi.

Fault-Tolerant Trajectory Tracking Control of a Quadrotor Helicopter Using Gain-Scheduled PID and Model Reference Adaptive Control

Iman Sadeghzadeh¹, Ankit Mehta², Youmin Zhang^{3*} and Camille-Alain Rabbath⁴

^{1,2,3}*Concordia University, Montreal, Quebec, H3G 1M8, Canada*

i_sade@encs.concordia.ca
an_mehta@encs.concordia.ca
**ymzhang@encs.concordia.ca*

⁴*Defence Research and Development Canada, Valcartier, G3J 1X5 Quebec, Canada*

Camille-Alain.Rabbath@drdc-rddc.gc.ca

ABSTRACT

Based on two successfully and widely used control techniques in many industrial applications under normal (fault-free) operation conditions, the Gain-Scheduled Proportional-Integral-Derivative (GS-PID) control and Model Reference Adaptive Control (MRAC) strategies have been extended, implemented, and experimentally tested on a quadrotor helicopter Unmanned Aerial Vehicle (UAV) test-bed available at Concordia University, for the purpose of investigation of these two typical and different control techniques as two useful Fault-Tolerant Control (FTC) approaches. Controllers are designed and implemented in order to track the desired trajectory of the helicopter in both normal and faulty scenarios of the flight. A Linear Quadratic Regulator (LQR) with integral action controller is also used to control the pitch and roll motion of the quadrotor helicopter. Square trajectory, together with specified autonomous and safe taking-off and landing path, is considered as the testing trajectory and the experimental flight testing results with both GS-PID and MRAC are presented and compared with tracking performance under partial loss of control power due to fault/damage in the propeller of the quadrotor UAV. The performance of both controllers showed to be good. Although GS-PID is easier for development and implementation, MRAC showed to be more robust to faults and noises, and is friendly to be applied to the quadrotor UAV.

1. INTRODUCTION

Safety, reliability and acceptable level of performance of dynamic control systems are key requirements in control systems not only in normal operation conditions but also in the presence of partial fault or failure in the components of the controlled system. Hence, the role of Fault-Tolerant

Control Systems (FTCS) is revealed evidently (Zhang & Jiang, 2008). In fact, when a fault occurs in a system, it suddenly starts to behave in an unanticipated manner with the originally designed baseline controller(s) under normal conditions. Therefore, fault-tolerant controller must be designed, implemented and executed on-line and in real-time to be able to handle the fault and to guarantee system stability and acceptable performance even in the presence of faults in actuators, sensors and other system components.

There are different techniques to handle such faults. As one of adaptive control techniques, Model Reference Adaptive Control (MRAC) is one of the recently widely investigated techniques for handling different fault situations with different types of aircraft applications as demonstrated in the recent AIAA Guidance, Navigation, and Control Conference (Bierling, Hocht, & Holzapfel, 2010; Crespo, Matsutani, & Annaswamy, 2010; Dydek & Annaswamy, 2010; Gadiant, Levin, & Lavretsky, 2010; Gregory, Gadiant, & Lavretsky, 2011; Guo & Tao, 2010; Jourdan et al, 2010; Lemon, Steck, & Hinson, 2010; Levin, 2010; Stepanyan, Campbell, & Krishnakumar, 2010; Whitehead & Bieniawski, 2010). MRAC is concerned with forcing the dynamic response of the controlled system to asymptotically approach that of reference system, despite parametric uncertainties in the plant. In fact, adaptive control is originally a control technique which bases on a concept that controllers must adapt to a controlled system with parameters which vary slowly, or are initially uncertain. For example, as an aircraft flies, its mass will slowly decrease as a result of fuel consumption. To maintain good control performance under such varying conditions, an adaptive control law is needed to adapt itself to such changing conditions. Based on its adaptive and self-tuning capability in the presence of system parameters changes, including such changes due to faults/damages, there are a trend for

¹Ph.D. Student, Department of Mechanical and Industrial Engineering

²M.Eng Student, Department of Mechanical and Industrial Engineering

³Engineering, Corresponding Author

⁴Defense Scientist, DRDC - Valcartier, 2459 Pie-XI Blvd. North

investigating the potential application of MRAC for fault-tolerant control of aircraft and UAVs recently. However, there is no published research result for using MRAC to fault-tolerant tracking control of quadrotor helicopter UAVs, which in fact motivated the work to be presented in this paper.

On the other hand, Proportional-Integral-Derivative (PID) controllers are the most widely used controllers in industry due to its unique feature without the need of a mathematical model of the controlled system for controller design, implementation and real-time execution. PID controllers are reliable and easy to use and can be used for linear and non-linear systems with certain level of robustness to the uncertainties and disturbances. Although one single PID controller can handle even wide range of system nonlinearities, to handle the possible fault conditions of a quadrotor helicopter UAV, multiple PIDs need to be designed to control the quadrotor helicopter UAV with acceptable performance under both normal and different faulty flight conditions. For such a purpose, the Gain-Scheduled PID (GS-PID) control strategy was initially proposed to be applied to a quadrotor helicopter UAV for achieving fault-tolerant control by Bani Milhim, Zhang, & Rabbath (2010). However, such a work was based only on simulation due to the lack of a physical UAV test-bed at that time. At the same conference of the 2010 AIAA Infotech@Aerospace, Johnson, Chowdhary, & Kimbrell (2010) also investigated a GS-PID scheme to their GTech Twinstar fixed-wing research vehicle.

In view of the advantages and potentials of using GS-PID for handling fault conditions, it motivated us to further investigate and most importantly to experimentally test the GS-PID controller in a physical quadrotor UAV test-bed at the Networked Autonomous Vehicles Lab of Concordia University, for fault-tolerant three-dimensional trajectory tracking control, instead of implementing the GS-PID only for one-dimensional height hold flight conditions. In this paper, GS-PID has been implemented for different sections of the entire flight envelope by properly tuning the PID controller gains for both normal and fault conditions. A Fault Detection and Diagnosis (FDD) scheme is assumed to be available for providing the time and the magnitude of the fault during the flight. Based on the decision of the FDD scheme about the fault occurring in the UAV during flight, the GS-PID controller will switch the controller gains under normal flight conditions to the pre-tuned and fault-related gains to handle the faults during the flight of the UAV.

During recent years, Unmanned Aerial Vehicles (UAVs) have proved to hold a significant role in the world of aviation. These UAVs also provide the academic and industrial researchers and developers feasible and low-cost test-beds for fault-tolerant control techniques development and flight testing verification (Jordan, et al, 2006; Jourdan et al, 2010; Gregory, Gadiant, & Lavretsky, 2011), which was

extremely difficult and costly by using manned aircraft, since flight testing verification with UAVs does not involve the main concern and the burden for flight testing the developed fault-tolerant control algorithms with human pilot sitting on the manned aircraft/aerial vehicles. These facts motivated also us for building and testing our developed fault-tolerant control algorithms with UAVs through financial supports of NSERC (Natural Sciences and Engineering Research Council of Canada) through a Strategic Project Grant (SPG) and a Discovery Project Grant (DPG) since 2007 leading by the third author. With consideration of an UAV with both in-door and out-door flying capability, a rotorcraft-type UAV, instead of a fixed-wing UAV as developed in the above-mentioned NASA (National Aeronautics and Space Administration) and DRAPA (Defense Advanced Research Projects Agency) sponsored projects in USA (Jordan, et al, 2006; Jourdan et al, 2010), was selected for such an UAV test-bed development and flight tests. Among the rotorcrafts, quadrotor helicopters can usually afford a larger payload than conventional helicopters due to their four-rotor configuration. Moreover, small quadrotor helicopters possess a great manoeuvrability and are potentially simpler to manufacture. For these advantages, quadrotor helicopters have received much and continuously increasing interest in UAV research, development, and applications. The quadrotor helicopter we consider in this work is an under-actuated system with six outputs and four inputs and the states are highly coupled. There are four fixed-pitch-angle blades whereas single-rotor helicopters have variable-pitch-angle (collective) blades.

Control of a quadrotor helicopter UAV is performed by varying the speed of each rotor. The configuration, structure, and related hardware/software of a quadrotor, especially the Quanser quadrotor unmanned helicopter, called as Qball-X4, which is used as the test-bed of this paper's work and was developed in collaboration between Concordia University and Quanser Inc. through an NSERC Strategic Project Grant (SPG), will be presented in the Section 2 of this paper. Nonlinear and linearized state-space models are presented in Section 3 for the purpose of controller design with MRAC. Descriptions of the GS-PID and MRAC with applications to the Qball-X4 are presented in Section 4 and Section 5, respectively. Experimental flight testing results and comparison between GS-PID and MRAC are presented in Section 6. The conclusion and our future work are outlined in Section 7.

2. GENERAL AND QBALL-X4 QUADROTOR HELICOPTER STRUCTURE

In Fig. 1, the conceptual demonstration of a quadrotor helicopter is shown. Each rotor produces a lift force and moment. The two pairs of rotors, i.e., rotors (1, 3) and rotors (2, 4) rotate in opposite directions so as to cancel the

moment produced by the other pair. To make a roll angle (ϕ) along the x -axis of the body frame, one can increase the angular velocity of rotor (2) and decrease the angular velocity of rotor (4) while keeping the whole thrust constant. Likewise, the angular velocity of rotor (3) is increased and the angular velocity of rotor (1) is decreased to produce a pitch angle (θ) along the y -axis of the body frame. In order to perform yawing motion (ψ) along the z -axis of the body frame, the speed of rotors (1, 3) is increased and the speed of rotors (2, 4) is decreased.

The quadrotor helicopter is assumed to be symmetric with respect to the x and y axes so that the center of gravity is located at the center of the quadrotor and each rotor is located at the end of bars.

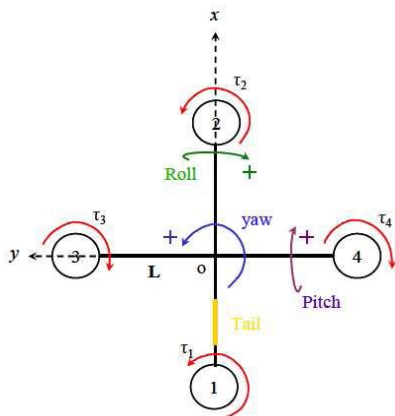


Figure 1. Quadrotor helicopter configuration with Roll-Pitch-Yaw Euler angles [ϕ , θ , ψ]

The quadrotor made by Quanser, known as Qball-X4 as shown in Fig. 2, is an innovative rotary-wing aerial vehicle platform suitable for a wide variety of UAV research and development applications. The Qball-X4 is a quadrotor helicopter propelled by four motors fitted with 10-inch propellers. The entire quadrotor is enclosed within a protective carbon fibre cage for the safety concern during flight to the quadrotor itself and for personnel using it in an in-door environment with limited flying space.

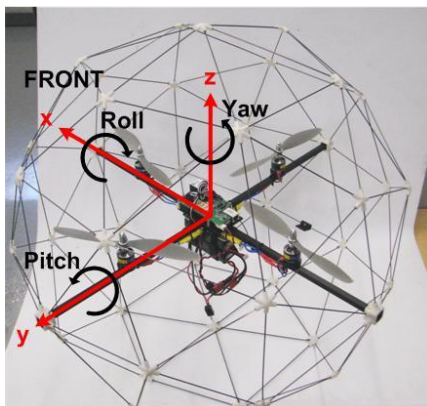


Figure 2. The Qball-X4 quadrotor UAV (Quanser, 2010)

The Qball-X4's proprietary design ensures safe operation as well as opens the possibilities for a variety of novel applications. The protective cage is a crucial feature since this unmanned aerial vehicle was designed for use in an indoor environment/laboratory, where there are typically many close-range hazards (including other vehicles) and personnel doing flight tests with the Qball-X4. The cage gives the Qball-X4 a decisive advantage over other vehicles that would suffer significant damage if contact occurs between the vehicle and an obstacle. To obtain the measurement from on-board sensors and to drive the motors connected to the four propellers, the Qball-X4 utilizes Quanser's onboard avionics Data Acquisition Card (DAQ), the HiQ, and the embedded Gumstix computer. The HiQ DAQ is a high-resolution Inertial Measurement Unit (IMU) and avionics Input/Output (I/O) card designed to accommodate a wide variety of research applications. QuaRC, Quanser's real-time control software, allows researchers and developers to rapidly develop and test controllers on actual hardware through a MATLAB/Simulink interface. QuaRC's open-architecture hardware and extensive Simulink blockset provides users with powerful control development tools. QuaRC can target the Gumstix embedded computer automatically to generate code and execute controllers on-board the vehicle. During flights, while the controller is executing on the Gumstix, users can tune parameters in real-time and observe sensor measurements from a host ground station computer (PC or laptop) (Quanser, 2010).

The interface to the Qball-X4 is MATLAB/Simulink with QuaRC. The controllers are developed in Simulink with QuaRC on the host computer, and these models are downloaded and compiled into executable codes on the target (Gumstix) seamlessly. A diagram of this configuration is shown in Figure 3.

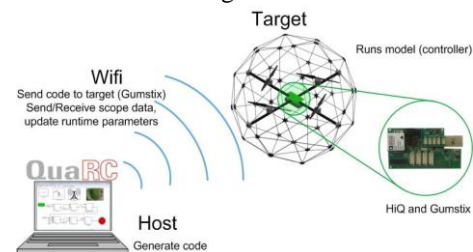


Figure 3. The Qball-X4 communication hierarchy and communication diagram (Quanser, 2010)

For Qball-X4, the following hardware and software are embedded:

- **Qball-X4:** as shown in the Figure 2.
- **HiQ:** QuaRC aerial vehicle data acquisition card (DAQ).
- **Gumstix:** The QuaRC target computer. An embedded, Linux-based system with QuaRC runtime software installed.
- **Batteries:** Two 3-cell, 2500 mAh Lithium-Polymer batteries.

- **Real-Time Control Software:** The QuaRC-Simulink configuration, as detailed in Quanser (2010).

3. MODELING OF THE QBALL-X4

3.1 Non-linear Model of the Qball-X4

In Qball-X4, there are four (E-flite Park 400) brushless motors, using a 10×4.7 inch propeller. As explained before, in order to cancel the moment of each pair of propellers, the motors 1 and 2 have clockwise rotation and the motors 3 and 4 have counterclockwise rotation.

For every attitude change the angular velocity of motors is changed, but the total thrust of all the four motors is constant in order to maintain the height. For instant, to make a pitch angle (θ) along the Y -axis of the body frame one can increase the angular velocity of motor (2) and increase the angular velocity of motor (1), while keeping the trust constant. Likewise the angular velocity of motor (3) is increased and the angular velocity of motor (4) is decreased in order to make a roll angle (ϕ) along the X -axis of the body frame.

It can be understood easily that yaw motion along the Z -axis of the body frame will be implemented by increasing total angular velocity of motors (1, 2) and decreasing the angular velocity of opposite rotation motors (3, 4). Motors of Qball-X4 are not exactly located at the end of the aluminum rods, but 6 inches from the end point for not to touch the fiber carbon cage by propellers and the L is the length between the rotational axis of each motor/rotor and the center of gravity (CoG) of the Qball-X4, as shown in Fig. 4.

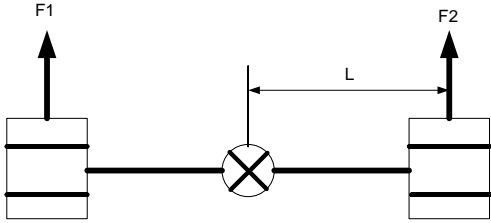


Figure 4. Roll/Pitch axis model

While flying there are four downwash thrust vectors generated by four propellers, if we neglect the drag of four propellers we can present the equations of motion of the Qball-X4 as follows:

$$\begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{bmatrix} = \frac{1}{m} \left(\sum_{i=1}^4 F_i \right) \text{Re}_3 + (g_r(z) - g) e_3 \quad (1)$$

$$\begin{aligned} \ddot{\phi} &= l(F_3 - F_4) / J_1 \\ \ddot{\theta} &= l(F_1 - F_2) / J_2 \\ \ddot{\psi} &= \rho(F_1 + F_2 - F_3 - F_4) / J_3 \end{aligned} \quad (2)$$

where J is the moment of inertia with respect to each axis and ρ is the force-to-moment scaling factor; $[x, y, z]$ are the position of the quadrotor in earth position and $[\phi, \theta, \Psi]$ are roll, pitch and yaw angle respectively.

As mentioned before, we need a transformation matrix which transforms variables from body frame to the Earth frame. Therefore, R represents the coordinate transformation matrix from body frame to earth frame and $e_3 = [0, 0, 1]^T$.

$$R = \begin{bmatrix} \cos \theta \cos \psi & \sin \phi \sin \theta \cos \psi - \cos \phi \sin \psi & \cos \phi \sin \theta \cos \psi + \sin \phi \sin \psi \\ \cos \theta \sin \psi & \sin \phi \sin \theta \sin \psi + \cos \phi \cos \psi & \cos \phi \sin \theta \sin \psi - \sin \phi \cos \psi \\ -\sin \theta & \sin \phi \cos \theta & \cos \phi \cos \theta \end{bmatrix} \quad (3)$$

We can assume that a certain height of the quadrotor, certain ground effects will affect Qball-X4 and we define $g_r(z)$ for such an effect as follows:

$$g_r(z) = \begin{cases} \frac{A}{(z + z_{cg})^2} - \frac{A}{(z_0 + z_{cg})^2} & 0 < z \leq z_0 \\ 0 & \text{else} \end{cases} \quad (4)$$

In this equation we consider A as ground effects and z_{cg} is the Z component of CoG. Because it is very difficult to derive the exact equations for the ground effects, the term $g_r(z)$ is considered as an unknown perturbation in control design, which requires compensation or adaptation. We can simplify (1) and (2), by defining input terms as in (5). u_1 represents the normalized total lift force, and u_2 , u_3 and u_4 correspond to the control inputs of roll, pitch and yaw moments, respectively.

$$\begin{aligned} u_1 &= (F_1 + F_2 + F_3 + F_4) / m \\ u_2 &= (F_3 - F_4) / J_1 \\ u_3 &= (F_1 - F_2) / J_2 \\ u_4 &= \rho(-F_1 - F_2 + F_3 + F_4) \end{aligned} \quad (5)$$

Then the equation of motion can be re-written as below:

$$\dot{x} = u_1 (\cos \phi \sin \theta \cos \psi + \sin \phi \sin \psi) \quad (6)$$

$$\dot{y} = u_1 (\cos \phi \sin \theta \sin \psi - \sin \phi \cos \psi) \quad (7)$$

$$\dot{z} = u_1 (\cos \phi \cos \theta) - g + g_r(z) \quad (8)$$

$$\dot{\phi} = u_2 l \quad (9)$$

$$\dot{\theta} = u_3 l \quad (10)$$

$$\dot{\psi} = u_4 \quad (11)$$

By defining state and input vectors as $\mathbf{x} = [x, y, z, \phi, \theta, \psi]$ and $\mathbf{u} = [u_1, u_2, u_3, u_4]$, the matrix-vector form of the above equations of motion can be represented as:

$$\ddot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})\mathbf{u} + \mathbf{f}_r(\mathbf{x}), \quad (12)$$

where

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} 0 \\ 0 \\ -g \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{f}_r(\mathbf{x}) = \begin{bmatrix} 0 \\ 0 \\ g_r(z) \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad (13)$$

and $\mathbf{g}(\mathbf{x})$ is defined as follows:

$$\mathbf{g}(\mathbf{x}) = \begin{bmatrix} \cos \phi \sin \theta \cos \psi + \sin \phi \sin \psi & 0 & 0 & 0 \\ \cos \phi \sin \theta \sin \psi - \sin \phi \cos \psi & 0 & 0 & 0 \\ \cos \phi \cos \theta & 0 & 0 & 0 \\ 0 & l & 0 & 0 \\ 0 & 0 & l & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (14)$$

3.2 Linearized State-Space Model of the Qball-X4

This section describes the linearized dynamic model of the Qball-X4 for the purpose of designing linear controller, such as MRAC. For the following discussion, the axes of the Qball-X4 are denoted as (x, y, z) and defined with respect to the configuration of the Qball-X4 as shown in Figure 2. Roll, pitch, and yaw are defined as the angles of rotation about the x , y , and z axis, respectively. The global workspace axes are denoted as (X, Y, Z) and defined with the same orientation as the Qball-X4 sitting upright on the ground.

Actuator Dynamics

To count into dynamics of the actuators in Qball-X4 modeling, the thrust generated by each propeller is modeled using the following first-order system:

$$F = k \frac{\omega}{s + \omega} u \quad (15)$$

where u is the PWM input to the DC-motor actuator, ω is the actuator bandwidth and K is a positive gain. These parameters were calculated and verified through experimental studies. A state variable, v , will be used to represent the actuator dynamics, which is defined as follows:

$$v = \frac{\omega}{s + \omega} u \quad (16)$$

Roll and Pitch Models

Assuming that rotations about the x and y axes are decoupled, the motion in roll/pitch axis can be modeled as shown in Figure 4. As illustrated in the figure, two propellers contribute to the motion in each axis. The thrust generated by each motor can be calculated from Eq. (15) and used as corresponding input. The rotation around the center of gravity is produced by the difference in the generated thrusts. The roll/pitch angle can be formulated using the following dynamics:

$$J \ddot{\theta} = \Delta FL \quad (17)$$

where

$$J = J_{roll} = J_{pitch} \quad (18)$$

are the rotational inertia of the device in roll and pitch axes. L is the distance between the propellers and the center of gravity, and

$$\Delta F = F_1 - F_2 \quad (19)$$

represents the difference between the forces generated by the propeller pair (1, 2).

By combining the dynamics of motion for the roll/pitch axis and the actuator dynamics for each propeller the following state-space equations can be derived:

$$\begin{bmatrix} \dot{\theta} \\ \ddot{\theta} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & \frac{KL}{J} \\ 0 & 0 & -\omega \end{bmatrix} \begin{bmatrix} \theta \\ \dot{\theta} \\ v \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Delta F \quad (20)$$

To facilitate the use of an integrator in the feedback structure a fourth state can be added to the state vector, which is defined as $\dot{s} = \theta$.

Height Model

The motion of the Qball-X4 in the vertical direction (along with the Z axis) is affected by all the four propellers. The dynamic model of the Qball-X4 in this case can be written as:

$$M \ddot{Z} = 4F \cos(r) \cos(p) - Mg \quad (21)$$

where F is the thrust generated by each propeller, M is the total mass of the propeller, Z is the height and r and p represent the roll and pitch angular rates, respectively. As expressed in this equation, if the roll and pitch angular rates are nonzero the overall thrust vector will not be perpendicular to the ground. Assuming that roll and pitch angles are close to zero, the dynamic equations can be linearized to the following state space form:

$$\begin{bmatrix} \dot{z} \\ \dot{z}_y \\ \dot{v} \\ \dot{s} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{4K}{M} & 0 \\ 0 & 0 & -\omega & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} z \\ z_y \\ v \\ s \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \omega \\ 0 \end{bmatrix} u + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} -g \quad (22)$$

X-Y Position Model

The motions of the Qball-X4 along the X and Y axes are caused by the total thrust and by changes of the roll/pitch angles. Assuming that the yaw angle is zero, the dynamics of motion in X and Y axes can be written as:

$$\begin{aligned} M\ddot{X} &= 4F \sin(p) \\ M\ddot{Y} &= -4F \sin(r) \end{aligned} \quad (23)$$

Assuming that the roll and pitch angle rates are close to zero, the following linear state-space equations can be derived for X and Y positions.

$$\begin{bmatrix} \dot{X} \\ \ddot{X} \\ \dot{v} \\ \dot{s} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{4K}{M} & p \\ 0 & 0 & -\omega & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} X \\ \dot{X} \\ v \\ s \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \omega \\ 0 \end{bmatrix} u \quad (24)$$

$$\begin{bmatrix} \dot{Y} \\ \ddot{Y} \\ \dot{v} \\ \dot{s} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{4K}{M} & r \\ 0 & 0 & -\omega & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} Y \\ \dot{Y} \\ v \\ s \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \omega \\ 0 \end{bmatrix} u \quad (25)$$

Yaw Model

The torque generated by each motor, τ , is assumed to have the following relationship with respect to the PWM input, u

$$\tau = K_y u \quad (26)$$

where K_y is a positive gain. The motion in the yaw axis is caused by the difference between the torques exerted by the two clockwise and the two counterclockwise rotating props. The motion in the yaw axis can be modeled by:

$$J_y \ddot{\theta}_y = \Delta \tau \quad (27)$$

where θ_y is the yaw angle and J_y is the rotational inertia about the z axis. The resultant torque of the motors, $\Delta \tau$, can be calculated by:

$$\Delta \tau_y = -\tau_1 - \tau_2 + \tau_3 + \tau_4 \quad (28)$$

The yaw axis dynamics can be rewritten in the state-space form as:

$$\begin{bmatrix} \dot{\theta}_y \\ \ddot{\theta}_y \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \theta_y \\ \dot{\theta}_y \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{K_y}{J_y} \end{bmatrix} \Delta \tau_y \quad (29)$$

4. GAIN-SCHEDULED PROPORTIONAL-DERIVATIVE-INTEGRAL (GS-PID) CONTROLLER

In view of the advantages of widely used Proportional-Integral-Derivative (PID) controller and gain scheduling control strategy in aerospace and industrial applications, a control strategy by using gain scheduling based PID controller is proposed for fault tolerant control (FTC) of our UAV test-bed Qball-X4.

As described previously, PID controllers are designed and tuned in both fault-free and faulty situations to control the Qball-X4 under normal and faulty flight conditions.

For GS-PID controller, several sets of pre-tuned gains are applied to the controllers in different flight conditions under both fault-free and faulty cases. In the next step, attempts to obtain the best stability and performance of Qball-X4 in trajectory tracking control under both cases and to switch the controller gains from one set of pre-tuned PID controller to another set of the gains in the presence of different levels of actuator faults are carried out.

One of the main parameters to consider in GS-PID is the switching time between the time of fault occurrence and the time of switching to new set of gains. In other words, if this transient (switching) time is held long (more than one second) it can cause the Qball-X4 to hit the ground and cause a crash, since the operating height was considered as 70 cm to 1 meter. The structure of a GS-PID controller implemented in the Qball-X4 software environment is shown in Figure 5.

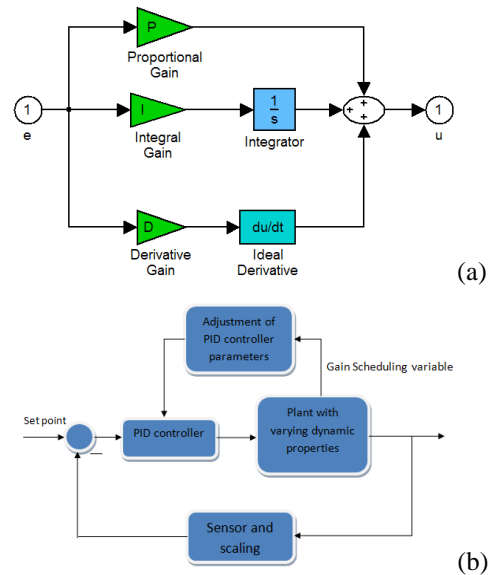


Figure 5. (a) PID and (b) GS-PID controller structures

5. MODEL REFERENCE ADAPTIVE FAULT/DAMAGE TOLERANT CONTROLLER

Model Reference Adaptive Control (MRAC) is concerned with forcing the dynamic response of the controlled system to asymptotically approach that of a reference system, despite parametric uncertainties (faults) in the system. Two major subcategories of MRAC are those of *indirect methods*, in which the uncertain plant parameters are estimated and the controller redesigned online based on the estimated parameters, and *direct methods*, in which the tracking error is forced to zero without regard to parameter estimation accuracy (though under certain conditions related to the level of excitation in the command signal, the adaptive laws often can converge to the proper values). MRAC for linear systems has received, and continues to receive, considerable attention in the literature. Based on the advantages of the direct method without the need of estimation of unknown parameters for implementing the adaptive controller as required by the indirect method, direct method is selected in this work for fault-tolerant control of the Qball-X4. The control structure of such a MRAC scheme can be represented as in Fig. 6.

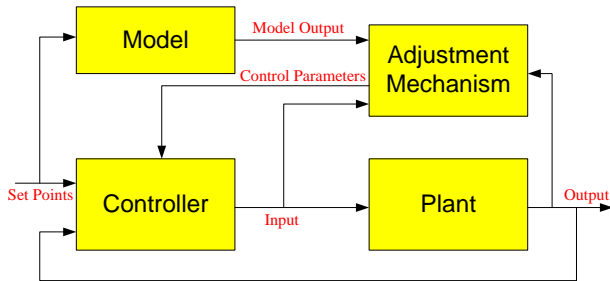


Figure 6. Model reference adaptive control structure

There are different approaches to MRAC design, such as:

- The MIT rule
- Lyapunov stability theory
- Hyperstability and passivity theory
- The error model
- Augmented error
- Model-following MRAC
- Modified-MRAC (M-MRAC)
- Conventional MRAC (C-MRAC)

In this paper, the MIT rule is used to design the MRAC for the height hold and trajectory tracking of the Qball-X4. However, the schemes based on the MIT rule and other approximations may go unstable. We illustrate the use of the MIT rule for the design of an MRAC scheme for the plant

$$\ddot{y} = -a_1\dot{y} - a_2y + u \quad (30)$$

where a_1 and a_2 are the unknown plant parameters, and \dot{y} and y are available for measurement. The reference model to be matched by the closed-loop plant is given by:

$$\ddot{y}_m = -2\dot{y}_m - y_m + r \quad (31)$$

The control law is then given by:

$$u = \theta_1^*\dot{y} + \theta_2^*y + r \quad (32)$$

where

$$\theta_1^* = a_1 - 2, \quad \theta_2^* = a_2 - 1 \quad (33)$$

will achieve perfect model following. The equation (33) is referred to as the matching equation. Because a_1 and a_2 are unknown, the desired values of the controller parameters θ_1^* and θ_2^* cannot be calculated from (33). Therefore, following control law are used instead:

$$u = \theta_1\dot{y} + \theta_2y + r \quad (34)$$

where θ_1 and θ_2 are adjusted using the MIT rule as:

$$\dot{\theta}_1 = -\gamma e_1 \frac{\partial y}{\partial \theta_1}, \quad \dot{\theta}_2 = -\gamma e_2 \frac{\partial y}{\partial \theta_2} \quad (35)$$

where $e_1 = y - y_m$. To implement (35), we need to generate the sensitivity functions $\frac{\partial y}{\partial \theta_1}, \frac{\partial y}{\partial \theta_2}$ online.

6. EXPERIMENTAL FLIGHT TESTING RESULTS

6.1 Flight Testing Results with GS-PID

For comparison purpose and as a baseline controller of the Qball-X4 under normal flight conditions, a single PID controller, which is tuned well for taking-off, hovering and landing scenario under normal flight condition is designed first. Such a controller is used also in a faulty scenario with an 18% of overall loss in power of all motors. Since the significantly deteriorated performance by using a single PID controller, in particular when the fault level increases, another set of PID gains is set for the fault case with gain scheduling strategy for a better handling of the fault comparing with a single PID controller mainly designed and turned for normal flight of the Qball-X4. To analyze the effect of time delay due to fault detection and diagnosis scheme, different levels of time delay were introduced when scheduling/switching the controller gains after a fault occurrence since such fault detection induced time delay is essential for maintaining the stability and the acceptable performance of the Qball-X4 after fault occurrence.

Flight tests with a one meter circuit leg square trajectory tracking scenario for cases with different time delays have been carried out. As shown in Fig. 7, acceptable tracking deviation from the desired square trajectory after the fault occurrence can be obtained with the case of 0.5 sec time delay. Better tracking performance with a shorter time delay can be achieved which verified the importance of fast and correct fault detection and control switching (reconfiguration) after fault occurrence.

To demonstrate the possible best performance without time delay, i.e. the fault occurrence and the switching of controller gains occur at the same time with the perfect fault detection and isolation, the best result can be achieved by the GS-PID is shown in Fig. 8 where the fault occurred and the PID controller is switched at the same time of 20s. Better tracking performance has been achieved compared to the case with 0.5 s time delay as shown in Fig. 7. Videos on the above flight testing results are available at <http://users.encs.concordia.ca/~ymzhang/UAVs.htm>.

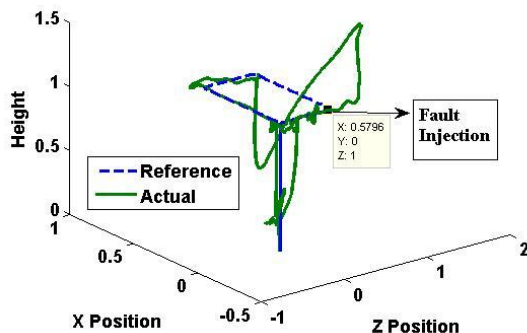


Figure 7. GS-PID with a time delay of 0.5 sec for controller switching in the presence of an actuator fault

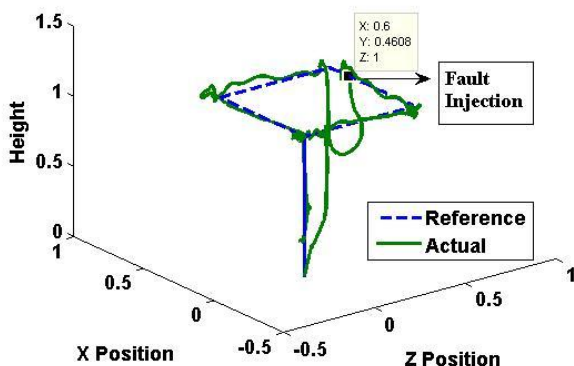


Figure 8. GS-PID without time delay for controller switching in faulty condition (the best performance can be achieved with the designed GS-PID)

6.2 Flight Testing Results with MRAC

Regarding MRAC, hovering control as well as square trajectory tracking controls with fault injection are applied to Qball-X4 and the experimental flight testing results are shown in Figs. 9 and 10. In Fig. 9, a fault-free condition is applied to the Qball-X4 and the MRAC was able to track the trajectory close to real one. In Fig. 10, a fault is injected to the left and back motors at 20 sec with a loss of 18% of power during the flight. As can be seen from Fig. 10, Qball-X4 can still track the desired trajectory with a safe landing. Relevant flight testing videos are also available at <http://users.encs.concordia.ca/~ymzhang/UAVs.htm>.

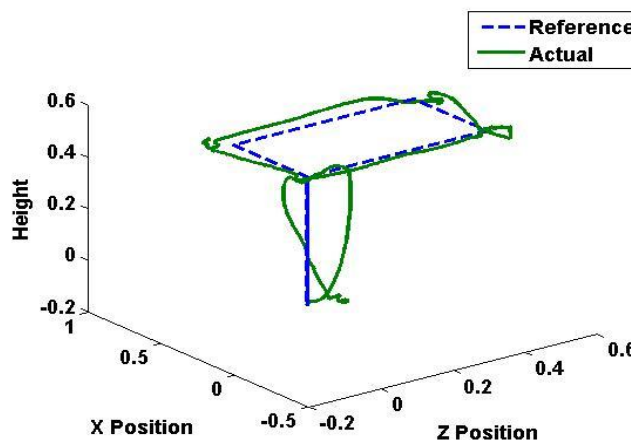


Figure 9. Square trajectory in fault-free condition with MRAC

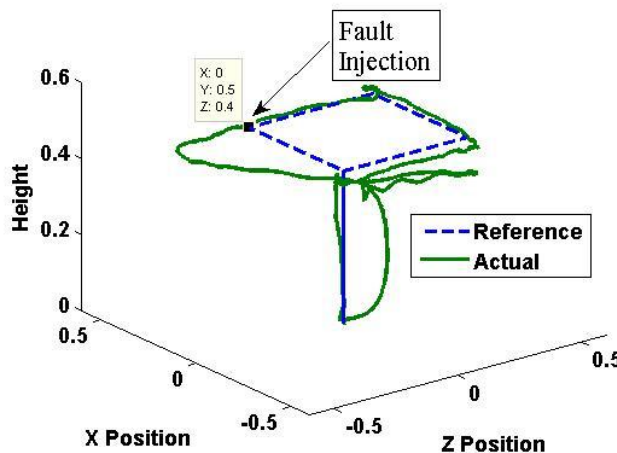


Figure 10. Square trajectory in faulty condition (left and back motors) with MRAC

6.3 Comparison and Comments Based on This Research

During this research many hours of flight tests have been spent at the Network Autonomous Vehicle Laboratory

(NAVL) of the Mechanical and Industrial Department at Concordia University in order to develop the GS-PID and MRAC for achieving the best fault-tolerant control performance of the Qball-X4 under fault flight conditions. By our experience and comparison of the flight testing results, it can be concluded that the MRAC yields a better response than GS-PID for trajectory tracking control although the GS-PID is easier to design and to implement in MATLAB/Simulink interface of the Qball-X4 as well as in the simulation environment. In fact, the GS-PID can give a better result if the tuning for controller gains at pre-fault and faulty cases be very precise. A good tuning for the GS-PID controller gains was very time consuming and gains could change from one flight to another even in our in-door lab environment. Any change in lab environment during flight could force the gains need to be tuned again. However, the MRAC is more reliable and robust to the lab noises and environment changes. Together with the advantages without the need of mathematical model in GS-PID design and implementation compared with MRAC (where a mathematical model is needed to design and implement the controller), GS-PID control technique can play an important role for fault-tolerant control of UAVs as the same as its wide and successful applications in normal/fault-free cases, with the support of an effective and efficient automatic control gains tuning techniques.

7. CONCLUSION AND FUTURE WORK

In this paper, two types of popular controllers, Proportional-Integral-Derivative (PID) controller with Gain Scheduling (GS) technique and Model Reference Adaptive Control (MRAC), are applied and tested, in a quadrotor helicopter UAV test-bed and the results are presented. Both controllers showed good results for height control of the quadrotor UAV: Qball-X4. Unlike the GS-PID, the single PID which is tuned for normal flight was not able to handle the faults with larger fault level.

The future work is considered to combine the GS-PID fault-tolerant control with an online fault detection and diagnosis scheme to achieve an entire active fault-tolerant GS-PID control of the Qball-X4 and other UAVs. Investigation and implementation of efficient auto-tuning strategies for GS-PID is also an important future work although these GS-PID controller gains do not need to be designed on-line in real-time.

REFERENCES

- A. Bani Milhim, Y. M. Zhang, and C.-A. Rabbath, "Gain Scheduling Based PID Controller for Fault Tolerant Control of a Quad-Rotor UAV," *AIAA Infotech@Aerospace 2010*, 20-22 April 2010, Atlanta, Georgia, USA.
- B. T. Whitehead, S. R. Bieniawskiy, "Model Reference Adaptive Control of a Quadrotor UAV," *AIAA Guidance, Navigation, and Control Conference*, 2-5 August 2010, Toronto, Ontario, Canada.
- D. Jourdan et al, "Enhancing UAV Survivability Through Damage Tolerant Control," *AIAA Guidance, Navigation, and Control Conference*, Toronto, Ontario, Canada, 2-5 Aug. 2010.
- E. N. Johnson, G. V. Chowdhary, and M. S. Kimbrell, "Guidance and Control of an Airplane under Severe Structural Damage," *AIAA Infotech@Aerospace 2010*, 20-22 April 2010, Atlanta, Georgia, USA.
- I. Gregory, R. Gadiant, and E. Lavretsky, "Flight Test of Composite Model Reference Adaptive Control (CMRAC) Augmentation Using NASA AirSTAR Infrastructure," *AIAA Guidance, Navigation, and Control Conference*, 8-11 August 2011, Portland, Oregon, USA.
- J. Guo and G. Tao, "A Multivariable MRAC Scheme Applied to the NASA GTM with Damage," *AIAA Guidance, Navigation, and Control Conference*, 2-5 August 2010, Toronto, Ontario, Canada.
- J. Levin, "Alternative Model Reference Adaptive Control," *AIAA Guidance, Navigation, and Control Conference*, 2-5 August 2010, Toronto, Ontario, Canada.
- K. A. Lemon, J. E. Steck, and B. T. Hinson, "Model Reference Adaptive Flight Control Adapted for General Aviation: Controller Gain Simulation and Preliminary Flight Testing on a Bonanza Fly-By-Wire Testbed," *AIAA Guidance, Navigation, and Control Conference*, 2-5 August 2010, Toronto, Ontario, Canada.
- L. G. Crespo, M. Matsutani and A. M. Annaswamy, "Design of a Model Reference Adaptive Controller for an Unmanned Air Vehicle," *AIAA Guidance, Navigation, and Control Conference*, 2-5 August 2010, Toronto, Ontario, Canada.
- Quanser Inc., Qball User Manual, available at http://www.quanser.com/english/html/UVS_Lab/fs_Qball_X4.htm
- R. Gadiant, J. Levin, and E. Lavretsky, "Comparison of Model Reference Adaptive Controller Designs Applied to the NASA Generic Transport Model," *AIAA Guidance, Navigation, and Control Conference*, 2-5 August 2010, Toronto, Ontario, Canada.
- T. Bierling, L. Hocht and F. Holzapfel, "Comparative Analysis of MRAC Architectures in a Unified Framework," *AIAA Guidance, Navigation, and Control Conference*, 2-5 August 2010, Toronto, Ontario, Canada.
- T. L. Jordan, J. V. Foster, R. M. Bailey, and C. M. Belcastro, "AirSTAR: A UAV Platform for Flight Dynamics and Control System Testing," *AIAA Aerodynamic Measurement Technology and Ground Testing Conference*, San Francisco, CA, 2006.

- V. Stepanyan, S. Campbell, and K. Krishnakumar, "Adaptive Control of a Damaged Transport Aircraft Using M-MRAC," *AIAA Guidance, Navigation, and Control Conference*, 2-5 August 2010, Toronto, Ontario, Canada.
- Y. M. Zhang and J. Jiang, "Bibliographical Review on Reconfigurable Fault-tolerant Control Systems," *Annual Reviews in Control*, 32(2), 2008, pp. 229-252.
- Z. T. Dydek and A. M. Annaswamy, "Combined/Composite Adaptive Control of Quadrotor UAV in the Presence of Actuator Uncertainty," *AIAA Guidance, Navigation, and Control Conference*, 2-5 August 2010, Toronto, Ontario, Canada.

Feature Selection and Categorization to Design Reliable Fault Detection Systems

H. Senoussi¹, B. Chebel-Morello², M. Denai³ and N. Zerhouni⁴

¹*University of Sciences and Technology, Mohamed Boudiaf, Oran, Algeria*
senoussih@yahoo.fr

^{2,4}*Automatic Control and Micro-Mechatronic Systems Department of the FEMTO-ST Institute, Besancon, France*
Brigitte.Morello@ens2m.fr
N. Zerhouni @ens2m.fr

³*Teesside University, Middlesbrough, England*
M.Denai@tees.ac.uk

ABSTRACT

In this work, we will develop a fault detection system which is identified as a classification task. The classes are the nominal or malfunctioning state. To develop a decision system it is important to select among the data collected by the supervision system, only those carrying relevant information related to the decision task. There are two objectives presented in this paper, the first one is to use data mining techniques to improve fault detection tasks. For this purpose, feature selection algorithms are applied before a classifier to select which measures are needed for a fault detection system. The second objective is to use STRASS (STRong Relevant Algorithm of Subset Selection), which gives a useful feature categorization: strong relevant features, weak relevant and/or redundant ones. This feature categorization permits to design reliable fault detection system. The algorithm is tested on real benchmarks in medical diagnosis and fault detection. Our results indicate that a small number of measures can accomplish and perform the classification task and shown our algorithm ability to detect the correlated features. Furthermore, the proposed feature selection and categorization permits to design reliable and efficient fault detection system.

1. INTRODUCTION

We work in conditional maintenance when the supervision system surveys the fault appearance. In a real supervision system, digital data collection devices and data storage technology allow organizations to store up huge data. The large amounts of data, has created a massive request for new

tools to transform data into task oriented knowledge (The knowledge data discovery, and data mining area). Our work concentrates on real-world problems and fault detection system, where the learner has to handle problems dealing with datasets containing large amounts of irrelevant information [9],[13],[14]. Initial features are often selected subjectively based on human experience. However, when large amount of data are being monitored, expert judgement may be subject to errors and biases. It is therefore desirable to use fully automated feature selection algorithm to overcome these shortcomings.

Over-instrumentation: monitoring too many metrics of a system poses significant problems, as a large number of threshold estimation, quantification, aggregation, situation identification and diagnostic rules exclude reliable manual design and maintenance, especially in evolving applications. On the other hand monitoring too many metrics also causes unnecessary performance overhead on the monitored systems, and data collection nodes especially in case of historic data collection.

Under-instrumentation: the improper reduction of the set of monitored metrics, on the other hand can significantly compromise the capabilities of supervision, manifesting in large reaction times to workload changes, significantly reduced availability due to late error detection and diagnosis. The selection of a compact, but sufficiently characteristic set of control variables is one of the core problems both for design and run-time complexity [43]. Dimension reduction methods are usually divided into two groups: feature extraction and feature selection approaches. Feature extraction aims at applying a projection of the multidimensional problem space into a space of fewer dimensions thus resulting in aggregate measures that did not exist in the measured environment while feature selection is

Senoussi et. al., This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

finding a subset of the measured variables or a subset of the transformed variables via feature extraction.

Many descriptive features may affect the precision of a classifier and some can even parasitize the processing of data. However, it should be noted that features do not have the same importance in the development of the classifier. Therefore it is very useful to be able to identify, within the whole training set, the appropriate features' types to discriminate between the fault detection concepts being considered. Yu et al [25] counted four (4) different features types namely irrelevant ones, strongly relevant, weakly relevant and redundant ones. An entire feature set can be conceptually divided into four (4) basic disjoint parts: irrelevant features (I), redundant features (part of weakly relevant features (WRr1 and WRr2)), weakly relevant but non-redundant features (WRnr), and strongly relevant features (predominant). Fig. 2 illustrates this hierarchy. The optimal subset essentially contains all predominant features, WRnr and WRr1 or WRr2. WRr1 is a subset of weakly relevant features having their redundant or equivalent features in WRr2.

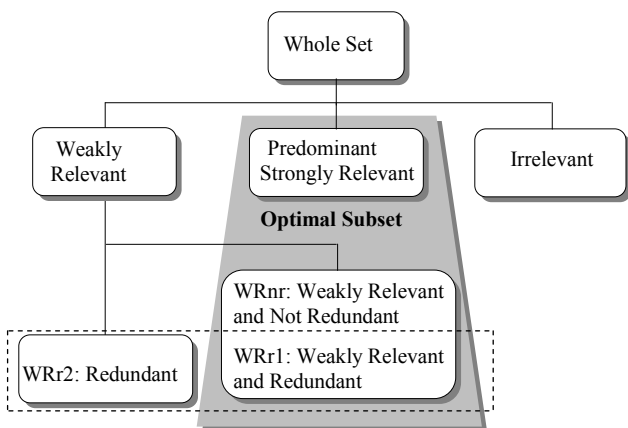


Figure 1. Hierarchy of feature's relevance and redundancy

First of all, we have to reduce the number of sensors/metrics considered in order to avoid over instrumentation and to simplify the classification problem. The filter algorithm STRASS (STRong Relevant Algorithm of Subset Selection) [22] is initially used to select relevant information, construct a robust fault detection model and speed up training time. Moreover the proposed feature selection algorithm is based on two criteria of relevance which provide a useful features' categorization (Fig. 1): the strongly, weakly relevant features and the redundant ones. This features' categorisation is based on criteria developed in [22], [35]. In our precedent study [22], we define two complementary criteria, one Myopic and the other Contextual, to take into account partially redundant feature and privilege the quality detection of relevant subset feature. The proposed criteria attempt to explicitly address feature interactions by finding

some low-order interactions 2-way (one feature and the class) and high order interactions k-way (k features and the class) interactions. Those criteria are associated with a greedy algorithm which is noted STRASS. STRASS proves its efficiency and effectiveness comparing with five representative algorithms on artificial benchmarks well known for their features interactions. The other paper's contribution is in the exploitation of redundant features to improve fault detection reliability by reducing false alarm and/or missed alarm. Reliability requires the minimization of undetectability and false alarm probability due to sensor readings, which is not only related with sensor readings but also affected by fault propagation. In engineering practice, sensors may often be faulty, meaning that they may fail to give adequate readings or the sensor may give an alarm for a normal operation state, known as a false alarm. We should therefore allow for some redundancy in sensors in case of failures. A robust strategy to identify faulty sensor readings and to discriminate among sensor failure and system failure has been developed.

The rest of the paper is organized as follows: Section 2 overviews the state of art of feature selection techniques for fault detection systems. The study highlights the importance of the pre-processing phase such as feature extraction and selection to improve the classifier. Section 3 introduces the features' categorization technique, the proposed criteria and STRASS features selection algorithm that take into account the type of features in a rather finer way than other methods. It is worth noting that the authors' contribution is not in the filtering algorithm, but rather in the features categorization that has been derived from it to build a reliable fault detection system. Section 4 is devoted to the proposed methodology using feature categorization to design reliable fault detection systems. In Section 5 the proposed algorithm is evaluated and compared with two well-known feature selection algorithms CFS (Correlation Based Feature Selection) [10] and FCBF (Fast Correlation Based Feature Selection) [25] and a feature extraction algorithm Principal Component Analysis (PCA). CFS and FCBF are considered to be among the best methods for their ability to treat different feature types and consequently provide a finer feature selection based on a minimal subset. Conclusions and recommendation for future work are summarized in Section 6.

2. A SURVEY OF RELATED WORK ON FEATURE SELECTION

Fault detection methods are generally based on either signal processing or physical models. Data-driven techniques for fault detection and diagnosis have also been extensively used. The following is a brief overview of some recently published papers on feature selection techniques for fault detection.

Paljak et al (2009) [31] considered the selection of a compact, but sufficiently characteristic set of control variables which can provide, in a simple way, good parameter estimators for predictive control. Their approach also provides the identification of the operational domain hence facilitating context-aware adaptive control, diagnostic and repair in large Infrastructure Monitoring. They used mRMR (minimum Redundancy Maximum Relevance) feature selection algorithm combined with linear approximation for selecting the few and most significant quantitative aspects of a system for the purpose of supervisory instrumentation. Yang et al (2008) [37] presented a survey on fault diagnosis using Support Vector Machine (SVM) classifiers combined with other methods. For the detection of faults in roller bearing, Jack et al (2002) used Genetic Algorithms to select an optimal feature subset for two SVM and artificial neural network based classifiers. Casimira et al (2006) [6] reviewed various pattern recognition methods for the diagnosis of faults in induction motors' stator and rotor. A set of 31 features were initially extracted by a time frequency analysis of stator currents and voltages and combined with others features. The most relevant features were selected using a sequential backward algorithm. The experimental results demonstrated the effectiveness of the proposed method to improve the k-nearest neighbours classification rate in condition monitoring. The work by Sugumara et al (2007) [38] focussed particularly on fault conditions in the roller bearing of a rotary machine. They used vibration signals from a piezoelectric transducer in different functional mode (good bearing, bearing with inner race fault, bearing with outer race fault, and inner and outer race fault). First, a set of 11 features were extracted by time frequency analysis. Among these, the 4 best features were selected from a given set of samples using the popular C4.5 decision tree algorithm. Second, Proximal Support Vector Machine (PSVM), was used to efficiently classify the faults using statistical features. Torkolan et al (2004) [23] constructed a driver's assistance system. This system uses feature selection to identify which sensors are needed for the classification of 12 manoeuvres (changing left, crossing shoulder, on road...). Sensor data like accelerator, brake, speed, etc. were collected from a driving simulator and a total of 138 features were extracted from this data set. The authors used Naïve Bayes and Random Forest classifiers. They combined CFS feature selection algorithm and Random Forest with various measures to calculate new features and evaluate which among the derived features were relevant to this problem in addition to selecting the best sensors. The results indicated that to some extent new sensor hardware can be exchanged with a software version by computing new variables based on existing ones. Feature selection in this case allows controlled collection of data using a desired number and type of sensors.

Among existing feature selection methods applied to fault detection system, earlier methods often evaluate variables without considering feature-feature correlation and interaction. They rank feature according to their individual relevance or discriminative power to the targeted classes and select top-ranked features. These methods are computationally efficient due to linear time complexity in terms of dimensionality. However, (1) they cannot give the feature categorization that we have cited and (2) they cannot remove partially redundant features.

3. FEATURE CATEGORISATION: CONCEPT AND CRITERIA OF RELEVANCE AND REDUNDANCY

3.1. Feature Categorisation

Definition 1: Irrelevant

A feature is useful if it is correlated with or predictive of the class; otherwise it is irrelevant [10].

Definition 2: Weakly relevant

A feature x_i is weakly relevant to a sample N of instances and distribution D if it is possible to remove a subset of the features so that x_i becomes strongly relevant (Blum and Langley [4]).

Definition 3: Strongly relevant

A feature x_k is strongly relevant to sample N if there exist examples A and B in N that differ only in their assignment to x_k and have different labels (or have different distributions of labels if they appear an N multiple of times). Similarly, x_k is strongly relevant to target c and distribution D if there exist examples A and B having non-zero probability over D that differ only in their assignment to x_k and satisfy $c(A) \neq c(B)$ (Blum and Langley definition's [4]).

Definition 4: Redundant

A feature is said to be redundant if several features taken together play the same role as the underlying feature (they discriminate the population studied by the considered feature).

3.2. Criteria of Relevance and Redundancy

Two criteria have been introduced to categorise a whole set of features (Senoussi et al [22]). These criteria were elaborated from the discriminatory power in a pair-wise data representation approach. The categorized features types depend on: predominant (strongly relevant), weakly relevant and redundant ones. These criteria are briefly described below.

3.2.1. Data representation

Giving the input data tabulated as Ω samples. A *signature* is a vector of r features x called pattern vector denoted by $x = \{x_i, i = 1, \dots, r\}$. The functional states are

represented by M classes $C = \{c_i, i = 1, \dots, M\}$ in an r -dimensional space. Making a decision consists in assigning an incoming input vector to the appropriate class. This decision consists in recognizing the functional state of the system. Let's associate to a feature x_k the function φ_{ij}^k relative to each pairs of instances $(\omega_i, \omega_j), i \neq j$.

$$\varphi^k(\omega_i, \omega_j) = \begin{cases} (\omega_i, \omega_j) \mapsto \varphi_{ij}^k = \\ 1 \Leftrightarrow x_k(\omega_i) = x_k(\omega_j) & i, j = 1, \dots, n \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The function φ_{ij}^c relative to each pair of instances and their corresponding labels is obtained in the way.

$$\varphi^c(\omega_i, \omega_j) = \begin{cases} (\omega_i, \omega_j) \mapsto \varphi_{ij}^c = \\ 1 \Leftrightarrow C(\omega_i) = C(\omega_j) & i, j = 1, \dots, n \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

3.2.2. Weak relevance measure

The weakly relevance of a set of feature is defined by the number of all pairs of objects who have at least one discriminating variable and different labels or different distributions of labels.

Proposition 1: *The discriminating capacity measure of a feature set $DC(L, \Omega)$:*

$$\text{On } (\Omega \times \Omega) \mapsto DC(L, \Omega) = \sum_{i=1}^n \sum_{j=1}^n \overline{\prod_{k=1}^m \varphi^k} \cdot \overline{\varphi^c} \quad (3)$$

Given a subset of m features $L = (x_1 \dots x_m)$; the subset of feature group relevance is the number of pairs that are discriminate at least with one feature for each class.

3.2.3. Strong relevance to the sample/distribution

To measure the exclusiveness of a feature, the equivalent of a "relevance gain" is defined as the measure related to a feature compared to a subset of features and is termed the Discriminating Capacity Gain (DCG).

First we define the relevance of a feature x_k compared to a relevant pre-selected features subset $L = (x_1 \dots x_m)$ on pairs of instances (ω_i, ω_j) .

The strong relevance (SR) of feature x_k on the data pair ω_i, ω_j is given by:

$$\text{On } (\omega_i, \omega_j) \mapsto SR(x_k, L, \omega_i, \omega_j) = \overline{\varphi^c} \cdot \overline{\varphi_{ij}^k} \cdot \prod_{l=1}^m \varphi^l \quad (4)$$

Proposition 2: *Discriminating capacity gain:* DCG

The aggregation of the Strong Relevance (SR) expression on the whole pairs will define the DCG as:

$$\text{On } (\Omega \times \Omega) \mapsto \text{DCG}(x_k, L, \Omega) = \sum_{i=1}^n \sum_{j=1}^n \overline{\varphi^c} \cdot \overline{\varphi_{ij}^k} \cdot \prod_{l=1}^m \varphi^l \quad (5)$$

The DCG of a feature x_k for a set of objects compared to a set of L features is equal to the number of object couples discriminated by only x_k and no other features.

3.2.4. Redundant feature

Let S be the current set of features if

$$\text{DC}(S, \Omega) - \text{DC}(S - \{x_l\}, \Omega) = 0 \quad (6)$$

Then x_l is a redundant or irrelevant feature compared to the feature subset S on Ω .

3.3. STRASS Algorithm

The criteria are associated with an algorithm related to the greedy type algorithms and noted STRASS (**Appendix A**). STRASS detects the strongly relevant features, the partially redundant features, selects a minimum feature subset and ranks the features' relevance. The algorithm breaks up into three stages depending on its initialisation:

- (1) Selection of strongly relevant features or predominant features which are impossible to exclude because they are the only ones which allow the discrimination of classes.
- (2) Selection of the remaining features or weakly relevant features which have the largest discriminating capacity and when combined with a subset of features, the resulting overall discriminating power is increased. The features having equivalent discriminating capacity are retained as weakly relevant and redundant and are denoted by WRr1 and WRr2.
- (3) Suppression of redundant features. At this stage, backward elimination is employed to detect the features that become redundant compared to the subset of the selected features when adding a new feature.

STRASS, presented in our previous study [22], has proved to be more efficient when compared to five (5) representative algorithms on artificial benchmarks well known for their features interactions and satisfactory performance for the selection of a minimal set of relevant features and handling the k-way features interaction [11]. Reference list entries should be alphabetized by the last name of the first author of each work.

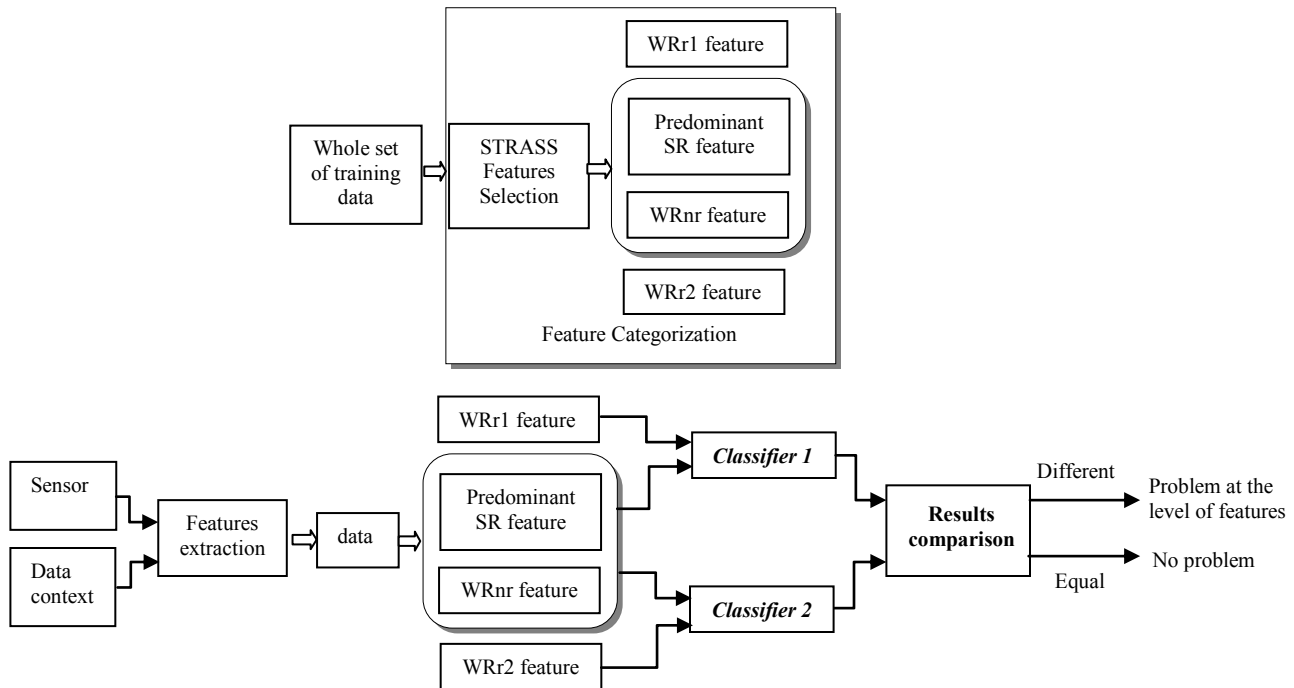


Figure. 2 The proposed fault detection system

4. FEATURES CATEGORIZATION TO CONSTRUCT A RELIABLE FAULT DETECTION SYSTEM

Reliability requires the minimization of undetectability and false alarm probability due to sensor readings or fault propagation. In this study the feature categorization will be used to design reliable fault detection system. Due to their natural discriminating characteristics the selected features make it practical to develop a measure of confidence for the fault detection system show in Fig. 2.

With reference to Fig. 2:

1. Firstly, a fault detection classifier is built using all predominant features (SR), weakly relevant but non-redundant features (WRnr) and weakly relevant features (WRr1).
2. Secondly, redundant features can be used with the predominant and the WRnr ones to build another classifier.
3. In the case of similar results, the second classifier confirms the result obtained with the first one. When the results obtained are different, it is an indication that there is a problem in the acquisition platform (a sensor is defiling) or in the data collection (a parameter is erroneous). The identification of the features is determined by a close examination of the redundant feature, or acquisition of another data.

We should therefore allow for some redundancy in sensors for the predominant measure in case of failures, and the examination of the redundant feature to relay the information. Therefore, missed alarms and false alarms can be detected.

5. EXPERIMENTS AND RESULTS

Our algorithm was implemented in MATLAB 7.5 environment. For the filtering algorithms and classifiers existing tools in WEKA machine learning platform [24] have been used. The experiments were run using WEKA with its default values.

5.1. Feature Selection and Categorization Results

The proposed algorithm has been evaluated on datasets from the UCI Machine Learning Repository [39]. Their characteristics are shown in Table 1.

Datasets	Instances	Features	Classes
Heart	270	14	2
Lung cancer	32	57	2
Hepatitis	20	20	2
Machine	12829	22	22
RFM	3519	72	22

Table 1. Summary of dataset

For the fault detection task Machine and RFM datasets (**Appendix B**) have been used. This data was originally taken at Texas Instruments as part of the SEMATECH J-88 project. For more information about this data set, please see [2][3].

Most existing feature selection algorithms are adapted for discretized (qualitative) data. Therefore for data sets with continuous features, the authors propose to use the MDL (Multi-interval discretization of continuous attributes) discretization algorithm proposed by Fayyad et al [8] also implemented in WEKA environment [24]. Table 2 presents the number of features selected by each features selection algorithm. The proposed algorithm has substantially reduced the number of features necessary to construct a classifier (18% in average in feature space). Table 3 gives STRASS selected features and their categorization. Heart and

Hepatitis have dominant features and redundant ones, thus make it possible to construct a second classifier to detect the same diagnosis and compare the results. For lung cancer and RFM datasets, the selected features are all predominant.

Datasets	ALL	STRASS	CFS	FCBF	ACP
Heart	13	8	6	5	12
Lcancer	56	3	8	6	25
Hepatitis	19	9	9	6	16
Machine	21	5	10	8	17
RFM	71	8	18	11	12
Average	36	6.6	10.2	7.2	16.4

Table 2. Number of features selected by each features selection algorithm

Data sets	STRASS Selected feature	SRp	WRnr	WRr1=WRr2
Heart	8 {3,7,8,1,2,12,9,13}	{3,7,8,1,2,12,13}		4=5=6=9 11=13
L cancer	3 {9,43,34}	{9,43,34}		3=7; 8=9
Hepatitis	9 {11,18,17,6,14,8,12,3,2}	{11,18}	{17,6,14,8,12,2}	3=7=10
Machine	5 {1,3,7,17,13}	{3,17, 13}	{1}	{7=11,12,14,16}
RFM	8 {35,26, 22,14, 44,66,9,4}	{35, 26, 22,14, 44,66,9,4}		

Table 3. STRASS feature categorization

5.2 Detection Results

For the classification task, three different classifiers have been used decision tree (C4.5), K-nearest-neighbor (IB_k), Support Vector Machines (SVM) and multilayer perceptron (MLP). In our experiments, k is set as 1. The classification results are obtained with 10-fold cross-validation. These results are compared with two Correlation-Based Feature Selection algorithms: CFS¹ [10] and FCBF² [25] and the Principal Component Analysis (PCA). TABLES 4-6 show results in both accuracy and kappa obtained with a two tailed test. The symbols “+” and “-” respectively identify significant improvement if an algorithm wins over or loses to the learning algorithm with the whole dataset.

Datasets	C4.5	C4.5+ STRASS	C4.5+ CFS	C4.5+ FCBF	C4.5+ ACP
heart	83.7	85.18+	83.3-	84.4+	81.67-
L cancer	78.12	84.35+	78.21	85.5+	57.92-
hepatitis	81.3	81.3	81.91+	80.6-	79.75-
machine	94.58	94.72+	94.81+	94.70+	93.22-
RFM	94.38	95.34+	94.07-	94.13-	86.79-
Average	86.41	88.17+	86.4	87.86+	79.87-
Win/Loss		4+/0-	2+/-	3+/-	5-/0+

Table 4. C4.5 Classifier precision with and without filtering

¹ CFS with best first search

² FCBF with the relevance threshold SU set to 0.

Datasets	IBk	IBk+ STRASS	IBk+ CFS	IBk+ FCBF	IBk+ ACP
Heart	83.2	82.5 -	82.5-	81.9 -	80.74-
L cancer	75	78.5 +	71.3-	71.8-	65.42-
Hepatitis	83.8	85.8+	77.38-	84.5+	83.96+
Machine	95.80	95.97+	93.3-	94.95-	95.3-
RFM	94.65	96.06+	95.84+	94.67	93.91-
Average	86.49	87.76+	84.06-	85.56-	83.86-
Win/Loss		4+/1-	1+/4-	1+/3-	1+/4-

Table 5. IBk Classifier precision with and without filtering

Datasets	SVM	SVM+ STRASS	SVM+ CFS	SVM+ FCBF	SVM+ ACP
heart	84	84.3+	84.44+	85.18+	84.26+
L cancer	65.62	81.25+	81.25+	87.5+	70.00+
Hepatitis	86.45	87.74+	85.16-	85.80-	83.25-
Machine	88.98	61.12-	73.40-	72.60-	78.12-
RFM	90.12	94.32+	89.78-	88.92-	87.45-
Average	83.03	81.74-	82.8-	84+	80.61-
Win/Loss		4+/1-	2+/3-	2+/3-	2+/3-

Table 6. SVM Classifier precision with and without filtering

Datasets	MLP	STRASS	CFS	FCBF	ACP
heart	80.43	83.12+	82.61+	79.35-	80.93+
L cancer	67.9	86.67+	85.42+	79.58+	59.17-
hepatitis	84.23	85.21+	84.46+	85.24+	82.23-
machine	79.28	59.87-	50.74-	58.90-	64.66-
RFM	90.51	90.5	89.87-	89.61-	89.16-
Average	80.47	81.07+	78.62-	78.53-	75.23-
Win/Loss		3+/1-	3+/2-	2+/3-	1+/4-

Table 7. MLP Classifier precision with and without filtering

From these results it can be concluded that STRASS leads to a better performance than CFS, FCBF and ACP classifiers. The combination of C4.5 and STRASS produced the best results. For both classifiers, the reduction of features by STRASS gives results comparable or even superior when using all features: average accuracy 88.17% (STRASS) vs. 86.41% (Full Set) for C4.5, 87.76% (STRASS) vs. 86.49% (Full Set) for IBk and 81.07% (STRASS) vs. 80.47% (Full Set) for MLP. The application on Machine and RFM process demonstrates that this method is very effective for feature selection and classification.

6. CONCLUSION AND FUTURE WORK

In this paper we proposed to use STRASS, a contextual-based feature selection algorithm for fault detection to categorize measures and to determine the interaction between features. This enabled us to detect the redundancy among measures. STRASS was initially evaluated in datasets related to medical diagnosis. The proposed feature selection algorithm was then applied to two well known fault detection benchmarks. STRASS has demonstrated its efficiency and effectiveness in reducing the dimensionality of datasets while maintaining or improving the performances of learning algorithms. The application of this feature categorization on Machine and RFM datasets has demonstrated that this method is very effective for fault detection.

STRASS is based on two criteria of relevance that permit to obtain a useful feature categorization. In fact the algorithm detects the strongly relevant features, the weakly relevant and their corresponding partially redundant feature and selects a minimum feature subset. Moreover the proposed criterion in this study provides a useful ranking incorporating the context of others features and detects the equivalent measures (partially redundant features). Future work will focus on exploiting this features categorization to construct a reliable fault detection system by adding redundant measures for the predominant ones and use the redundant information from the redundant measures to construct an equivalent classifier to relay the information (in the case of same result for both classifier) or to point out a problem in the acquisition platform or in the data collection (in the case classifiers give different results).

REFERENCES

- Almuallim, H., Dietterich T. G. Learning with Many Irrelevant Features, Proc. of the Ninth National Conference on Artificial Intelligence, pp. 547-552. (1991).
- Barry, M. Wise and B. Neal. PARAFAC2 Part III. Application to Fault Detection and Diagnosis in Semiconductor Etch. Gallagher Eigenvector Research, Inc. Manson, WA USA. (1999).
- Barna, G.G. Procedures for Implementing Sensor-Based Fault Detection and Classification (FDC) for Advanced Process Control (APC), SEMATECH Technical Transfer Document # 97013235A-XFR (1997).
- A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning, *Artificial Intelligence* 97(1-2), 1997. pp. 245-271.
- Casillas, O. Cordón, M.J. del Jesus, F. Herrera. Genetic Feature Selection in a Fuzzy Rule-Based Classification System Learning Process. *Information Sciences* 136:1-4 135-157 (2001)
- Casimira, R., E. Boutleuxa, G. Clercb, A. Yahouib. The use of features selection and nearest neighbors rule for faults diagnostic in induction motors. *Engineering Applications of Artificial Intelligence* 19 169–177 (2006).
- Dash, M., H. Liu, Hiroshi Motoda. Consistency Based Feature Selection. *PAKDD*: 98-109 (2000).
- Fayyad, U. M., K. B. Irani (1993). Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. *IJCAI*: 1022-1029.
- Guyon, I. and A. Elisseeff, An introduction to variable and feature selection, *The Journal of Machine Learning Research*, 3:1157-1182 (2003).
- Hall, M. Correlation-based feature selection of discrete and numeric class machine learning. In *Proceedings of the International Conference on Machine Learning*, pages 359-366, San Francisco, CA. Morgan Kaufmann Publishers. (2000).
- Jakulin, A., Ivan Bratko, Testing the significance of attribute interactions, *Proceedings of the twenty-first international conference on Machine learning*, p.52, July 04-08, Banff, Alberta, Canada. (2004).
- John, G. H., R. Kohavi, and K. Pflieger. Irrelevant Features and the Subset Selection Problem. *Proceedings of the Eleventh International Conference on Machine Learning*. New Brunswick, NJ: Morgan Kaufmann, pp. 121-129. (1994).
- Kohavi, R and G. H. John. Wrappers for feature subset selection. *AIJ issue on relevance*. (1995).
- Kira, K. and L. Rendell. The feature selection problem: Traditional methods and a new algorithm. *Proceedings*

- of the Tenth National Conference on Artificial Intelligence (pp. 129-134). Menlo Park: AAAI Press/The MIT Press. (1992).
15. Kira, K., L. A. RENDELL (1992), A Practical Approach to Feature Selection, in Proc. of the Ninth International Workshop, ML, , pp. 249-255.
 16. Kononenko, I., S.E. Hong. Attribute selection for modelling, *Future Generation Computer Systems*, 13, pp 181 – 195. (1997).
 17. Langley, P. Selection of relevant features in machine learning, Proc of the AAAI, Fall Symposium on relevance, New Orleans pp 399 – 406. (1994)
 18. Lanzi, P.L. Fast Feature Selection With Genetic Algorithms: A Filter Approach. IEEE International Conference on Evolutionary Computation. Indianapolis. Indianapolis 537-540. (1997).
 19. Li, W., D. Li, J. Ni. Diagnosis of tapping process using spindle motor current. *International Journal of Machine Tools & Manufacture* 43 73–79 (2003).
 20. Liu, H. et L. Yu. Toward Integrating Feature Selection Algorithms for Classification and Clustering, *IEEE Trans on Knowledge and Data Engineering*, VOL. 17, NO. 4. (2005).
 21. Pudil, P., J. Navovicova, J. Kittler, Floating search methods in feature selection. *Pattern Recognition Letters* 15, 1119–1125. (1994).
 22. Senoussi, H. and Chebel-Morello. A New Contextual Based Feature Selection. *IEEE International Joint Conference on Neural Networks, IJCNN 2008 and WCCI 2008 (IEEE World Congress on Computational Intelligence)*. Hong Kong June 1-6. (2008).
 23. Torkola, K., S. Venkatesan. and H. Liu. Sensor Selection for Maneuver Classification. *IEEE Intelligent TranspOltation Systems Conference Washington, D.C., USA, October 36.* (2004)
 24. Witten, I. H. and E. Frank, *Data Mining—Practical Machine Learning Tools and Techniques with JAVA Implementations*", Morgan Kaufmann, San Francisco, CA. (2000).
 25. Yu, L. and H. Liu, Efficient Feature Selection via Analysis of Relevance and Redundancy. *Journal of Machine Learning Research* 5 1205–1224. (2004).
 26. Zhao, Z. and H. Liu, Searching for Interacting Features, *IJCAI2007.* (2007).
 27. Zio, E., P. Baraldi, D. Roverso. An extended classifiability index for feature selection in nuclear transients. *Annals of Nuclear Energy.* 32 1632–1649. (2005).
 28. Sylvain Verron, Teodor Tiplica, Abdessamad Kobi. Fault detection and identification with a new feature selection based on mutual information. *Journal of Process Control* 18 (2008) 479–490.
 29. J. Downs, E. Vogel, Plant-wide industrial process control problem, *Computers and Chemical Engineering* 17 (3) 245–255. (1993).
 30. N. Ricker, Decentralized control of the tennessee eastman challenge process, *Journal of Process Control* 6 (4) 205–221. (1996).
 31. Paljak, I. Kocsis, Z. Egel, D. Toth, and A. Pataricza, Sensor Selection for IT Infrastructure Monitoring, in *Third International ICST Conference on Autonomic Computing and Communication Systems*, 2009.
 32. H. Peng, F. Long, Chris Ding, Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp 1226-1238, Aug., 2005.
 33. Tyan C., Wang, P., Bahler, D. An application on intelligent control using neural network and fuzzy logic *Neurocomputing* 12(4): 345-363 (1996).
 34. Widodo A, Yang, B application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors, *Expert system with application*, volume 33(1), pp 241-25. (2007).
 35. Chebel Morello B., Michaut D, Baptiste P. (2001) A knowledge discovery process for a flexible manufacturing system. *Proc. of the 8th IEEE, ETFA'2001*, vol 1, pp.652-659, octobre, Antibes, Juan les Pins.
 36. Riverol, C., Carosi, C.. integration of fault diagnosis based on case based reasoning in brewing *Sens. & Instrumen. Food Qual2:15-20 Springer.*
 37. Yang, B and Widodo, A. support Vector Machine for Machine Fault Diagnosis, *journal of system design and dynamics* vol 2 n°1 pp 12-23 (2008).
 38. Sugumaran, V., Muralidharan, V. Ramachandran K.I. Feature selection using Decision Tree and classification through Proximal Support Vector Machine for fault diagnostics of roller bearing. *Mechanical Systems and Signal Processing* 21 930–942 (2007).
 39. UCI Knowledge Discovery in Databases Archive: www.ics.uci.edu/~mllearn/MLRepository.html.
 40. L. Chiang, M. Kotanchek, A. Kordon, Fault diagnosis based on fisher discriminant analysis and support vector machines, *Computers and Chemical Engineering* 28 (8) (2004) 1389–1401.
 41. T. Jockenhövel, L. T. Biegler, and A. Wächter, Tennessee Eastman Plant-wide Industrial Process Challenge Problem. Complete Model. *Computers and Chemical Engineering.*, 27, 1513-1531, 2003.
 42. Mostafa Noruzi Nashalji, Mahdi Aliyari Shoorehdeli, Mohammad Teshnehlab. Fault Detection of the Tennessee Eastman Process Using Improved PCA and

Neural Classifier. International Journal of Electrical & Computer Sciences IJECS Vol: 9 No: 9. 2009.

43. G. Paljak, I. Kocsis, Z. Egel, D. Toth, and A. Pataricza, Sensor Selection for IT Infrastructure Monitoring, AUTONOMICS 2009, LNICST 23, pp. 130–143, 2010.

Appendix A: Algorithm STRASS

E	The whole set of data pairs $\Omega \times \Omega$.
$L = \{x_1, x_2 \dots x_m\}$	A set of features to be treated
$S = \emptyset$	Selected features
$SR_p = \emptyset$	Strongly relevant predominant features
$DC_{Tot} = DC(L)$	
$DC_{max} = 0$	
$WR_{nr} = \emptyset$	Weakly relevant and not redundant features
$WR_{r1} = \emptyset$	Weakly relevant and redundant features
$WR_{r2} = \emptyset$	Weakly relevant and redundant features

Table A1 STRASS algorithm pseudo-code

```

1. Selection of predominant features
   for each feature  $x_k$  of  $L$  do
     scan the examples space  $E$ 

       if  $DCG(x_k, L-x_k) \neq 0$ 
          $S = S + x_k$ ;  $L = L - x_k$ ;
          $SR_p = S$ ;
          $E = E - \{\text{discriminated pairs}\}$ 

2. Selection of weak relevant features
   while  $DC(S) < DC_{tot}$  do
     for each feature  $x_k$  of  $L$  do
       scan the examples space  $E$ 

         if  $DC(x_k + S) > DC_{max}$ 

            $DC_{max} = DC(\{x_k\} + S)$ 
            $x_{k\_max} = x_k$ ;  $S = S + \{x_{k\_max}\}$ ;
            $L = L - \{x_{k\_max}\}$ 
            $WR_{nr} = WR_{nr} + \{x_{k\_max}\}$ 

         if  $DC(x_k + S) = DC_{max}$ 

            $WR_{r1} = WR_{r1} + \{x_{k\_max}\}$ 
            $WR_{r2} = WR_{r2} + \{x_k\}$  // detection of redundant features
            $E = E - \{\text{discriminated pairs}\}$ 

3. Detection of the partially redundant features
   for each feature  $x_k$  of  $S$  do
     if  $DC(x_k, S - \{x_k\}) = 0$ 
        $S = S - \{x_k\}$ ; // suppression of the redundant features
        $WR_{r2} = WR_{r2} + \{x_k\}$ ; // detection of redundant features

return  $S, SR_p, WR_{nr}, WR_{r1}, WR_{r2}$ 

```

Appendix B: Machine and RFM Datasets

Machine and RFM datasets are elaborated for fault detection and diagnosis in semiconductor etch [2]. The data comes from the metal etch step in semiconductor processing, specifically the Al-stack etch process. Data was collected on the commercially available Lam 9600 plasma etch tool [3]. The metal etcher was equipped with the machine state sensors, built into the processing tool; it collects machine data during wafer processing. The machine data consists of measured and controlled variables sampled at 1 second intervals during the etch. These are engineering variables, such as gas flow rates, chamber pressure and RF power. These variables are listed in Table B1. The RFM sensors measure the voltage, current and phase relationships at the fundamental frequency of 13.56 MHz and the next four harmonics at four locations in the RF control system. The resulting 70 values are sampled every 3 seconds.

Table B1 Machine state variables used for process monitoring.

x1 : Time	x12 : Phase Error
x2 : Step Number	x13 : RF Power
x3 : BCl3 Flow	x14 : RF Impedance
x4 : Cl2 Flow	x15 : TCP Tuner
x5 : RF Bottom Power	x16 : TCP Phase Error
x6 : RFB Reflected Power	x17 : TCP Impedance
x7 : Endpoint A Detector	x18 : TCP Top Power
x8 : Helium Pressure	x19 : TCP Reflected Power
x9 : Chamber Pressure	X20 : TCP Load
x10 : RF Tuner	X21 : Vat Valve
x11 : RF Load	

Hafida Senoussi received the State Engineering of Electronic degree and the Master these from the University of Sciences and Technology of Oran (Algeria), in 1997 and 2001 respectively. She has been a Lecturer and an Associate Researcher at the Automatic Control and Systems laboratory of Electrotechnic Department, University of Sciences and Technology, Mohamed Boudiaf, Oran (Algeria) from 2002 to 2006. During June-July 2004, she has been a visiting scientist at Automatic Control and Micro -Mechatronic Systems Department of the FEMTO-ST Institute, Besançon (France). At 2006, she obtained a bursary to follow her PHD at the Automatic Control and Micro-Mechatronic Systems Department of the FEMTO-ST Institute, Besançon (France). Since 2009, she is a Lecturer at the University of Sciences and Technology of Oran (Algeria). Her research interests are data selection and decision support system, knowledge data discovery in detection faults system, signal processing.

Brigitte Chebel Morello is an assistant professor at the University of Franche Comté France. She received the Engineer Diploma in Electrical Engineering in 1979 from de University of Science and Technology of Algiers and Ecole Nationale Polytechnique d'Alger (Algeria) and his PhD in Automatic Engineering from the University of Lille (France) in 1983. Since 1983, she worked as Associate professor at the Department of Electrical Engineering and at the department of computer Engineering in Institute of technology of Calais University of Lille (France), and came to the university of Besancon (France) and is a researcher at the Automatic Control and Micro-Mechatronic Systems Department of the FEMTO-ST Institute. His current research interests are experience feedback from maintenance process in the enterprise especially focused in three complementary methods: Knowledge capitalization for industrial equipment diagnosis and repair; Knowledge data discovery in detection faults system; Case based reasoning in a decision support system. She is involved in many research and development of industrial projects.

Mouloud Denai received his Bachelor in Electrical Engineering in 1982 from de University of Science and Technology of Algiers and Ecole Nationale Polytechnique d'Alger (Algeria) and his PhD in Control Engineering from the University of Sheffield (UK) in 1988. He has been with the University of Science and the Technology of Oran (Algeria) from 1988-2004. He worked as Research Associate at the Dept of Automatic Control and Systems Engineering, University of Sheffield (UK) from 2004-2009. Since 2010 he is Senior Lecturer at the Teesside University, UK. His research interests include hybrid (physically-based, data-driven and qualitative) modeling, optimisation and control of life science systems, decision support systems for diagnosis and therapy planning in critical care medicine, data modeling and knowledge elicitation (using neuro-fuzzy, probabilistic reasoning, evolutionary techniques) for medical decision-making. His other research interests include intelligent control design for efficiency optimisation in the field of renewable energies systems, investigation of power electronics interface for renewable energy systems, fault detection and isolation in electric power networks and drives.

Noureddine Zerhouni is a Professor at ENSMM, Besançon, France. He received his MEng from Algiers University, Algeria, and his PhD from Grenoble National Polytechnical Institute (INPG), France. He is also a Researcher at the Automatic Control and Micro-Mechatronic Systems Department of the FEMTO-ST Institute, Besancon (France). His research interests are Petri net applications and AI methods for diagnostic, prognostic and scheduling approaches. He is involved in many research and development projects on prognostic and e-maintenance.

From measurement collection to remaining useful life estimation: defining a diagnostic-prognostic frame for optimal maintenance scheduling of choke valves undergoing erosion

Giulio Gola^{1,2}, Bent Helge Nystad¹

¹*Institute for Energy Technology, OECD Halden Reactor Project, Norway*

²*IO-center for Integrated Operations, Trondheim, Norway*
giulio.gola@hrp.no

ABSTRACT

Condition Based Maintenance (CBM) aims at regulating maintenance scheduling based on data analyses and system condition monitoring. Clear advantages of optimizing maintenance scheduling include relevant cost savings and improved safety and plant availability. A critical aspect is the integration of CBM strategies with condition monitoring technologies for handling a wide range of information sources and eventually making optimal decisions on when and what to repair. In this work, a practical case study concerning maintenance of choke valves in offshore oil platforms has been investigated. Choke valves used in offshore oil platforms undergo erosion caused by the sand grains transported by the oil-water-gas mixture extracted from the well. Erosion is a critical problem which can affect the correct functioning of the valves, result in revenue losses and cause environmental hazards. In this respect, this work proposes a diagnostic-prognostic scheme for assessing the actual health state of a choke valve and eventually estimating its Remaining Useful Life (RUL). In particular, the focus has been on the identification of those parameters which contribute to the actual erosion of the choke valve, the development of a model-based approach for calculating a reliable indicator of the choke valve health state, the actual estimation of the choke RUL based on that indicator using statistical approaches and, finally, the investigation of methods to reduce the uncertainty of the RUL estimation by adding

highly meaningful knowledge on the erosion state of the choke valve*.

1. INTRODUCTION

In oil and gas industries, choke valves are normally located on top of each well and are used to balance the pressure on several wells into a common manifold to control oil, gas and water flow rates and protect the equipment from unusual pressure fluctuations. Figure 1 sketches a choke valve.

The throttle mechanism consists of two circular disks, each with a pair of circular openings to create variable flow areas. One of the disks is fixed in the valve body, whereas the other is rotated either by manual operation or by actuator, to vary or close the opening. For large pressure drops, the well stream containing gas, liquid and sand particles can reach 400-500 m/s and produce heavy metal loss mainly due to solids, liquid droplets, cavitation and combined mechanisms of erosion-corrosion, resulting in choke lifetimes of less than a year. Erosion management is vital to avoid failures that may result in loss of containment, production being held back, and increased maintenance costs. Moreover, several chokes are located subsea, where the replacement cost is high (Andrews *et al.*, 2005; Bringedal *et al.*, 2010; Haugen *et al.*, 1995; Hovda and Andrews, 2007; Hovda and Lejon, 2010; Jarrel *et al.*, 2004; Ngkleberg, and Sontvedt, 1995; Wallace *et al.*, 2004).

For these reasons, attention has focused on the maintenance of choke valves. Currently, fixed maintenance is the most common way to manage choke replacement. A

* This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

more effective way to handle maintenance is to base it on indications of the actual condition (i.e. health state) of the choke valve and possibly on the estimations of its remaining useful life (RUL) (Gola and Nystad, 2011; Kiddy, 2003; Nystad *et al.*, 2010; van Noortwijk and Pandey, 2003).

In general, condition-based maintenance (CBM) approaches rely on data analysis and condition monitoring systems. In fact, the measurements of those parameters considered relevant to assess the health state of a component are first processed by condition monitoring systems which return the diagnostic indication of the current health state. This indication can be then used within prognostic systems to eventually estimate the remaining useful life of the component (Fig. 2).

The integration of condition monitoring systems with CBM strategies is critical for handling a wide range of information sources and providing a reliable indication upon which optimal decisions can be made on when and what to repair.

In this work, the diagnostic-prognostic scheme sketched above is applied to a real case study of choke valve erosion. In this respect, an empirical, model-based condition monitoring system is developed to process the collected measurements in order to give a reliable indication of the erosion state of the choke. A statistical prognostics system based on the gamma process is then used for the estimation of the remaining useful life of the choke.

The work is organized as follows: Section 2 describes the parameters used to assess the choke valve erosion state; Section 3 reports the case study under analysis. Section 4 illustrates the diagnostic-prognostic scheme hereby proposed to assess the choke erosion state and to estimate its remaining useful life. Conclusions are drawn in the last Section.

2. CHOKE VALVE EROSION ASSESSMENT

In the generic choke valve fluid dynamic model, the total flow w through the choke is proportional to the pressure drop Δp through the choke:

$$w = C_V \sqrt{\frac{\Delta p}{\rho}} \quad (1)$$

where ρ is the average mixture density and C_V is called valve flow coefficient. C_V is related to the effective flow cross-section of the valve and is proportional to the choke opening according to a function depending on the type of choke valve and given by the valve constructors, i.e. for a given choke opening, C_V is expected to be constant (Metso Automation, 2005).

When erosion occurs, a gradual increase of the valve area available for flow transit is observed even at constant pressure drop. Such phenomenon is therefore related to an

abnormal increase of the valve flow coefficient with respect to its expected theoretical value, hereby denoted as C_V^{th} .

For this reason, for a given choke opening the difference δ_{C_V} between the actual value of the valve flow coefficient, hereby simply denoted as C_V , and its theoretical value C_V^{th} is retained as an indication of the choke erosion. The difference $\delta_{C_V} = C_V - C_V^{th}$ is expected to monotonically increase throughout the choke life since it should reflect the physical behaviour of the erosion process. When δ_{C_V} eventually reaches a pre-defined failure threshold, the choke must be replaced.

The actual valve flow coefficient C_V cannot be directly measured, but it can be calculated from the following analytical expression which accounts for the physical parameters involved in the process:

$$C_V = \frac{w_o + w_w + w_g}{N_6 F_p \sqrt{p_{in} - p_{out}}} \sqrt{\frac{f_o}{\rho_o} + \frac{f_w}{\rho_w} + \frac{f_g}{\rho_g} J^2} \quad (2)$$

where p_{in} and p_{out} are the pressures upstream and downstream of the choke, w_o , w_w and w_g are the flow rates of oil, water and gas, f_o , f_w and f_g the corresponding fractions with respect to the total flow rate and ρ_o , ρ_w , ρ_g and the corresponding densities, J is the gas expansion factor, F_p is the piping geometry factor and N_6 is a constant equal to 27.3 (Andrews *et al.*, 2005; Gola and Nystad, 2011; Hovda and Andrews, 2007; Metso Automation, 2005; Nystad *et al.*, 2010).

3. CHOKE VALVE EROSION: THE CASE STUDY

A case study on a choke valve located top side on the Norwegian continental shelf is here considered.

Measurements and calculations related to the physical parameters involved in the process are available as daily values. In particular, the pressures upstream and downstream of the choke are directly measured, whereas oil, gas and water flow rates are calculated based on the daily production rates of other wells of the same field. Pressure measurements are considered reliable since they are directly related to the well under analysis, whereas the calculations of oil, gas and water flow rates expected from that well might not be realistic and therefore might not reflect the actual physical composition of the extracted mixture. In addition to the daily measurements and calculations, seven well tests are carried out throughout the valve life at regular intervals, during which oil, gas and water flow rates are accurately measured using a multi-phase fluid separator. The valve choke opening is also provided as a parameter.

Since oil, gas and water flow rates are used to compute the actual C_V (Eq. 2), inaccuracies in their calculation might negatively affect the C_V calculation itself and thus the quality of the erosion indication δ_{C_V} .

Figures 3 and 4 illustrate the parameters used to compute the actual C_V and the resulting erosion indication δ_{C_V} , respectively.

The mismatch between the values of oil, water and gas flow rates daily calculated accounting for the other wells and the values of the same three parameters measured during the well tests is evident in the bottom graphs in Figure 3. Notice that there is instead no mismatch for the pressure drop and, obviously, for the choke opening indication (top graphs in Fig. 3).

As a consequence of the inaccurate daily calculations of oil, water and gas flow rates, the daily erosion indication δ_{C_V} (black line in Fig. 4) results non-monotonic and very noisy, generally showing an unphysical behaviour. On the other hand, when δ_{C_V} is computed using the well test measurements of oil, water and gas flow rates, its behaviour results monotonic and provide a reliable information on the physical erosion process.

Nevertheless, a diagnostic assessment on the erosion state of the valve and a prognostic estimation of its remaining useful life cannot be made based on the daily erosion indications. In the next Section, an empirical model-based approach is used to produce a reliable daily calculation of the erosion state which is then fed to a prognostic system for estimating the choke remaining useful life.

4. IMPROVING THE EROSION STATE CALCULATION FOR ASSESSING THE CHOKE REMAINING USEFUL LIFE

A method developed at the Norwegian Institute for Energy Technology and called Virtual Sensor is here used (PCT/NO2008/00293, 2008). Virtual Sensor is an empirical method based on the use of an ensemble of feed-forward Artificial Neural Networks (ANNs). In general, given a number of input parameters correlated to a quantity of interest, the Virtual Sensor aims at providing a reliable estimate of that quantity.

In general, a subset of the available data (in the format input-parameters/output-target) is used to train the ANN models, i.e. to tune its parameters, with the goal of learning the ANN to estimate the output target. Once the model is trained, it can be used on-line by providing a stream of input measurements in order to obtain an estimate of the (unknown) output.

Virtual Sensor exploits the concepts of ensemble modelling which bear the advantages of ensuring high accuracy and robustness of the estimation without the need of developing one single optimal model. Critical aspects of ensemble modelling are the diversity of the individual

models, hereby ensured by randomizing the training initial conditions of the ANNs, and the aggregation of the outcomes of the individual models, hereby performed by retaining the median of the individual estimates.

In this work, Virtual Sensor is used to provide a reliable estimation of the actual C_V based on the set of available input parameters, namely the pressure drop, the choke opening and the oil, water and gas flow rates. Given the limited amount of available data, the Virtual Sensor has been trained by using as output target a C_V obtained by the linear interpolation of the C_V values calculated with the well test measurements. Figure 5 shows the erosion indication δ_{C_V} obtained with the Virtual Sensor daily estimations of C_V compared with the one obtained using the Equation (2). Despite the erosion indication obtained with the Virtual Sensor is still not completely monotonic, the improvement with respect to the one obtained using Equation (2) is evident.

The erosion indication obtained with the Virtual Sensor conveys a more physically reliable indication of the erosion state of the choke and can be used both within a diagnostic frame to assess the valve performance in the present and within a prognostic system for predicting the temporal evolution of the erosion, eventually estimating when the erosion will cross the failure threshold and the valve needs to be replaced.

To this aim, a statistical approach based on gamma process (van Noortwijk and Pandey, 2003) is here used. Gamma process is a statistical analysis based on Markovian principles and gamma probabilistic distribution.

In a generic prognostic problem, the gamma process exploits the knowledge embedded in the health state indications to calculate the parameters of the temporal evolution of such indication. According to the gamma process, the increments of the health indications are gamma-distributed and can therefore be only positive representing a monotonic quantity. This makes the approach suitable to model the choke valve erosion process which is naturally monotonic.

The expected temporal trend of the health indicator h at time t (i.e. the expected value of the gamma distribution at time t) is $h(t) = \frac{a}{c} t^b$, where b is the parameter which regulates the concavity/convexity of the trend shape and a and c determine the spread of the gamma probability distribution.

Given a failure threshold for the health indicator, the gamma process calculates the conditional probability that the component fails at time $t > T$ given that it has survived up to time T (hereby called time-based approach). The quality of this additional information is critical to define the failure time probability distribution.

In this work, a different approach (hereby called state-based approach) has been adopted which accounts for the knowledge of the actual valve health state. In this view, the gamma process calculates the conditional probability that the component fails at time $t > T$ given the knowledge of its current health state is $h(T) = H$. This approach exploits information of noticeably higher quality, given that a pre-defined list of discrete health states for a component is available based on expert analysis (Gola and Nystad, 2011; Kiddy, 2003).

Another critical issue is the calculation of the parameters of the expected gamma function. In particular, the accurate determination of b is fundamental to obtain meaningful values of the remaining useful life.

Different methods can be used to calculate b . In this work, b is determined by a weighted least-square optimization. Given a time series of health state calculations $b(t)$, $t = 1, \dots, T$, b at current time T is determined by the least-square method using the log-transformed expression for $h(t)$, i.e. $\ln(h(t)) = \ln\left(\frac{a}{c}\right) + b \ln(t)$. Parameter b is therefore the angular coefficient of the straight line which best interpolates the log-transformed health state calculations up to time T given the condition that the interpolation passes by the last available health state calculation, i.e. $\ln(h(T)) = \ln\left(\frac{a}{c}\right) + b \ln(T)$.

The so-called weighted least-square optimization amounts to improving the calculation of b by assigning more importance to the most recent health state calculations which are conjectured to be the most informative. In practice, this is done by artificially adding to the time series of the health state calculations a number K of replicates of the last N health state calculations, i.e. $h(t)$, $t = T - N, \dots, T$. This way of proceeding forces the least-square optimization to better approximate those health state calculations considered most relevant to determine the shape of the gamma function. Once the value of b is set, parameters a and c can be analytically determined using the method of moments (van Noortwijk and Pandey, 2003).

In this case study, measurements corresponding to 305 operational days are available. Approximately 235 operational days of measurements are collected and processed with the Virtual Sensor to produce reliable erosion state indications δ_{C_v} before the gamma process is devised to estimate the choke remaining useful life. This amount of measurements is conjectured to be sufficient to achieve reliable calculations of parameters a , b and c . The weighted least-square optimization is done by considering $K = 1000$ replicates of the last $N = 50$ erosion state indications. This augmented virtual measurement set forces the gamma process to provide the best fit, in terms of

least-square error, for the last 50 collected measurements, which indeed bear the most recent and therefore valuable information on the valve erosion state.

The estimation of the RUL and its uncertainty is then carried out every operational day until the choke is actually replaced. The failure threshold for the erosion indicator δ_{C_v} is set equal to 16. Since the gamma process requires a monotonic data series, the erosion indicator δ_{C_v} is first filtered with a combination of moving average and moving maxima.

Results of the remaining useful life estimation are shown in Figure 6 and compared to those obtained when the b parameter is set constant and equal to 2.2 which is the value that best fits the last 50 available erosion state indications δ_{C_v} in terms of least-square error.

The slowly increasing values calculated for the erosion indicator δ_{C_v} up to 273 operational days (Fig. 5) lead to having values of b with the weighted least-square optimization smaller than 1. As a consequence, the resulting convex shape of the expected gamma function hits the failure threshold at considerably large times, thus returning an overestimated value of the choke remaining useful life.

On the other hand, when values of the erosion indicator δ_{C_v} show a sharp increase towards the end of the choke life, the weighted least-square optimization allows to quickly update the value of b with the effect of obtaining a more precise estimation of the remaining useful life, which, after 290 operational days is comparable to that obtained by fixing b equal to the value which best fits the last 50 measurements.

5. CONCLUSIONS

In this paper, a practical case study concerning erosion in choke valves used in oil industries has been analysed with the aim of defining a diagnostic-prognostic frame for optimizing maintenance scheduling of such components.

Two objectives have been identified: 1) the development of a condition monitoring system capable of providing reliable calculations of the erosion state based on collected measurements of physical parameters related to the choke erosion and 2) the development of a prognostic system to accurately estimate the remaining useful life of the choke.

An empirical, model-based approach has been used to fulfil the diagnostic objective of providing reliable calculations of the erosion state, whereas a statistical method based on the gamma probability distribution has been adopted to reach the prognostic goal of accurately estimating the remaining useful life of the choke.

Although the results obtained so far are encouraging with respect to the goal of defining a diagnostic-prognostic frame for optimizing maintenance scheduling of choke valves, a strong limitation of the proposed procedure has

been envisioned in the amount and the quality of the available data. In fact, it is evident that having data corresponding to one single valve considerably affect the general applicability of the approach which has not been yet demonstrated. With a larger amount of data related to many similar valves one could in fact perform a more consistent training of the Virtual Sensor and eventually define an optimal value for the shape-parameter of the gamma function. In this respect, more measurements are currently collected and further analysis and research is planned.

ACKNOWLEDGEMENTS

The authors wish to thank Erling Lunde and Morten Løes at Statoil ASA for proving us with the operational choke valve data and the IO Center for Integrated Operations in the Petroleum Industry (www.ntnu.no/iocenter) for funding this research project.

REFERENCES

- (Andrews *et al.*, 2005) J. Andrews, H. Kjørholt, and H. Jøranson, Production Enhancement from Sand Management Philosophy: a Case Study from Statfjord and Gullfaks, *SPE European Formation Damage Conference*, Sheveningen, The Netherlands, 2005.
- (Bringedal *et al.*, 2010) B. Bringedal, K. Hovda, P. Ujang, H.M. With, and G. Kjørrefjord, Using Online Dynamic Virtual Flow Metering and Sand Erosion Monitoring for Integrity Management and Production Optimization, *Deep Offshore Technology Conference*, Houston, Texas, US, 2010.
- (Gola and Nystad, 2011) G. Gola, B.H. Nystad. Comparison of Time- and State-Space Non-Stationary Gamma Processes for Estimating the Remaining Useful Life of Choke Valves undergoing Erosion. *Proceedings of COMADEM*, Stavanger, Norway, 2011.
- (Haugen *et al.*, 1995) K. Haugen, O. Kvernfold, A. Ronold, and R. Sandberg. Sand Erosion of Wear Resistant Materials: Erosion in Choke Valves. *Wear 186-187*, pp. 179-188, 1995.
- (Hovda and Andrews, 2007) K. Hovda and J.S. Andrews, Using C_v Models to Detect Choke Erosion - a Case study on Choke Erosion at Statfjord C-39, *SPE Applied Technology Workshop on Sound Control*, Phuket, Thailand, 2007.
- (Hovda and Lejon, 2010) K. Hovda and K. Lejon, Effective Sand Erosion Management in Chokes and Pipelines - Case studies from Statoil, *4th European Sand management Forum*, Aberdeen, Scotland, UK, 2010.
- (Jarrel *et al.*, 2004) D.B Jarrell, D.R Sisk. and L.J. Bond, Prognostics and Condition-Based Maintenance: A New Approach to Precursive Metrics, *Nuclear Technology*, 145, pp. 275-286, 2004.
- (Kiddy, 2003) J.S. Kiddy, Remaining Useful Life Prediction based on Known Usage Data. *Proceedings of SPIE*, 5046(11), 2003.
- (Metso Automation, 2005) Metso Automation. Flow Control Manual. *Metso Automation*, 4th edition, 2005.
- (Ngkleberg, and Sontvedt, 1995) L. Ngkleberg, T. Sontvedt. Erosion in Choke Valves - Oil and Gas Industry Applications. *Wear*, 186-187, Part 2, pp. 401-412 1995.
- (Nystad *et al.*, 2010) B.H. Nystad, G. Gola, J.E. Hulsund, and D. Roverso. Technical Condition Assessment and Remaining Useful Life Estimation of Choke Valves subject to Erosion. *Proceedings of the Annual Conference of the Prognostics and Health Management Society*, Portland, Oregon, US, 2010.
- (PCT/NO2008/00293, 2008) PCT/NO2008/00293, *System and Method for Empirical Ensemble-based Virtual Sensing*.
- (van Noortwijk and Pandey, 2003) J.M. van Noortwijk and M.D. Pandey. A Stochastic Deterioration Process for Time-dependent Reliability Analysis, in *Proceeding of IFIP WG 7.5 Working Conference on Reliability and Optimization of Structural Systems*, Banff, Canada, 2003.
- (Wallace *et al.*, 2004) M.S. Wallace, W.M. Dempster, T. Scanlon, J. Peters, and S. McCulloch. Prediction of Impact Erosion in Valve Geometries. *Wear* 256, pp. 927-936, 2004.

Giulio Gola MSc in Nuclear Engineering, PhD in Nuclear Engineering, Polytechnic of Milan, Italy. He is currently working as a Research Scientist at the Institute for Energy Technology (IFE) and OECD Halden Reactor Project (HRP) within the Computerized Operations and Support Systems department. His research topics deal with the development of artificial intelligence-based methods for on-line, large-scale signal validation, condition monitoring, instrument calibration, system diagnostics and prognostics.

Bent Helge Nystad was awarded an MSc in Cybernetics by RWTH, Aachen in Germany, 1993, and a PhD in Marine Technology by the University of Trondheim (NTNU), Norway, 2008. He has work experience as a condition monitoring expert from Raufoss ASA (a Norwegian missile and ammunition producer) and he has been a Principal Research Scientist at the Institute for Energy Technology (IFE) and OECD Halden Reactor Project (HRP) since 1998. He is the author of 15 publications in international journals and conference proceedings. His experience and research interests have ranged from data-driven algorithms and first principle models for prognostics, performance evaluation of prognostic algorithms, requirement specification for prognostics, technical health assessment and prognostics in control applications.

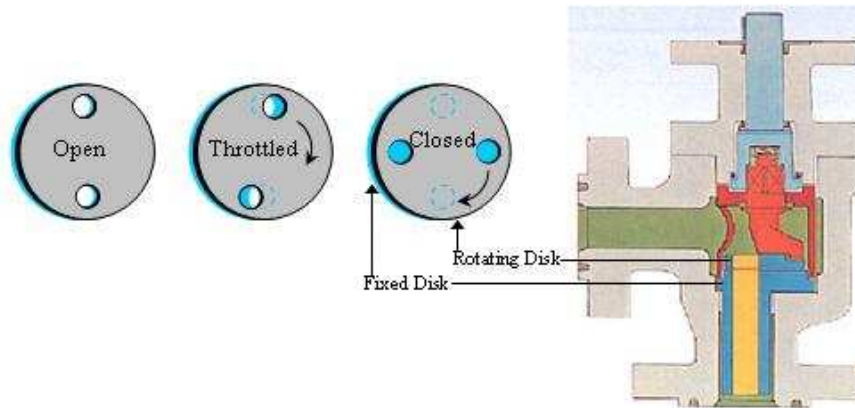


Figure 1: Typical choke valve of rotating disk type: by rotating the disk the flow will be throttled (picture taken from www.vonkchokes.nl)

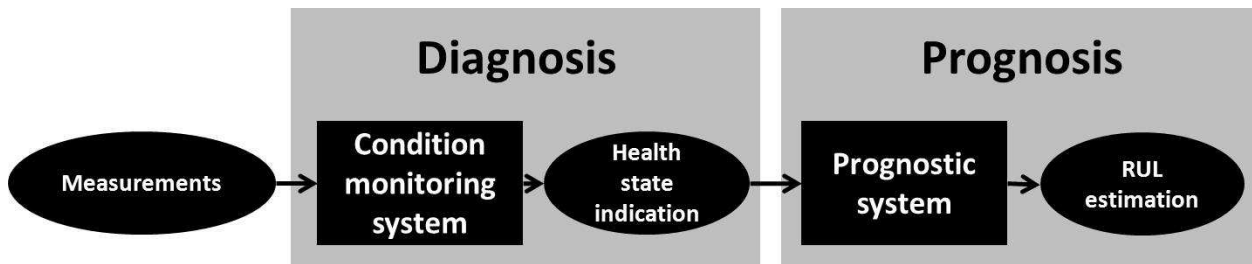


Figure 2. General diagnostic-prognostic frame.

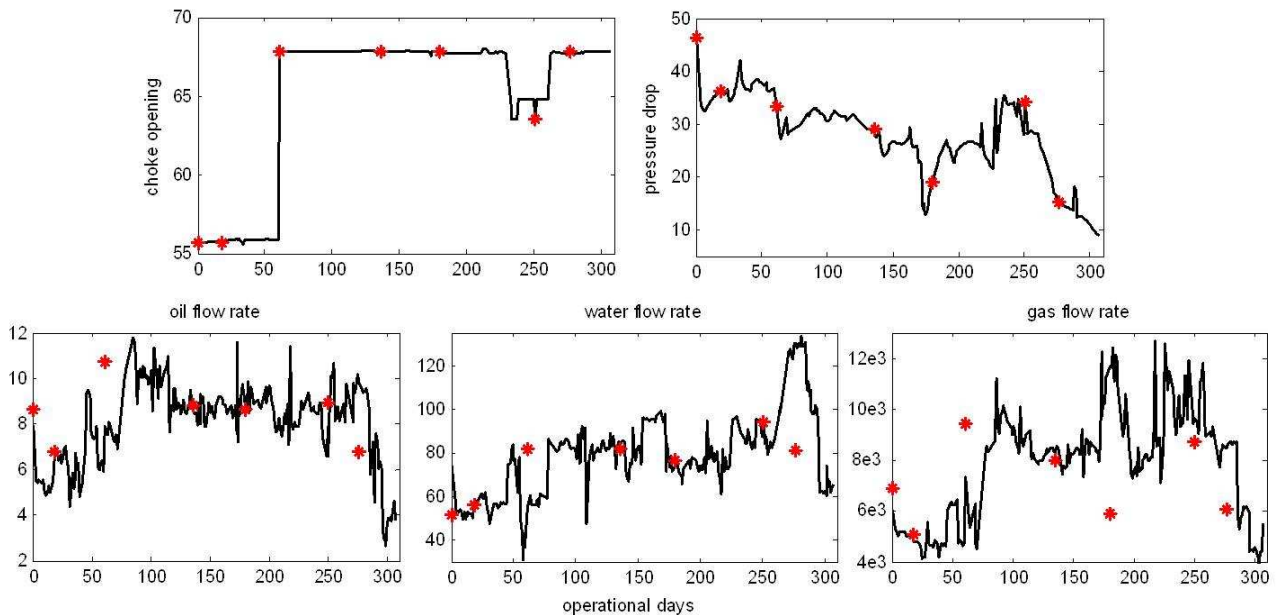


Figure 3. Choke opening and pressure drop (top graphs) and oil, water and gas flow rates (bottom graphs) during daily measurements (black line) and well tests (red stars).

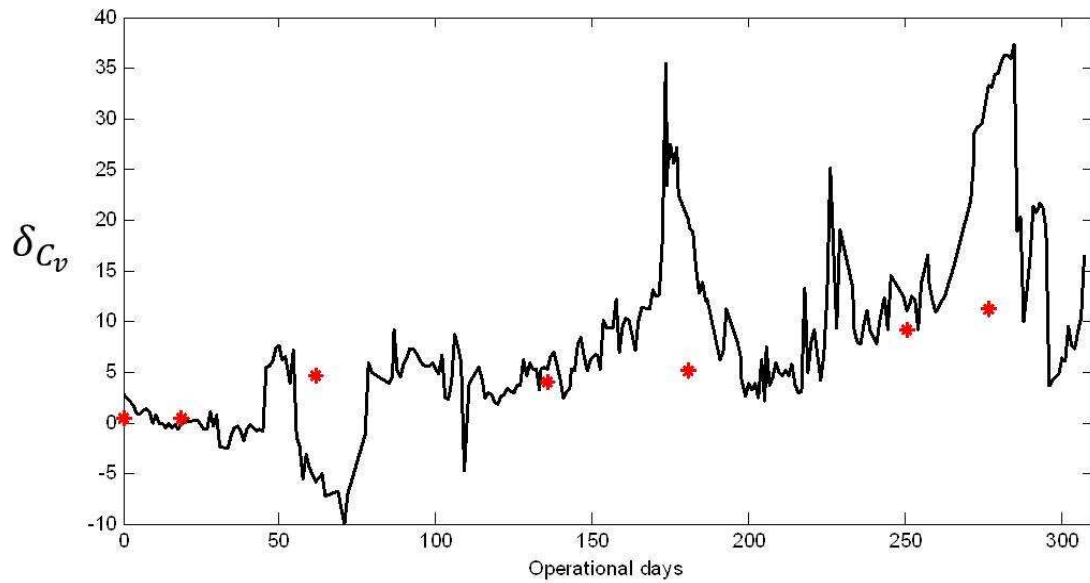


Figure 4. Erosion indication (δ_{C_v}) obtained with C_v calculations based on daily measurements and calculations (black line) and computed using the measurements of the well tests (red stars).

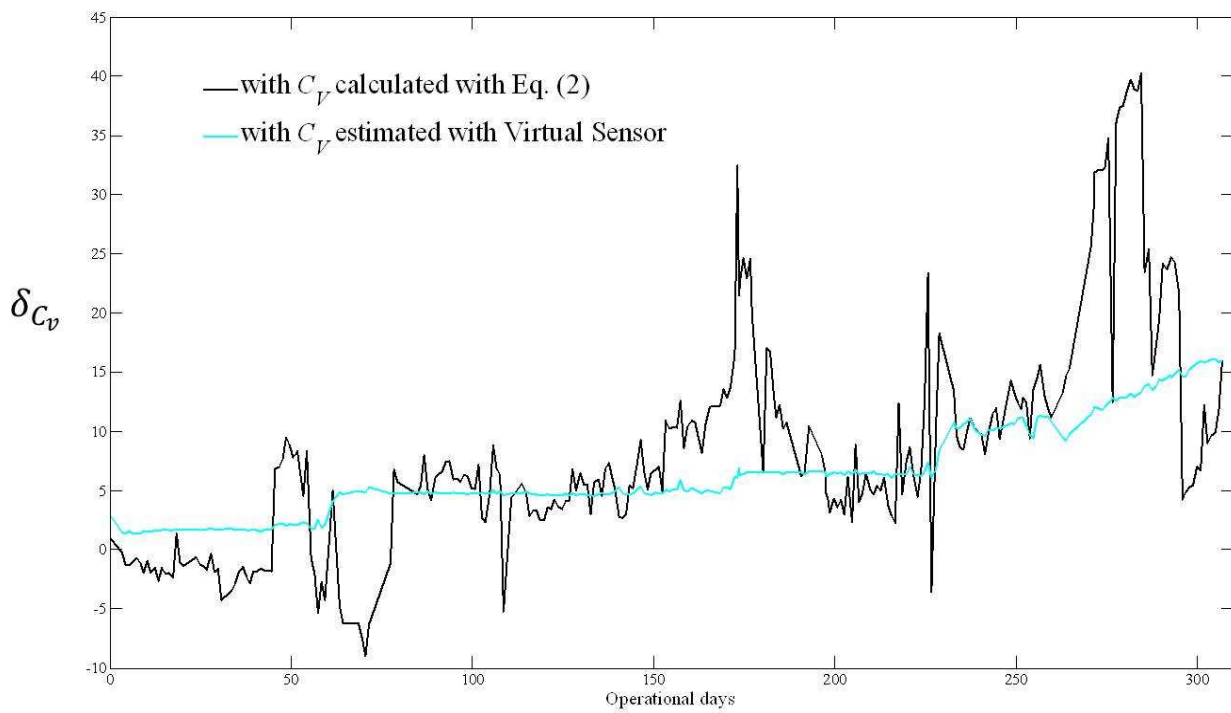


Figure 5. Erosion indication (δ_{C_v}) obtained with C_v calculated with Eq. (2) (black line) and with the Virtual Sensor (light blue line).

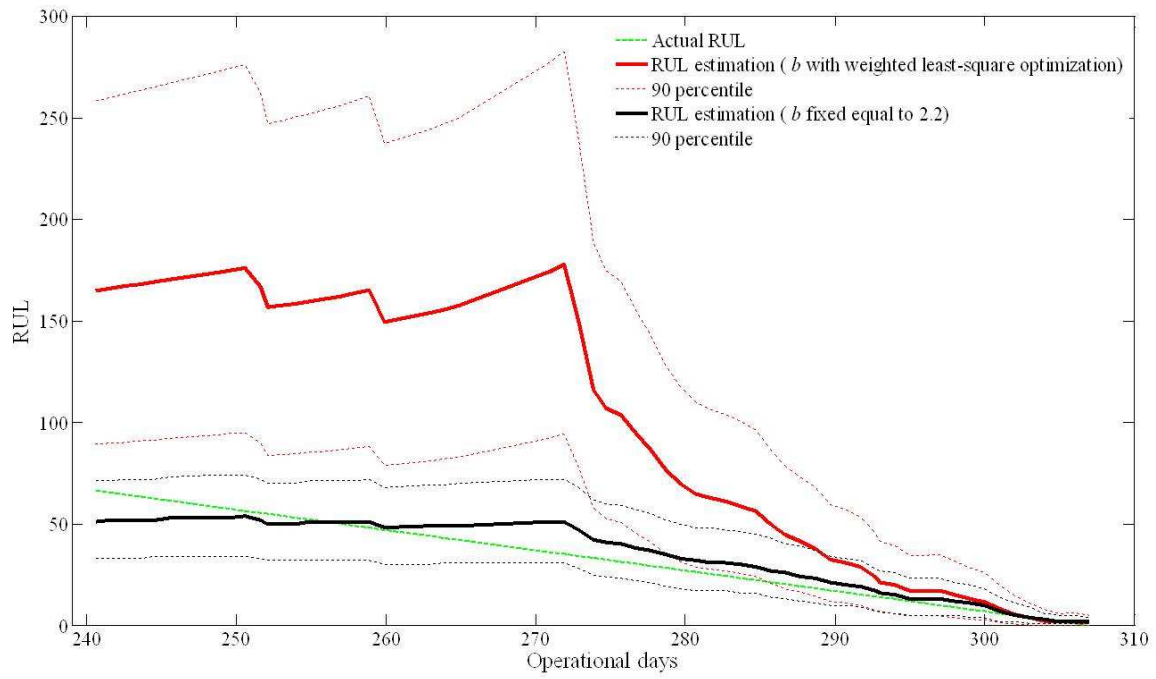


Figure 6. RUL estimation and uncertainty obtained with the gamma process when parameter b is calculated with the weighted least-square optimization (red lines) and when it is fixed to 2.2 (black line). The actual RUL is indicated by the green dashed line.

Gear Health Threshold Setting Based On a Probability of False Alarm

Eric Bechhoefer¹, David He², and Paula Dempsey³

¹ NRG Systems, Hinesburg, VT, 05461
erb@nrgsystems.com

² University of Illinois at Chicago, Department of Mechanical and Industrial Engineering, Chicago, IL, 69697, USA
davidhe@uic.edu

³ NASA, Glenn Research Center, Cleveland, OH, 69697, USA
paula.j.dempsey@nasa.gov

ABSTRACT

There is no established threshold or limit for gear vibration based condition indicators (CI) that indicates when a gear is in need of maintenance. The best we can do is set CI thresholds statistically, based on some small probability of false alarm. Further, to the best of our knowledge, there is no single CI that is sensitive to every failure mode of a gear. This suggests that any condition based maintenance system for gears will have some form of sensor fusion.

Three statistical models were developed to define a gear health indicator (HI) as a function of CI: order statistics (max of n CIs), sum of CIs and normalized energy. Since CIs tend to be correlated, a whitening process was developed to ensure the HI threshold is consistent with a defined probability of false alarm. These models were developed for CIs with Gaussian or Rayleigh (skewed) distributions. Finally, these functions, used to generate HIs, were tested on gear test stand data and their performance evaluated as compared to the end state of the gear (e.g. photos of damage). Results show the HIs performed well detecting pitting damage to gears.*

1 INTRODUCTION

Vibration based gear fault detection algorithms have been developed to successfully detect damaged on gears (McFadden and Smith 1985). Significant effort has also been expended to validate the efficacy of these

* Bechhoefer *et al.* This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

algorithms (Zakrajsek 1993, Lewicki *et al.* 2010). These studies have demonstrated the ability of gear CI algorithms to detect damage. However, they have not established standardized threshold values for a given quantified level of damage. Additionally, it has been shown (Wemhoff *et al.* 2007, Lewicki *et al.* 2010) that different algorithms are sensitive to different fault modes (Tooth Crack, Tooth Spacing Error, Tooth surfacing Pitting).

The concept of thresholding was explored by Byington *et al.* (2003), where for a given, single CI, a Probability Density Function (PDF) for the Rician/Rice statistical distribution was used to set a threshold based on a probability of false alarm (PFA). No single CI has been identified that works with all fault modes. This suggests that any functioning condition monitoring will use n number of CIs in the evaluation of gear health. A need exists for a procedure to set a PFA for a function using n number of CIs.

All CIs have a probability distribution (PDF). Any operation on the CI to form a health index (HI), is then a function of distributions (Wackerly *et al.* 1996). Functions such as:

- The maximum of n CI (the order statistics)
- The sum of n CIs, or
- The norm of n CIs (energy)

are valid if and only if the distribution (e.g. CIs) are independent and identical (Wackerly *et al.* 1996). For Gaussian distribution, subtracting the mean and dividing by the standard deviation will give identical Z distributions. The issue of independence is much more difficult.

Two CIs are independent if the probability (P) of CI_1 and CI_2 are equal to:

$$P(CI_1 \cap CI_2) = P(CI_1)P(CI_2) \quad (1)$$

Equivalently, CI_1 and CI_2 are independent random variables if the covariance of CI_1 , CI_2 is 0. This is, in general, not the case, where the correlation coefficient is defined as the covariance divided by the standard deviation:

$$\rho = \frac{Cov(CI_1, CI_2)}{\sigma_1 \sigma_2} \quad (2)$$

The range of correlation coefficients used in this study for pairs of gear CIs are listed in Table 1.

ρ_{ij}	CI 1	CI 2	CI 3	CI 4	CI 5	CI 6
CI 1	1	0.84	0.79	0.66	-0.47	0.74
CI 2		1	0.46	0.27	-0.59	0.36
CI 3			1	0.96	-0.03	0.97
CI 4				1	0.11	0.98
CI 5					1	0.05
CI 6						1

Table 1: Correlation Coefficients for the Six CI Used in the Study

This correlation between CIs implies that for a given function of distributions to have a threshold that operationally meets the design PFA, the CIs must be whitened (e.g. de-correlated). Fukinaga (1990) presents a whitening transform using the Eigenvector matrix multiplied by the square root for the Eigenvalues (diagonal matrix) of the covariance of the CIs.

$$\mathbf{A} = \Lambda^{1/2} \Phi^T \quad (3)$$

where Φ^T is the transpose of the eigenvector matrix, and Λ is the eigenvalue matrix. The transform can be shown to not be orthonormal, illustrating that the Euclidean distances are not preserved in the transform. While ideal for maximizing the distance (separation) between classes (such as in a Bayesian classifier), the distribution of the original CI is not preserved. This property of the transform makes it inappropriate for threshold setting.

If the CIs represented a metric such as shaft order acceleration, then one can construct an HI which is the square of the normalized power (e.g. square root of the acceleration squared). This can be defined as normalized energy, where the health index is:

$$HI = \sqrt{CI \times cov(CI)^{-1} \times CI^T} \quad (4)$$

Bechhoefer and Bernhard (2007) were able to whiten the CI and establish a threshold for a given PFA.

The objective of this analysis is to broaden the diagnostic capability available for gear health indexes by generalizing a method to develop HIs across CIs with other functions and statistical distributions.

1. GENERALIZED FUNCTION OF DISTRIBUTIONS

The desired linear transformation operates on the vector CI such that:

$$\begin{aligned} \mathbf{Y} &= \mathbf{L} \times \mathbf{CI}^T, \\ 0 &= \rho = correlation(\mathbf{Y}) \end{aligned} \quad (5)$$

where \mathbf{Y} preserves the original distribution of the CIs.

The Cholesky Decomposition of Hermitian, positive definite matrix results in $\mathbf{A} = \mathbf{L}\mathbf{L}^*$, where \mathbf{L} is a lower triangular, and \mathbf{L}^* is its conjugate transpose. By definition, the inverse covariance is positive definite Hermitian. It then follows that:

$$\mathbf{L}\mathbf{L}^* = \Sigma^{-1} \quad (6)$$

and

$$\mathbf{Y} = \mathbf{L} \times \mathbf{CI}^T \quad (7)$$

where \mathbf{Y} is 1 to n independent CI with unit variance (one CI representing the trivial case). The Cholesky Decomposition, in effect, creates the square root of the inverse covariance. This in turn is analogous to dividing the CI by its standard deviation (the trivial case of one CI). In turn, Eq. (7) creates the necessary independent and identical distributions required to calculate the critical values for a function of distributions.

1.1 Gear Health as a Function of Distributions

Prior to detailing the mathematical methods used to develop the HI, background information will be discussed. A common nomenclature for the user/operator of the condition monitoring system will be presented, such that the health index (HI) has a common meaning. The critical values (threshold) will be different for each monitored component, because the measured CI statistics (e.g. covariance) will be unique for each component type. The threshold will be normalized, such that the HI is independent of the component. Further, using guidance from GL Renewables (2007), the HI will be designed such that there are two alert levels: warning and alarm. Then the HI is defined such that the range is:

- 0 to 1, where the probability of exceeding an HI of 0.5 is the PFA
- A warning alert is generated when the HI is greater than or equal to 0.75
- An alarm alert is generated when the HI is greater than or equal to 1.0

2. HI BASED ON RAYLEIGH PDFs

The PDF for the Rayleigh distribution uses a single parameter, β , resulting in the mean ($\mu = \beta^*(\pi/2)^{0.5}$) and variance ($\sigma^2 = (2 - \pi/2) * \beta^2$) being a function of β . Note that when applying Eq. (7) to normalize and whiten the vector of CI data, the value for β for each CI will then be:

$$\begin{aligned} \sigma^2 &= 1, \\ \beta &= \sigma^2 / \sqrt{2 - \pi/2} = 1.5264 \end{aligned} \quad (8)$$

The PDF of the Rayleigh is:

$$f(x) = x/\beta^2 \exp(-x/2\beta^2) \quad (9)$$

The cumulative distribution function, the integral of (9) is:

$$F(x) = 1 - \exp(-x^2/2\beta^2) \quad (10)$$

It can be shown that the PDF of the magnitude of a frequency of a random signal is a Rayleigh PDF (Bechhoefer and Bernhard 2006). This property makes the Rayleigh an appropriate model for thresholds for shaft (Shaft order 1, etc) and bearing energies. The next section will demonstrate how this can be used appropriately for gears.

2.1 The Rayleigh Order Statistic

Consider a HI function which takes the maximum of n CIs. If the CIs are Independent and Identical (IID), then the function defines the order statistic. Given the order statistic PDF as (Wackerly 1996):

$$g(x) = n[F(x)]^{n-1} f(x) \quad (11)$$

The threshold is then calculated for t , from the inverse Cumulative distribution function (CDF):

$$1 - PFA = \int_{x=-\infty}^t n[F(x)]^{n-1} f(x) dx \quad (12)$$

For $n = 3$, PFA of 10⁻³, after solving the inverse CDF (Eq 12), the threshold t equals 6.1 (Note, the solution to Eq 12 can sometime require significant effort. See the Appendix for solution strategies). The HI algorithm, referred to as the Rayleigh Order Statistics (OS) is then:

$$HI = \max\{\mathbf{Y}\} \times 0.5 / 6.1 \quad (13)$$

Here $\mathbf{Y} = \mathbf{L} \times \text{CIT}$ (e.g. whitening and normalizing the CIs by applying Eq (7), which is scaled by 0.5 over the threshold. This then is consistent with the definition of the HI presented in 2.1, or a HI of 0.5 for the defined PFA.

2.2 The Sum of n Rayleigh

Consider a HI function which takes the sum of n CIs. If the CIs are Independent and Identical (IID), then the function defines a distribution with a Nakagami PDF (Bechhoefer and Bernhard 2007). Given the mean and variance for the Rayleigh, the sum of n normalized Rayleigh distributions is $n * \beta^*(\pi/2)^{0.5}$, with variance $\sigma^2 = n$. Given the Nakagami PDF as:

$$2\left(\frac{\eta}{\omega}\right)^\eta \frac{1}{\Gamma(\eta)} x^{(2\eta-1)} e^{-\eta/\omega x^2} \quad (14)$$

where Γ is the gamma function. Then, the statistics for the Nakagami are calculated as:

$$\eta = E[x^2]^2 / \text{Var}[x^2], \quad \omega = E[x^2] \quad (15)$$

which are used in the inverse Cumulative distribution function (CDF) to calculate the threshold.

For $n = 3$ CIs, the threshold is 10.125 and the HI algorithm, referred to as the Rayleigh normalized energy (NE) is then:

$$HI = 0.5 / 10.125 \sum_{i=1}^3 \mathbf{Y}_i \quad (16)$$

For a more in depth treatment of the Nakagami, see Bechhoefer and Bernhard (2007). Again, the dividing 0.5/10.125 allows Eq (15) to be consistent with the HI paradigm.

2.3 The Total Energy of n Rayleigh

Consider a HI function which takes the norm of n CIs, which represents the normalized energy. If the CIs are IID, it can be shown that the function defines a Nakagami PDF (Bechhoefer and Bernhard 2007). The mean is now $2*n*1/(2-\pi/2)^{0.5}$. Then, the statistics for the Nakagami are calculated as:

$$\eta = n, \quad \omega = 1 / (2 - \pi/2) * 2 * n \quad (17)$$

which are used in the inverse CDF to calculate the threshold. For our $n = 3$ CIs, the threshold is 6.259 and the HI algorithm, referred to as the Sum of Rayleigh (SR) is then:

$$HI = 0.5 / 6.259 \sqrt{\sum_{i=1}^3 \mathbf{Y}_i^2} \quad (18)$$

3. HI BASED ON GAUSSIAN PDFs

If it is found that the distribution of the CI data follows a Gaussian distribution a comparable mathematical process can be applied. Using similar constructs as applied to the Rayleigh PDF, we can generate

thresholds for the Gaussian distribution. The PDF of the Gaussian is:

$$f(x) = x/\sigma\sqrt{2\pi} \exp\left(-(x - \mu)^2/2\sigma^2\right) \quad (19)$$

The cumulative distribution function, the integral of Eq (19) is

$$F(x) = x/\sigma\sqrt{2\pi} \int_{-\infty}^x \exp\left(-(t - \mu)^2/2\sigma^2\right) dt \quad (20)$$

3.1 The Gaussian Order Statistic

Eq. 11 can be applied to the Gaussian PDF and CDF to derive the order statistic PDF of the Gaussian HI function:

$$f(x) = 3 \left[x/\sigma\sqrt{2\pi} \int_{-\infty}^x \exp\left(-(t - \mu)^2/2\sigma^2\right) dt \right]^2 \times x/\sigma\sqrt{2\pi} \exp\left(-(x - \mu)^2/2\sigma^2\right) \quad (21)$$

Again, we find the threshold by solving the inverse CDF of Eq (12). The PDF of the order statistic (OS) for a zero mean Gaussian is not bounded at zero, such as the Rayleigh. As such, to be consistent without the HI paradigm of lower HI range of 0, the OS PDF is shifted such the probability of the HI being less than or equal to zero is small. In this example, that probability is defined at 0.05%, corresponding to a PFA of 0.95 (e.g. a lower threshold). For $n = 3$, for a PFA of 0.95, lower threshold, t is -0.335, and upper threshold for a PFA of 10^{-3} , the threshold t is 3.41 (for HI of 0.5). The CIs are now a z distribution (Gaussian normalized with zero mean and unit variance). An additional rule is set such that any HI less than the lower 5% (corresponding to a PFA of 0.95) is an HI of zero. The HI algorithm is:

$$\mathbf{Y} = \mathbf{L} \times (\mathbf{CI}^T - \mathbf{m})$$

$$HI = (\max\{\mathbf{Y}\} + .34) \times 0.5 / (3.41 + 0.34) \quad (22)$$

where \mathbf{m} is the mean of the CIs. Subtracting the mean and multiplying by \mathbf{L} transforms the CIs into n , Z distributions (zero mean, IID Gaussian distributions).

3.2 The Sum of n Gaussian

Consider a HI function that takes the sum of n Gaussian CIs. Then the mean and variance of the sum of the CI are:

$$\mu = \sum_{i=1}^3 E[\mathbf{L}_i] \quad \sigma^2 = n \quad (23)$$

Again the inverse normal CDF is used to calculate the threshold. Similar to (22), an offset and scale value is

needed to ensure the HI is lower bounded to 0. For $n = 3$ CI, the mean, $\mu = 3$ and variance $\sigma^2 = 3$. Using the inverse normal CDF, the lower threshold (PFA of .95) is -0.15 and the and upper threshold (PFA 10^{-3}), is 8.352, then the HI algorithm is then:

$$\mathbf{Y} = \mathbf{L} \times \mathbf{CI}^T$$

$$HI = 0.5 / (8.352 - 0.15) \left(-0.15 + \sum_{i=1}^3 \mathbf{Y}_i \right) \quad (24)$$

3.3 The Total Energy of n Gaussian

Finally, we will consider a HI function that takes the norm of n Gaussian CIs. Again it can be shown that the function defines a Nakagami PDF (Bechhoefer and Bernhard 2007). The mean is $2*n*1/\text{sqrt}(2-\pi^2)$, with $\omega = n$ and η is $\omega/2$. Using the inverse Nakagami CDF to calculate the threshold for $n = 3$ CIs and a PFA of 10^{-3} , the threshold is: 3.368. The HI algorithm is then:

$$\mathbf{Y} = \mathbf{L} \times \mathbf{CI}^T$$

$$HI = 0.5 / 3.368 \sqrt{\sum_{i=1}^3 \mathbf{Y}_i^2} \quad (25)$$

4. APPLICATION TO GEAR FAULT

Vibration data from experiments performed in the Spiral Bevel Gear Test facility at NASA Glenn was reprocessed for this analysis. A description of the test rig and test procedure is given in Dempsey *et al.* (2002). The rig is used to quantify the performance of gear material, gear tooth design and lubrication additives on the fatigue strength of gears. During this testing, CIs and oil debris monitoring were used to detect pitting damage on spiral bevel gears (**Figure 1 Test Rig and Gears (Dempsey et al. 2002)**).

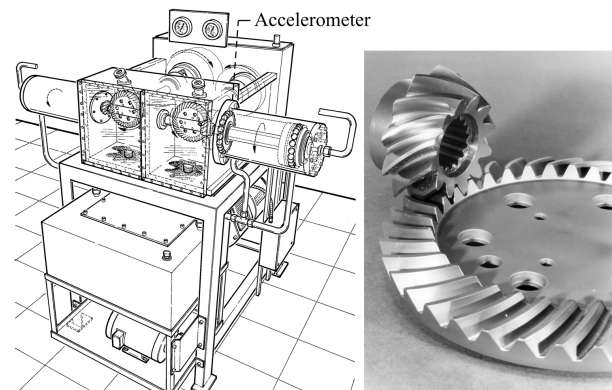


Figure 1 Test Rig and Gears (Dempsey et al. 2002)

The tests consisted of running the gears under load through a “back to back” configuration, with acquisitions made at 1 minute intervals, generating time synchronous averages (TSA) on the gear shaft (36

teeth). The pinion, on which the damage occurred, has 12 teeth.

TSA data was re-processed with gear CI algorithms presented in Zakrajsek *et al.* (1993) and Wemhoff *et al.* (2007), to include:

- TSA: RMS, Kurtosis (KT), Peak-to-Peak (P2P), Crest Factor (CF)
- Residual RMS, KT, P2P, CF
- Energy Operator RMS, KT
- Energy Ratio
- FM0
- Sideband Level factor
- Narrowband (NB) RMS, KT, CF
- Amplitude Modulation (AM) RMS, KT
- Derivative AM KT
- Frequency Modulation (FM) RMS, KT

From these CIs, a total of six CIs were used for the HI calculation: Residual RMS, Energy Operator RMS, FM0, NB KT, AM KT and FM RMS. These CIs were chosen because they exhibited good sensitivity to the fault. Residual Kurtosis and Energy Ratio also were good indicators, but were not chosen because;

- It has been the researcher's experience that these CIs become ineffective when used in complex gear boxes, and
- As the faults progresses, these CIs lose effectiveness. The residual kurtosis can in fact decrease, while the energy ratio will approach 1.

Covariance and mean values for the six CI were calculated by sampling healthy data from four gears prior to the fault propagating. This was done by randomly selecting 100 data points from each gear, and calculating the covariance and means over the resulting 400 data points.

The selected CI's PDF were not Gaussian, but exhibited a high degree of skewness. Because of this, the PDFs were "left shifted" by subtracting an offset such that the PDFs exhibited Rayleigh like distributions. Then, the threshold setting algorithms were tested for:

- Rayleigh order statistic (OS): threshold 8.37 for $n = 6$ and a PFA of 10^{-6} ,
- Rayleigh normalized energy (NE): threshold 10.88 for $n = 6$ and a PFA of 10^{-6} ,
- Sum of Rayleigh (SR): threshold 24.96 for $n = 6$ and a PFA of 10^{-6} ,

Figures 2, 4 and 6 are HI plots that compare the OS, NE and SR algorithms during three experiments in the test rig. The HI trend (in black) is plotted on top of the raw HI values (in blue). Figures 3, 5 and 7 show the amount of pitting damage on the pinion teeth at each test completion.

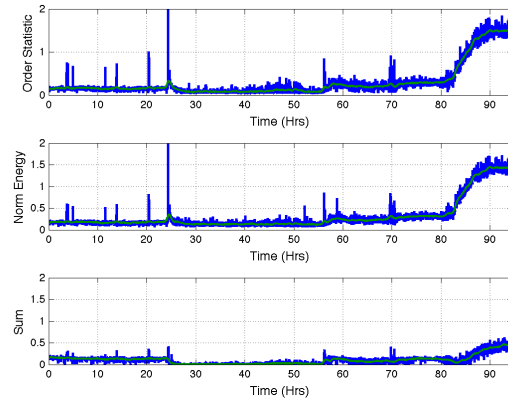


Figure 2 Test BV2_10_15_01EX4

Note that the spikes corresponded to changes in torque on the rig. All the HI algorithms were sensitive to damage, although in general, the best system response was from both the OS and NE.



Figure 3 Pitting Damage on EX4

Note that the decrease in the HI rate of change corresponds to a decrease in torque load towards the end of the test.

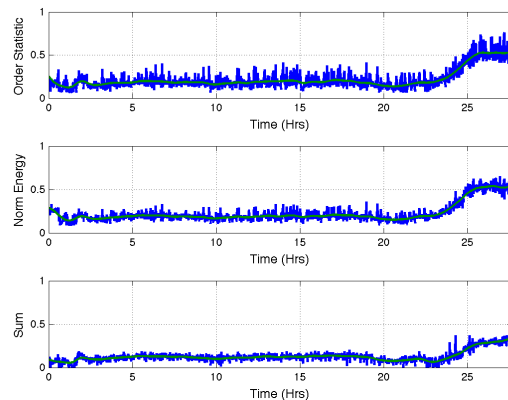


Figure 4 Test BV2_10_26_01EX5

For the data plotted in figure 4, this test appears to have been halted prior to heavy pitting damage, as the gear HI is reach only 0.5. However, the photo of gear EX5 (Figure 5) shows extensive pitting damage.



Figure 5 Damage on Gear EX5

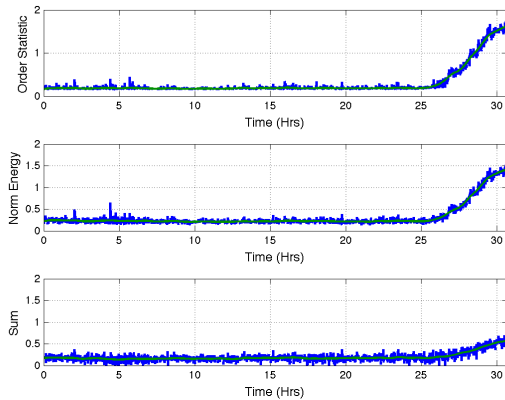


Figure 6 Test BV2_1_4_02EX6



Figure 7 Damage on Gear EX6

5. DISSCUSSION AND OBSERVATIONS

After the three statistical models were applied to the test rig CI data, it was observed that each HI algorithm performed well, although the OS and NE is clearly more sensitive to fault than the SR algorithm. Additionally, the measured RMS noise of the OS was 15% to 25% higher than the NE, that RMS value being approximately 0.05 HI. However, the most important

contribution is that a process has been developed to whiten CI data so that different HI algorithms can be explored with some assurance that, mathematically, the PFA performance was being met.

Additionally, it is encouraging that, based solely on nominal data (statistics taken prior to fault propagation), it was observed that:

- An HI of 1 displays damage warranting maintenance.
- That nominal data is approximately 0.1 to 0.2 HI, where the PFA was set for 0.5 HI
- That while no one CI seemed to work for every gear tested, the HI function captured the damage consistently (even for a small sample set).
- The HI trends were low noise. This can facilitate prognostics.

6. CONCLUSION

Thresholding is critical for the operation of a condition monitoring system. If the probability of false alarm (PFA) is too high, then the operator is flooded with numerous false alarms and tends to disregard alerts. Unfortunately, some of the alerts will be true, resulting in collateral equipment damage. If the PFA is low, but the probability of fault detection is low, then the operator cannot perform maintenance “on condition”. Again, there are missed faults resulting in collateral damage.

Because the condition indicators (CI) are correlated, without some pre-processing, it is difficult to operationally achieve the design PFA. A method was presented for whitening the CIs used in gear fault detection. The whitening was achieved by a linear transformation of the CI using the Cholesky decomposition of the inverse of the CIs covariance.

With this transformed, whitened CI data, a health indexed based on a specified PFA was demonstrated. Three candidate HI algorithms (order statistics, normalized energy and sum of CI) for two different CI probability distribution functions (Gaussian and Rayleigh), were presented and tested on three data sets of pitted gears from a test stand.

It was observed that the HI algorithms performed as designed: low PFA (e.g. noise) and good fault detection capability. Given this process, we will now expand the class of distributions that this can be applied to, for example, the Rice and Weibull distribution.

APPENDIX: Monte Carlo Techniques to Solve the Inverse CDF

The solution of the inverse CDF can be difficult for none standard distribution. In fact, most function of distributions are non-standard. Solutions for order statistic on Gaussians distribution are very problematic:

even solving using optimization techniques is nontrivial.

Alternatively, Monte Carlo techniques are relatively simple to set up, and give accuracy limited only by patients. For example, since the order statistic is defined as the maximum of n IID distribution, it is relatively easy to call 10 million random tuples of n distribution, take the maximum of each tuple, and sort to generate the CDF. The critical value corresponds to the index of the sorted values at 10 million x (1-PFA).

As an experiment, find the inverse CDF for the normal Gaussian with a PFA of 10⁻³. For 10 million, the index is 9990000. Running 100 experiments, the estimated critical value was: 3.090155199948529 vs. the actual value of 3.090232306167824. The PFA calculate from the Monte Carlo generated threshold was: 0.00100025, or an error of .025%.

REFERENCES

- McFadden, P., Smith, J., (1985), *A Signal Processing Technique for detecting local defects in a gear from a signal average of the vibration*. Proc Instn Mech Engrs Vol 199 No C4
- Zakrajsek, J. Townsend, D., Decker, H., (1993). *An Analysis of Gear Fault Detection Method as Applied to Pitting Fatigue Failure Damage*. NASA Technical Memorandum 105950.
- Lewicki, D., Dempsey, P., Heath, G., and Shanthakumaran P. (2010), *Gear Fault Detection Effectiveness as Applied to Tooth Surface Pitting Fatigue Damage*, Gear Technology, November/December 2010.
- Wemhoff, E., Chin, H., Begin, M., (2007), *Gearbox Diagnostics Development Using Dynamic Modeling*, AHS 63rd Annual Forum, Virginia Beach, 2007
- Byington, C., Safa-Bakhsh, R., Watson, M., Kalgren, P., (2003), *Metrics Evaluation and Tool Development for Health and Usage Monitoring System Technology*, HUMS 2003 Conference, DSTO-GD-0348
- Wackerly, D., Mendenhall, W., Scheaffer, R.,(1996), *Mathematical Statistics with Applications*, Buxbury Press, Belmont, 1996
- Fukunaga, K., (1990), *Introduction to Statistical Pattern Recognition*, Academic Press, London, 1990, page 75.
- Bechhoefer, E., Bernhard, A., (2007), *A Generalized Process for Optimal Threshold Setting in HUMS*, IEEE Aerospace Conference, Big Sky.
- GL Renewables, (2007), *Guidelines for the Certification of Condition Monitoring Systems for Wind Turbines*, [http://www.gl-](http://www.gl-group.com/en/certification/renewables/CertificationGuidelines.php)

[group.com/en/certification/renewables/CertificationGuidelines.php](http://www.gl-group.com/en/certification/renewables/CertificationGuidelines.php)

- Bechhoefer, E., Bernhard, A., (2006), *Use of Non-Gaussian Distribution for Analysis of Shaft Components*, IEEE Aerospace Conference, Big Sky.
- Dempsey, P., Handschuh, R., Afjeh, A. (2002), *Spiral Bevel Gear Damage Detection Using Decision Fusion Analysis*, NASA/TM-2002-211814

Gearbox Vibration Source Separation by Integration of Time Synchronous Averaged Signals

Guicai Zhang and Joshua Isom

United Technologies Research Center, East Hartford, CT 06108, USA

zhangg@utrc.utc.com

isomjd@utrc.utc.com

ABSTRACT

This paper describes a simple approach for integrating all the time synchronous average (TSA) signals from multiple shafts of a gearbox to generate a composite time synchronous average which can be subtracted from the original signal to generate a second-order cyclostationary residual. This approach is compared with other techniques including an all-shaft TSA over the least common multiple of shaft rotation periods, high-pass filtering, and self-adaptive noise cancellation (SANC). The results demonstrate that the proposed approach produces an integrated TSA signal that includes only the shaft components, gear mesh components and the sidebands associated with all the shafts, while the residual contains the random vibration components and noise. The results produced by three alternative techniques do not separate the components as well or have a lower signal-to-noise ratio.*

1. INTRODUCTION

Gearboxes are an important component in a wide variety of machinery including helicopters, wind turbines, aero-engines, and automobiles. Gearboxes tend to be complex with ever increasing needs of power transmission, speed change and compact size of modern equipments. A complex gearbox (e.g., the planetary gearboxes used in wind turbines and helicopters) may have several dozen gears, as many bearings, and five or more shafts rotating at different speeds. Failures in any of the components may cause the malfunction of the entire gearbox and the maintenance or replacement of the gearbox is of very high cost.

Fault diagnostics, prognostics and health management (PHM) for gearboxes is a great challenge and has attracted a lot of attention for the past over

thirty years (Welbourn, 1977; Randall, 1982; McFadden, 1986). The separation of vibration sources in a complex gearbox is critical for effectively and accurately diagnosing gearbox failures. Due to the complexity of the gearbox, there are typically more vibration sources than sensors; thus the use of fully determined source-separation techniques like independent component analysis (ICA) is limited.

In this work, we propose a source separation method based on the single shaft TSA. Specifically, it integrates the TSA components from each shaft and produces a composite signal including vibration sources from all the shafts and gears. When the resulting composite signal is subtracted from the original signal, one obtains a residual signal that contains the vibration components from bearings and random noise. The paper is organized as follows. Section 2 is a brief review of existing work on gearbox vibration source separation. Section 3 describes the algorithm for the proposed method and a demonstration using gearbox vibration data. Section 4 provides a justification for the proposed algorithm. Section 5 compares the source separation results obtained by the proposed method with those produced by other existing techniques. Section 6 and Section 7 contain a discussion and conclusion.

2. BRIEF REVIEW OF GEARBOX VIBRATION SOURCE SEPARATION TECHNIQUES

Vibration monitoring is the most widely used health-monitoring method for gearbox and other rotating machinery. A basic source separation technique long employed for gearbox health monitoring is the time synchronous average. The time synchronous average extracts periodic waveforms from a vibration mixture by averaging the vibration signal over several revolutions of the shaft of interest. This can be done in either the time or frequency domain. The time synchronous average technique enhances vibration features that are synchronous with a particular shaft, and attenuates features that are not synchronous with that shaft. The technique has proved to be useful for monitoring of gear and shaft health.

* This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Techniques for the separation of bearing vibration from other gearbox vibration sources are also available. The high-frequency resonance technique is one specifically designed for extracting features of local defects in bearings.

Adaptive noise cancellation (ANC) can be used to extract a faulty bearing signal in cases where the primary signal can be measured near the faulty bearing of a gearbox, and a secondary reference signal measured near another remote healthy bearing is also available. When one of the two components to be separated is deterministic (gear and shaft signals) and the other random (bearing signal), the reference signal can be made a delayed version of the primary signal. This is based on the fact that the random signal has a short correlation length while deterministic signal has long correlation length. Thus the adaptive filter will find the transfer function between the deterministic part of the signal and the delayed version of itself. The separation of the deterministic and random parts can also be achieved using one signal only, and this technique is called self-adaptive noise cancellation (SANC). The performance of the SANC algorithm depends on the choice of three parameters: the time delay, the filter length and the forgetting factor (Antoni and Randall, 2004, Zeidler, 1990). In practice, there are trade-offs between the parameter settings and signal properties such as the length of the measured signal.

Independent component analysis (ICA) is a standard technique for blind source separation. It has been applied to separate signals from two independent vibration sources recorded at two separate locations on a gearbox (Zhong-sheng, et al., 2004). The utility of this technique is limited to cases where the number of sensors is equal to or greater than the number of vibration sources, a condition that does not generally hold for gearbox vibration monitoring.

Principal component analysis (PCA) has also been used to identify the number of gearbox vibration sources (Gelle et al., 2003, Serviere et al., 2004). The utility of this dimensionality-reduction technique is limited by the fact that the reduced-dimension vibration features may not have physical significance and thus it is difficult to create a mapping between features and physical vibration sources.

3. THE INTEGRATED TIME SYNCHRONOUS AVERAGE

Although the TSA is a powerful tool for isolating gear and shaft components synchronous with a particular shaft, it fails to isolate random vibration components because the subtraction of a single TSA from the original signal results in a combination of random signals and other shaft-synchronous components.

A natural way to deal with the issue of separating shaft/gear components from random signals is to extend the standard single-shaft TSA to multiple shafts by conducting an average over the least common multiple of shaft rotation periods to include the components associated with all the shafts. Thus the residual only contains random components with the exclusion of the deterministic parts. However, this technique is impractical because the time period corresponding to the least common multiple of the shaft revolutions for an actual gearbox is usually several hours.

Recently, two methods have been described for subtracting a composite TSA from a vibration signal to produce a residual signal (Randall and Antoni, 2011). A frequency domain method consists of computing the FFT of the entirety of a signal, and simply removing spectral peaks at the discrete harmonics of the known periodic signals.

A second method, a time domain method, consists of multiple resampling and subtraction of time synchronous averages from a signal (Randall and Antoni, 2011).

Both of these issues have shortcomings. The frequency-domain approach must deal with the fact that the frequency of the harmonics leaks into adjacent bins, except for the very special case in which the numbers of sample per period and the number of periods are powers of two. In the case where there is leakage, removal of discrete peaks at shaft harmonics will not completely remove the periodic signal.

The time-domain approach described in (Randall and Antoni, 2011) has the issue that there are certain time-domain features that are common to the time synchronous average of two or more shafts – the signal corresponding to the gear mesh frequency being a universal example. Thus, repeated subtraction of individual TSAs will “over-subtract” certain features of the periodic signals.

In this work, an alternative method is proposed to integrate all the single shaft TSA signals to obtain a combination of the components synchronous with each shaft in a gearbox. The proposed method overcomes the limitations of the existing methods based on a time synchronous average, is simple, and performs better than other techniques not based on the time synchronous average.

The new algorithm is presented in Table 1 and justified in Section 4. To obtain the integration of the TSAs, all the single shaft TSA signals should have the same number of data points (this is actually the same spatial angle of one chosen reference shaft after angular resampling), and this can be achieved by interpolating and/or repeating the TSA time series. An FFT is then applied to each of the TSAs to get a complex series in the Fourier domain. Next, the magnitudes of each complex series are computed and a new series with the

same length is formed by taking the complex value which has the maximum magnitudes of all the single shaft TSA signals. The maximum magnitude series is used to create a time series using an inverse FFT operation. The new time series contains all the shaft components, mesh frequencies and their sidebands, and we call it the *integrated TSA*. A residual signal can be obtained by subtracting the integrated TSA from the original signal.

-
1. Read the original vibration data and tachometer data.
 2. Conduct TSA for each shaft: $Tsa1, Tsa2, Tsa3, \dots$
 3. Interpolate and repeat data to obtain $TSA1, TSA2, TSA3, \dots$, of the same length N .
 4. FFT to get complex series: $C_{TSA1} = \text{fft}(TSA1), C_{TSA2} = \text{fft}(TSA2), C_{TSA3} = \text{fft}(TSA3), \dots$
 5. Compute magnitude series: $A_{TSA1} = \text{abs}(C_{TSA1}), A_{TSA2} = \text{abs}(C_{TSA2}), A_{TSA3} = \text{abs}(C_{TSA3}), \dots$
 6. Obtain maximum magnitude series: $MaxA_{TSA} = \text{max}(A_{TSA1}, A_{TSA2}, A_{TSA3}, \dots)$
 7. For $i = 1:N$
 - if $A_{TSA1}(i) == MaxA_{TSA}(i)$
 $C_{Total}(i) = C_{TSA1}(i)$
 - elseif $A_{TSA2}(i) == MaxA_{TSA}(i)$
 $C_{Total}(i) = C_{TSA2}(i)$
 - elseif $A_{TSA3}(i) == MaxA_{TSA}(i)$
 $C_{Total}(i) = C_{TSA3}(i)$
 -
 - end
 - End
 8. Inverse FFT: $TC_{Total} = \text{iFFT}(C_{Total})$
 9. Integrated TSA time waveform = $\text{real}(TC_{Total})$
-

Table 1. Integrated TSA algorithm

4. ALGORITHM JUSTIFICATION

The objective of the time synchronous average is to extract a periodic signal synchronous with a particular shaft from a mixture of signals. If $y(t)$ is a signal that is periodic with period T , then it can be represented with the Fourier series expansion

$$y(t) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi t}{T} + b_n \sin \frac{n\pi t}{T} \right) \quad (1)$$

Conceptually, the objective of the time synchronous average is to extract only those portions of the signal that have frequency $\frac{n\pi}{T}$, $n = 1, \dots, \infty$.

In actuality, the time synchronous average consisting of the average of N periods for a signal with period T is equivalent to a comb filter with a frequency response given by $|H(f)| = \frac{1}{N} \frac{\sin(\pi N f T)}{\sin(\pi f T)}$ (Braun, 2011), which is plotted in Figure 1 as a function of f/T . The lobes of the comb filter naturally address the leakage issue. As N becomes large, the lobes of the filter become more tightly centered on the frequencies

$$f = \frac{n\pi}{T}, n = 1, \dots, \infty. \quad (2)$$

A filter selective for K periodic components corresponding to different shafts, each with period T_i , should be selective for the frequencies

$$f = \frac{n\pi}{T_i}, n = 1, \dots, \infty; i = 1, \dots, K. \quad (3)$$

This filter should be windowed to address leakage. Such a filter can be formed by merging multiple comb filters with a maximum-select rule,

$$|H(f)| = \max_i \left\{ \frac{1}{N_i} \frac{\sin(\pi N_i f T_i)}{\sin(\pi f T)} \right\}. \quad (4)$$

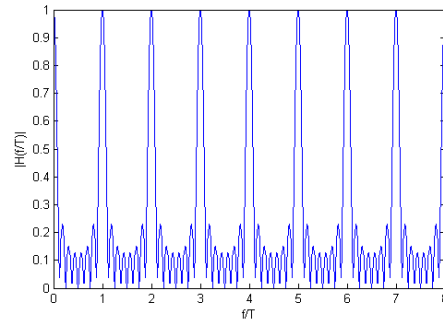


Figure 1. Frequency response of a comb filter equivalent to the time-synchronous average, as a function of f/T

The spectrum of this maximum-select comb filter is equivalent to the spectrum produced by Step 7 of the algorithm presented in Table 1. The maximum-select algorithm is actually equivalent to that selection of one of the non-zero complex values from the FFT series of the single-shaft TSAs along the frequency axis. This operation avoids producing redundant components in the integrated TSA.

5. EXPERIMENT

5.1 Application to Gearbox Vibration Data

In this section, the integrated TSA method described above is applied to vibration signals collected from a two-stage gearbox. This method is also compared with other source separation techniques, namely, the all-shaft TSA in which an average is conducted over the least common multiple of shaft revolutions; high/low-pass filtering; and self-adaptive noise cancellation (SANC).

5.2 Data

The data used for the demonstration of the proposed method is from the 2009 PHM Challenge.

The gearbox has three shafts (an input shaft, an idler shaft, and an output shaft), each with input side and output side bearings, and a total of four gears, one on the input shaft, two on the idler shaft, and one on the output shaft. During the experiments, a tachometer signal was collected from the input shaft with 10 pulses per revolution and two accelerometers were mounted on the input side and output side to collect vibration acceleration signals. The gear mesh configuration and sensor locations are illustrated in Figure 2.

For the method demonstration, selected data sets from the 2009 PHM Challenge were used. The data sets collected from the gearbox include both spur gear pair and helical gear pair configuration. The operating condition for the data sets used in this paper was the spur gear configuration, operating at high torque, with an input shaft speed is 3000 rpm (50Hz). The number of teeth for the two spur gear pairs are 32/96 and 48/80, respectively. The sampling frequency is 66.667 kHz and the sampling period is about 4 seconds for each data set.

The feature frequencies of the shafts and gears are listed as follows: the 1st and the 2nd mesh frequency are 1598 Hz and 799 Hz respectively, and the three shaft frequencies are 50 Hz, 16.7 Hz, and 10 Hz for the input, idler, and output shafts, respectively.

5.3 Demonstration of the integrated TSA method

In the following the results were generated from data set Spur 3 in which the gear of 48 teeth on the idler shaft is eccentric.

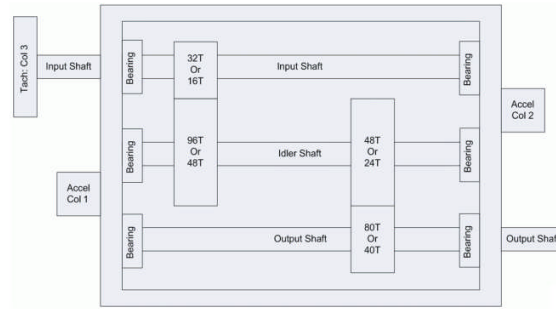


Figure 2. Configuration of the testing gearbox

Figure 3 shows the single shaft TSA waveforms for the input shaft, idler shaft and the output shaft (one revolution for each shaft), respectively. The corresponding time periods for the three shafts are different due to the different rotation speeds of the shafts. The eccentric feature is clearly evident in the TSA waveform of the idler shaft.

Figure 4 shows the original time waveform of the same acceleration signal (Spur 3), the integrated TSA waveform and the residual waveform.

Figure 5 shows magnitude spectra corresponding to the time waveforms shown in Figure 4. Only the frequency range below 2000 Hz is shown which covers the lower orders of the shaft harmonics, the two mesh frequencies (1598 Hz and 799 Hz) and their sidebands in Figure 5. From Figure 5 it can be seen that in the residual signal the shaft components and the gear mesh components are removed or attenuated significantly. It can also be seen from the magnitude spectrum of the original signal that the peak at the second mesh frequency (799 Hz) is significantly larger than that of the first mesh frequency (1598 Hz), and the peak at the input shaft frequency (49.9 Hz) is also one of the dominant components. The spectrum of the integrated TSA basically contains the shaft frequencies and their harmonics, as well as the mesh frequencies with the sidebands from the three shafts, and this can be seen most clearly in the following zoomed-in plots.

Figure 6 provides a comparison of the spectra of the original signal, the integrated TSA, and the residual signal at lower frequency band. From Figure 6 it can be seen that the integrated TSA basically contains the harmonics of the three shafts and in the residual signal the major periodic components (mainly the three shaft frequencies and their harmonics) are removed.

Figure 7 show the spectral comparison between the original signal and the integrated TSA zoomed-in around the two mesh frequencies. From Figure 7 it can be seen that the integrated TSA mainly contains the mesh frequencies and the sidebands associated with all the three shafts. And the dominant sidebands are caused by the idler shaft on which there is an eccentric gear.

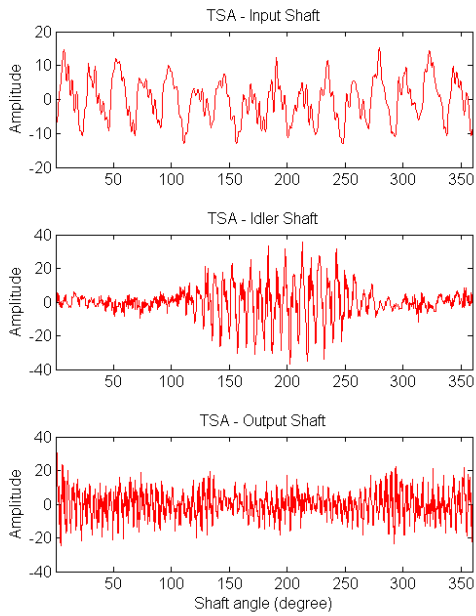


Figure 3. Single-shaft TSA waveforms

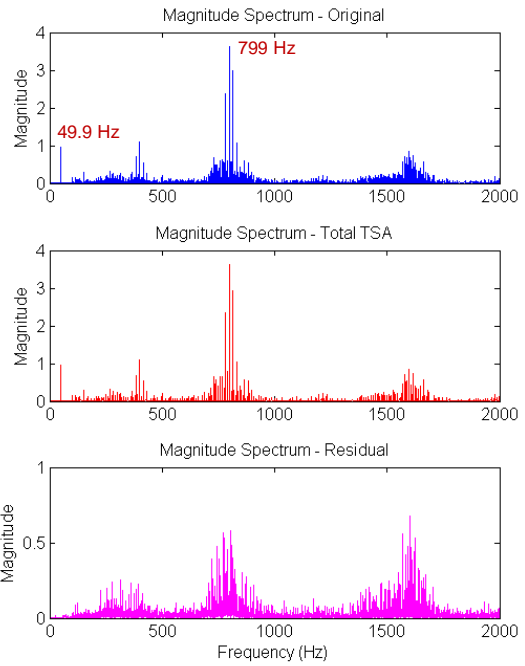


Figure 5. Magnitude spectra of the signals shown in Figure 4

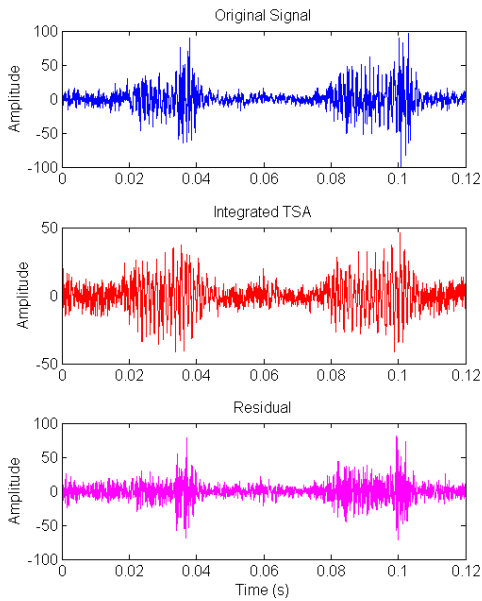


Figure 4. Waveforms for the original signal, the integrated TSA, and the residual

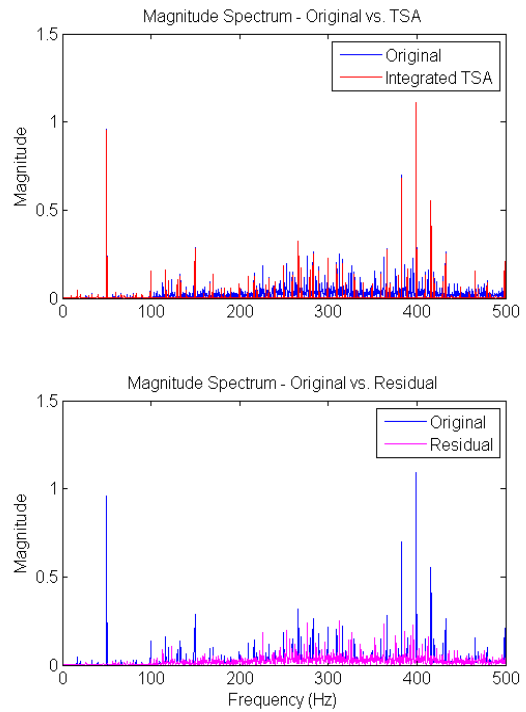


Figure 6. Spectrum comparison at lower frequency band

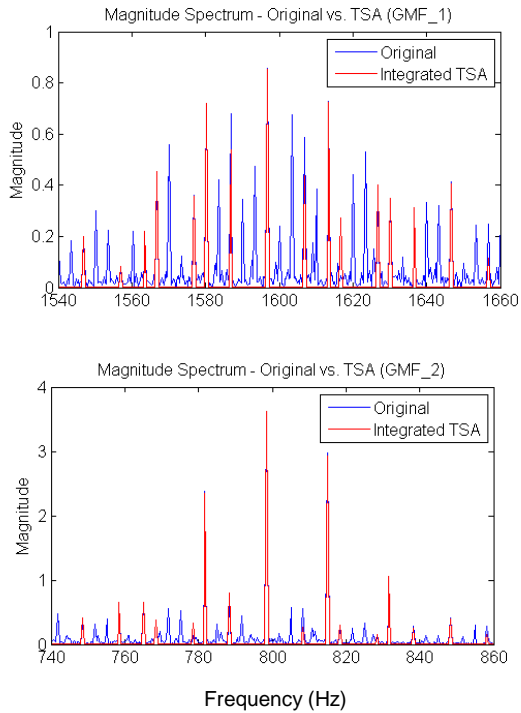


Figure 7. Spectrum comparison around the two gear mesh frequencies

5.4 Comparison with other techniques

In this section, the integrated TSA is compared with other techniques, namely, all-shaft TSA, high/low-pass filtering and SANC, using data from the PHM 2009 Challenge.

In these comparisons, data set Spur 6 with a rotation speed of 3000 rpm for the input shaft and a high load condition is used. There are some compound faults seeded in this data set, specifically, a broken tooth on the gear (80 teeth) installed on the output shaft, a defect on the bearing inner race, a ball defect, and an outer race defect in the bearings supporting the input side of the input shaft, idler shaft and the output shaft respectively. The input shaft is also imbalanced.

The cut-off frequency used in the high/low-pass filtering method is set to 5000 Hz which is roughly equal to three times the gear mesh frequency of 1598 Hz plus the fourth order sideband of the input shaft frequency 50 Hz (the highest shaft frequency among the three shafts).

Figure 8 shows the time-domain waveform of the data set, which evidences strong impulses caused by the broken tooth.

Figure 9 shows the magnitude spectrum of the original signal. From Figure 9 it is seen that the rotation frequency of the input shaft is the dominant component for this data set and has a much higher magnitude than the gear mesh vibration components.

Figure 10 (a), (b), (c) and (d) show the separated periodic components and the random transient components produced by the integrated TSA, all-shaft TSA, high/low-pass filtering and SANC, respectively. From Figure 10 it can be seen that the results from the separated results by using integrated TSA and all-shaft TSA look very close to each other. The other two methods -- high/low-pass filtering and SANC -- cannot filter out the broadband impulses from the periodic portion of the signal. One also notes some over-attenuation at the start of the filtered signal in the results of the SANC.

Figure 11 (a), (b), (c) and (d) provides a comparison of the magnitude spectra of the separated components in the lower frequency bands (<500Hz). It can be seen that the all-shaft TSA includes a more components than that of the integrated TSA in the low frequency band, while in the filtered parts of the other two filtering approaches almost all the components in low frequency bands are kept. The magnitudes of the residual spectra for high/low-pass filtering and SANC are close to zero, while some larger peaks (random components) could be found in the residual spectra of the integrated TSA and all-shaft TSA.

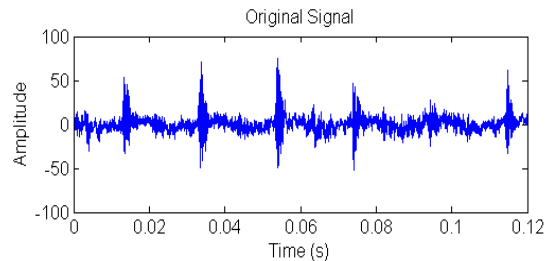


Figure 8. Original time waveform (Spur 6)

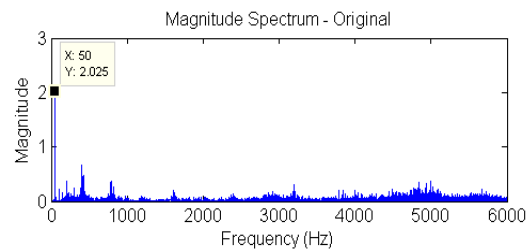
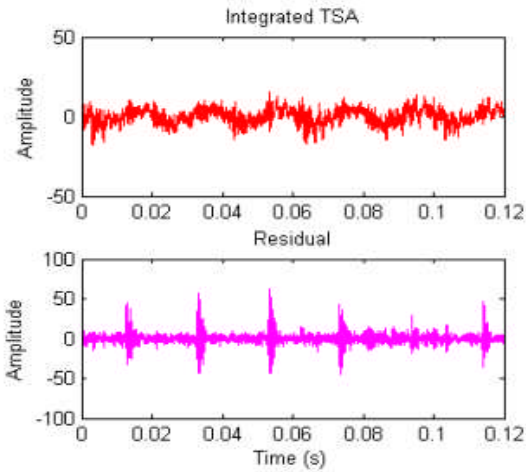
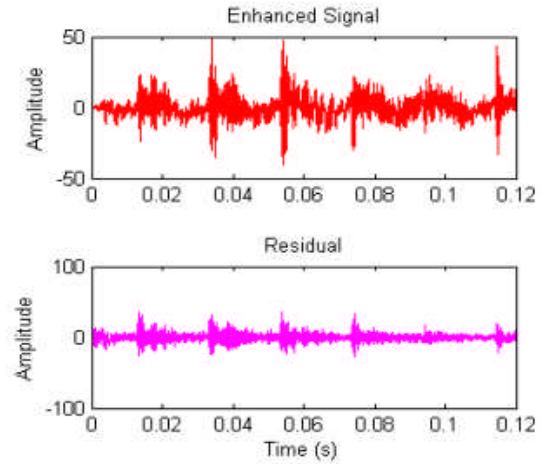


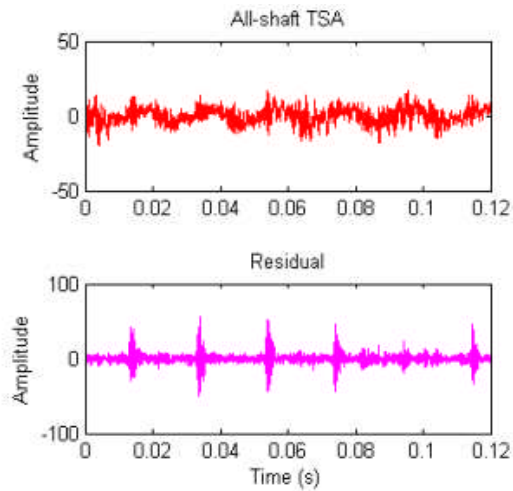
Figure 9. Magnitude spectrum of the original signal (Spur 6)



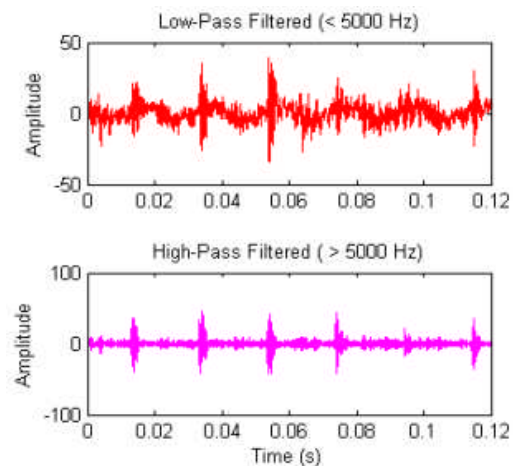
(a) Integrated TSA



(d) SANC



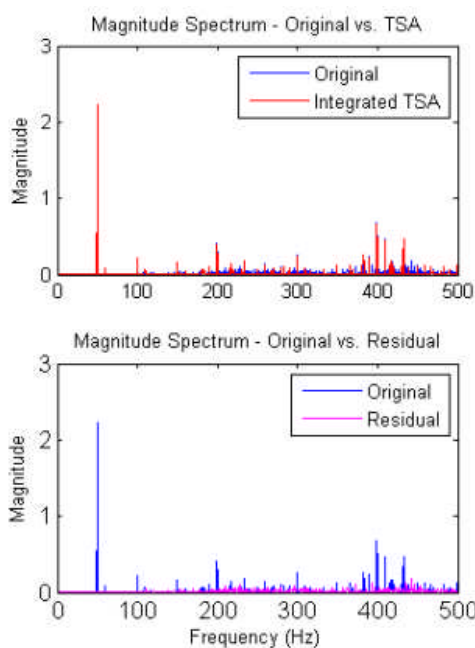
(b) All-shaft TSA



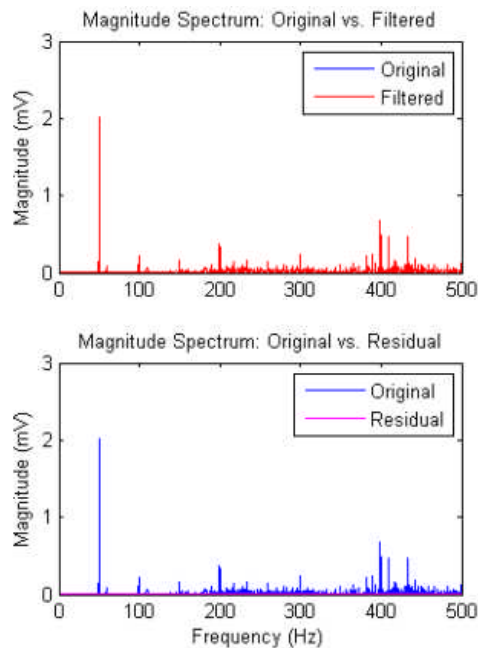
(c) High-Pass filtering

Figure 10. Comparison of the separated signals

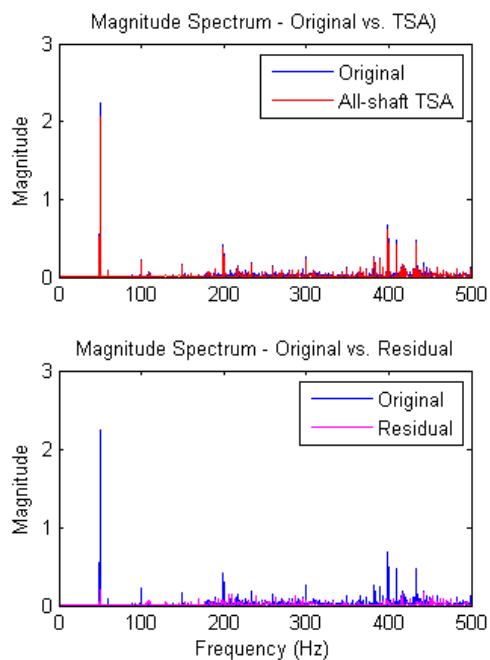
Figure 12 provides a comparison of the magnitude spectra of the separated components for the integrated TSA, all-shaft TSA and SANC methods around the second gear-pair mesh frequency ($GMF2 = 799$ Hz). And Figure 13 provides a comparison of the magnitude spectra of the separated components for the integrated TSA, all-shaft TSA and SANC methods around the first gear-pair mesh frequency ($GMF1 = 1598$ Hz). The result of high-pass filtering are not shown here as the cut-off frequency of 5000 Hz is much higher than these frequency bands and thus all the original components are kept in the filtered signals and the magnitudes of the residual signals are basically zero in the frequency bands compared here. It can be seen from Figure 12 and Figure 13 that the integrated TSA contains the mesh frequencies and their sidebands from all three shafts. The all-shaft TSA includes more periodic components than that of the integrated TSA and it is seen that there are some peaks between the mesh frequencies and the sidebands in the all-shaft TSA spectrum. From the spectrum of the filtered signal by SANC, it is seen that it covers all the major peaks and its residual signal is of very small magnitudes. It can also be seen from these spectra that the residual is complementary with the TSA or filtered signal to form the original signal, as expected.



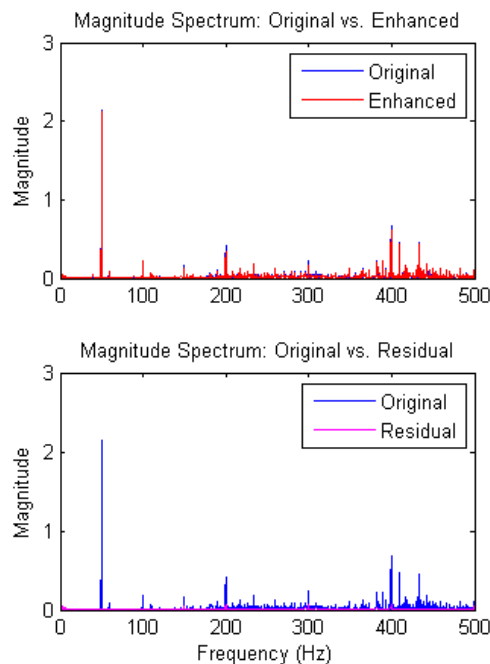
(a) Integrated TSA



(c) High-Pass filtering

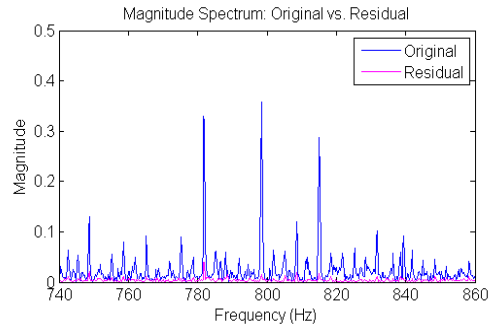
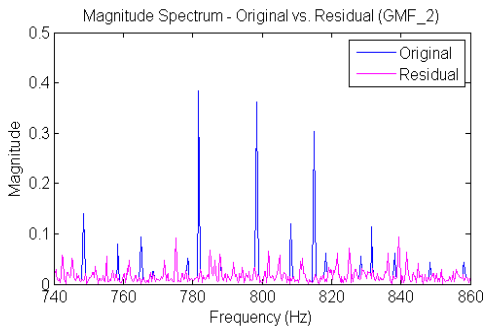
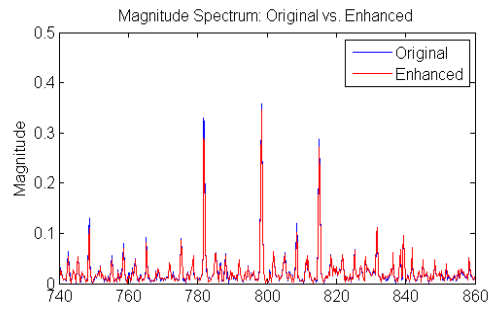
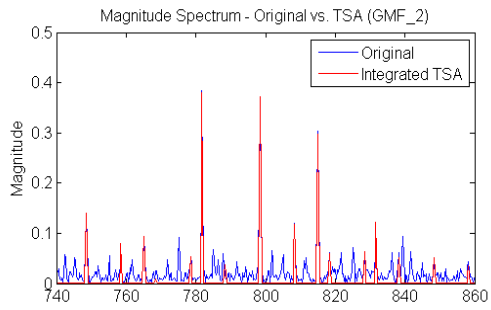


(b) All-shaft TSA



(d) SANC

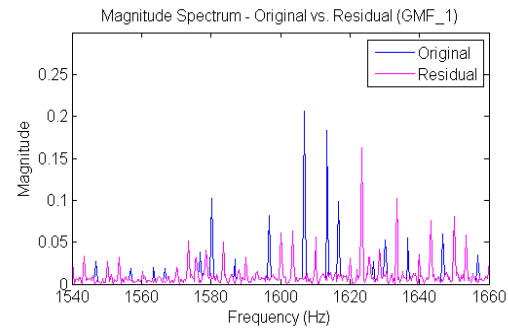
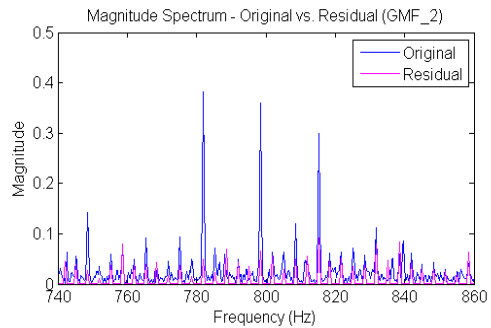
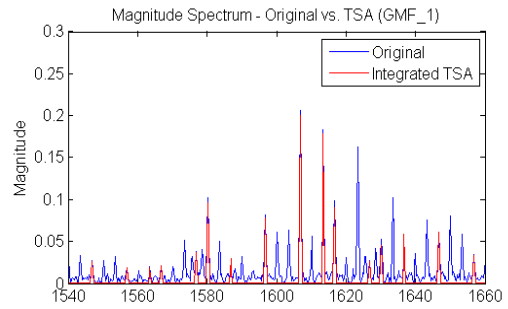
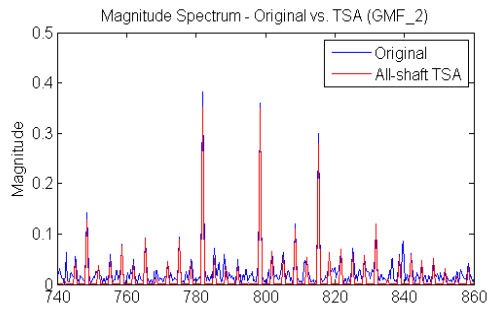
Figure 11. Spectrum comparison of the separated signals (in low frequency bands)



(a) Integrated TSA

(c) SANC

Figure 12. Spectrum comparison of the separated signals (zoomed in around GMF2)



(b) All-shaft TSA

(a) Integrated TSA

Figure 13. Spectrum comparison of the separated signals (zoomed in around GMF1)

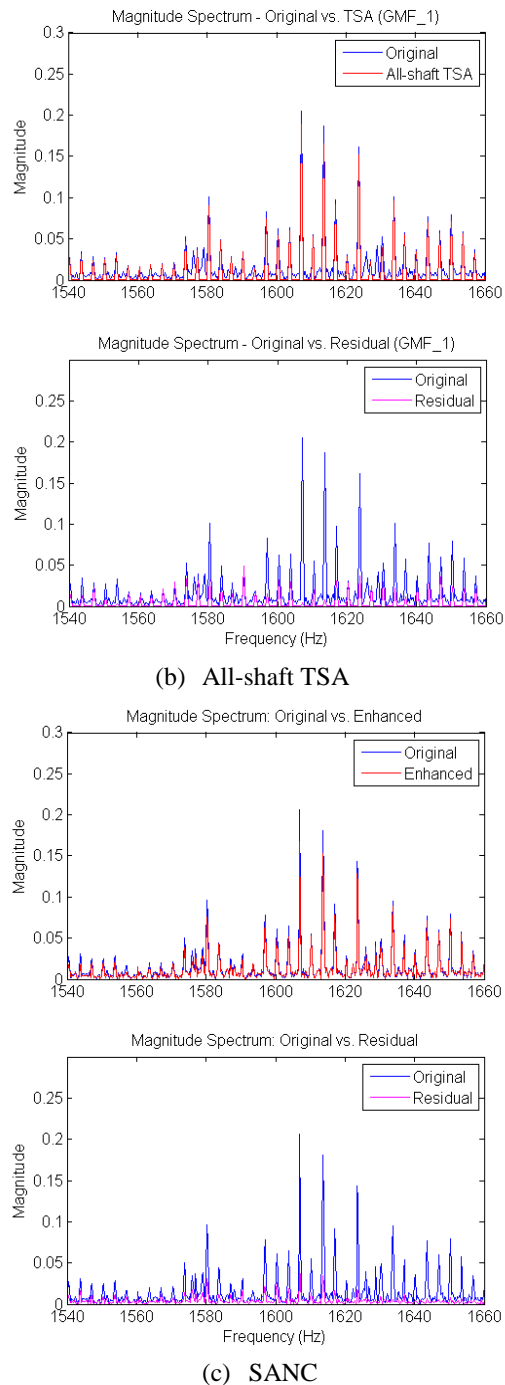


Figure 13 (continued). Spectrum comparison of the separated signals (zoomed in around GMF1)

6. DISCUSSION

Based on the analysis results in the above section, the following observations can be made:

1. The integrated TSA keeps all the shaft synchronous components including the shaft rotation frequencies and their harmonics, gear mesh frequencies and their harmonics with the sidebands from all the shafts. Vibration components other than these are significantly attenuated or eliminated.

2. The all-shaft TSA includes more vibration components than the integrated TSA. There are basically two reasons for this: (a) averaging over the least common multiple of the shaft revolutions includes the components with period of the least common multiple of the periods of the shafts (the peaks between the mesh frequencies and the shaft sidebands); (b) averaging over longer period also makes the number of averages smaller and this makes the signal-to-noise ratio lower (equivalent to higher side lobes of the comb filter). The data used in this paper is a special case in which all-shaft TSA can be applied. However, as mentioned in Section 3, for industrial gearboxes the least common multiple of the shaft revolution periods is usually several hours, making the technique impractical.

3. The high/low-pass filtering method separates the data according to the frequency range. Ideally the low frequency part should include the vibration components of the shafts and the gear mesh. However, the low frequency part surely includes all the periodic components not associated with the shafts and gears and noise components in the frequency bands lower than the cut-off frequency.

4. The filtered part of SANC is determined by the time delay factor, forgetting factor, and filter length. These should be optimized with the specific data type and data length to reach some trade-offs and in practice this is somewhat arbitrary and difficult to obtain satisfactory results.

5. Both high/low pass filtering and SANC fail to filter out broadband impulses. The transient feature is clearly seen in the deterministic part.

7. CONCLUSIONS

This paper describes a simple approach for integrating the TSA of individual shafts to generate a composite time synchronous average which can be subtracted from the original signal to generate a second-order cyclostationary residual. This approach is applied to vibration signals collected from a two-stage gearbox and compared with other techniques including an all-shaft TSA in which angular re-sampling over the least common multiple of shaft revolutions is conducted, high/low pass filtering and self-adaptive noise cancellation.

The results demonstrate that by using the proposed approach, the integrated parts contain only the components synchronous with each of the shafts

including the shaft frequencies and their harmonics, mesh frequencies and their harmonics, and the sidebands caused by all the shafts around the harmonics of the mesh frequencies. The all-shaft TSA signal contains more components than that of synchronous with each shaft and has a lower signal-to-noise ratio than the integrated TSA. High/low pass filtering and SANC induce more noise in the filtered part and the results are less satisfactory.

The integrated TSA is a simple but powerful approach for the separation of a gearbox vibration signal into first-order and second-order cyclostationary components. The new technique will facilitate the diagnosis of faults in complex gearbox systems.

REFERENCES

- Antoni, J., Randall, R.B. (2004). Unsupervised noise cancellation for vibration signals: part I - evaluation of adaptive algorithms, *Mechanical Systems and Signal Processing* 18 (2004) 89–101
- Braun, S. (2011). The synchronous (time domain) average revisited, *Mechanical Systems and Signal Processing*, 25 (2011) 1087-1102
- Gelle, G., Colas, M., Serviere, C. (2003). Blind source separation: a new pre-processing tool for rotating machines monitoring, *IEEE Transactions on Instrumentation and Measurement*, vol. 52, pp. 790-795.
- McFadden, P.D. (1986). Detecting fatigue cracks in gears by amplitude and phase demodulation of the meshing vibration. *ASME Transactions Journal of Vibration, Acoustics, Stress and Reliability in Design* 108, 165–170.
- PHM Challenge 2009 Data Sets:
<http://www.phmsociety.org/references/datasets>
- Randall, R.B. (1982). A new method of modeling gear faults. *ASME Journal of Mechanical Design*, 104, 259–267.
- Randall, R.B., Antoni, J. (2011). Rolling Element Bearing Diagnostics — A Tutorial, *Mechanical Systems and Signal Processing*, 25(2011), 485–520.
- Serviere, C., Fabry, P. (2004). Blind source separation of noisy harmonic signals for rotating machine diagnosis, *Journal of Sound and Vibration*, vol. 272, pp. 317-339.
- Tan, C.C., Dawson, B. (1987). An adaptive noise cancellation approach for condition monitoring of gearbox bearings, *Proceedings of the International Tribology Conference*, Melbourne, 1987.
- Welbourn, D.B. (1977). Gear Noise Spectra - a Rational Explanation, *ASME 1977 International Power Transmission and Gearing Conference*, Chicago, 28-30 Sept 1977
- Zeidler, J.R. (1990). Performance analysis of LMS adaptive prediction filters, *Proceedings of the IEEE*, 1990, 78 (12), 1781–1806.
- Zhong-sheng, C., Yong-min, Y., Guo-ji, S. (2004). Application of independent component analysis to early diagnosis of helicopter gearboxes. *Mechanical Science and Technology*, 2004, 23(4), 481-484.

Guicai Zhang received the M.S. from Xi'an JiaoTong University, Xi'an, China in 1993, the Ph.D. from Huazhong University of Science and Technology, Wuhan, China in 2000, both in Mechanical Engineering. He is a Staff Research Engineer in the area of prognostics and health management at United Technologies Research Center (China). Before joining United Technologies Research Center in 2005, he was an Associate Professor at Shanghai JiaoTong University from 2003 to 2005 and a Research Associate at the Chinese University of Hong Kong from 2000 to 2003.

Joshua D. Isom received the B.S. from Yale University in 2000, the M.S. from Rensselaer at Hartford in 2002 and the Ph.D. from the University of Illinois at Urbana-Champaign in 2009. He is a Principal Engineer in the area of prognostics and health management at United Technologies Research Center. Before joining United Technologies Research Center in 2010, he was the technical lead for prognostics and health management at Sikorsky Aircraft Corporation from 2007 to 2010. He was a lead systems engineer at UTC Power in South Windsor, CT from 2000 through 2007.

Health Monitoring of an Auxiliary Power Unit Using a Classification Tree

Wlamir O. L. Vianna¹, João P. P. Gomes¹, Roberto K. H. Galvão², and Takashi Yoneyama² and Jackson P. Matsuura²

¹ *EMBRAER, São Jose dos Campos, São Paulo, 12227-901, Brazil*
wlamir.vianna@embraer.com.br
joao.pordeus@embraer.com.br

² *ITA – Instituto Tecnológico de Aeronáutica, São José dos Campos, São Paulo, 12228-900, Brazil*
kawakami@ita.br
takashi@ita.br
jackson@ita.br

ABSTRACT

The objective of this work is to present a method to monitor the health of Auxiliary Power Units (APU) using a Dynamic Computational Model, Gas Path Analysis and Classification and Regression Trees (CART). The main data used to train the CART consists of measurements of the exhaust gas temperature, the bleed pressure and the fuel flow.

The proposed method was tested using actual APU data collected from a prototype aircraft. The method succeeded in classifying several relevant fault conditions. The few misclassification errors were found to be due to the insufficiency of the information content of the measurement data.*

1. INTRODUCTION

Increased aircraft availability is one of the most desirable fleet characteristics to an airliner. Delays due to unanticipated system components failures cause prohibitive expenses, especially when failures occur on sites without proper maintenance staff and equipments. In recent years researches have focused on providing new technologies which could prevent some failures or notify maintenance staff in advance when any component is about to fail. Health Monitoring (HM) provides this knowledge by estimating the current health state of components. This may guide the maintenance activities and spare parts logistics to properly remove or fix the component at the most suitable time and place.

* Vianna, W. O. L. et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Since the Auxiliary Power Unit (APU) represents a significant maintenance cost to an airliner, several HM studies have been conducted on this component (Vieira et al, 2009; Urban, 1967; Jones, 2007). Many of them exploit similar approaches to methods devoted to the main engines, due to the similarities in physical behavior.

Methods based on thermodynamic models, or gas path analysis, may provide more precise information as compared to data-driven methods. However, the use of model-based techniques still presents challenges when dealing with a large and heterogeneous fleet.

This paper aims to provide a HM solution based on a classification and regression tree (CART) employing data obtained from a mathematical model of an APU derived from thermodynamic principles. The proposed method is validated with APU field data.

The work is organized as follows. Section 2 contains a brief description of the system under analysis. Section 3 presents the methodology adopted. Section 4 contains the model description used on the implementation. Section 5 presents the implementation steps and the results of the method applied on the APU performance data. The last section presents the conclusion of the study and some remarks.

2. SYSTEM DESCRIPTION

An APU is a gas turbine device on a vehicle with the purpose of providing power to other systems apart from engines. This power can either be of pneumatic nature, extracted from a bleed system, or of electrical type, extracted from the generator. APUs are commonly found on large aircraft, as well as some large land vehicles. Its primary purpose is to provide bleed to start the main engines. It is also used to run accessories such

as air conditioning and electric pumps. It is usually located at the tail end of the aircraft as represented in Figure 1.

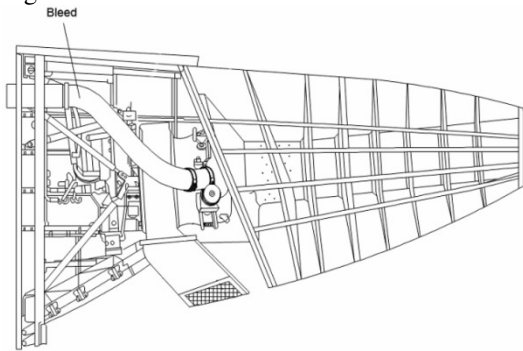


Figure 1: APU at the tail end of an aircraft.

A typical gas turbine APU contains a compressor, a burner and a turbine as every conventional gas turbine. It also has a bleed system that controls the amount of extracted pneumatic power, a fuel system, a gearbox and a generator. Protective components such as anti-surge, and guide vane may also be present. The logics and control are executed by the Full Authority Digital Engine Control (FADEC). A simplified APU representation is illustrated in Figure 2.

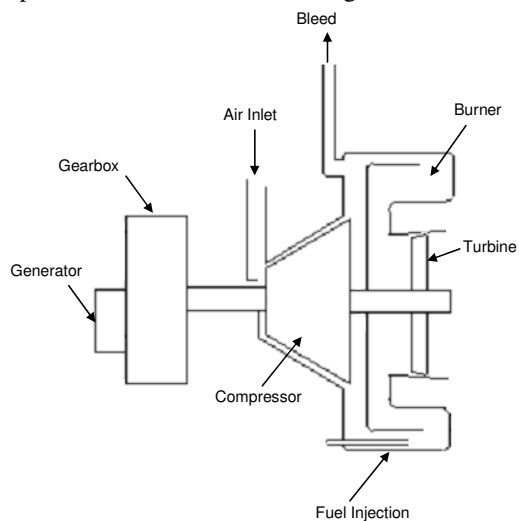


Figure 2: Simplified APU representation.

In order to provide proper information to the FADEC, the system must contain sensors for several variables, such as speed, exhaust gas temperature (EGT) and bleed pressure. A fuel flow meter may also be valuable but it is not an essential sensor. The EGT is a useful parameter for health monitoring and can indicate several failures such as core degradation and inlet blockage (SAE, 2006). A typical EGT profile during APU operation is indicated in Figure 3.

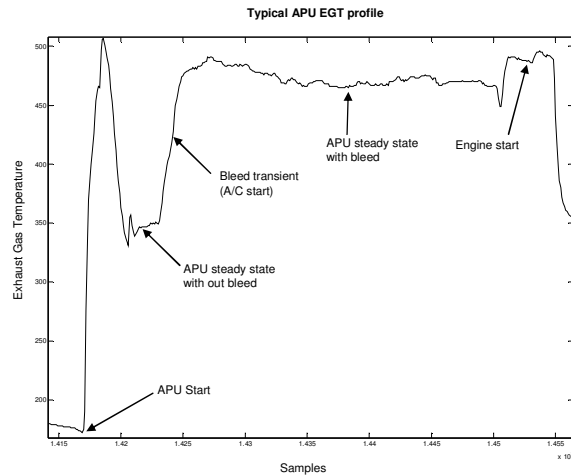


Figure 3: Typical EGT profile during the operation of an APU.

During the APU start an EGT peak is observed and can be used as a health monitoring feature (SAE, 2006). After the speed has reached its operational value the EGT stabilizes until the air conditioning system turns on. This produces an increase in EGT, which then reaches another steady-state value. When the engine starts, usually all other pneumatic sources are turned off so the APU can provide the required bleed.

3. HEALTH MONITORING METHODOLOGY

Gas path analysis is a methodology for monitoring gas turbines proposed by (Urban, 1967), which has been used in several studies for the purpose of health performance analysis (Saravanamuttoo et al, 1986) , (Li, 2003). Within the scope of APU monitoring, one of the main challenges consists of discriminating among possible failure modes affecting different components. In this context, promising results have been obtained with the use of classification methods (Vieira et al, 2009), (Sabyasachi, 2008).

Classification and Regression Trees (CART) are a popular set of classification methods that have as one of its key characteristics the easiness of interpretation of the results. This feature facilitates the validation of the results or the adjustment of the classification rules on the basis of the knowledge of a system specialist.

CART uses a “learning sample” of historical data with assigned classes for building a “decision tree”, which expresses a set of classification rules in terms of a sequence of questions. The use of CART involves three stages (Timofeev, 2004):

1. Construction of the maximum tree
2. Selection of an appropriate tree size
3. Classification of new data using the resulting tree

The classification tree uses some rules to split the “learning sample” into smaller parts, thus creating the nodes and the tree itself. Such rules are called “splitting

rules”. Some examples are the “Gini splitting rule” and the “Twoing splitting rule”. The first one is the most broadly used (Timofeev, 2004) and uses the following “impurity” function:

$$i(t) = \sum_{k \neq l} p(k | t) \quad (1)$$

where k and l are class indexes, t is the node under consideration and $p(k|t)$ is the conditional probability of class k provided in node t .

At each node the CART solves a maximization problem of the form:

$$\arg \max_{x_j \leq x_j^k, j=1, \dots, M} [i(t_p) - P_l i(t_l) - P_r i(t_r)] \quad (2)$$

where P_l and P_r are probabilities of the left and right node respectively.

Using the Gini impurity function, the following maximization problem must be solved to isolate the larger class from other data

$$\arg \max_{x_j \leq x_j^k, j=1, \dots, M} [-\sum_{k=1}^K p^2(k | t_p) + P_l \sum_{k=1}^K p^2(k | t_l) + P_r \sum_{k=1}^K p^2(k | t_r)] \quad (3)$$

The health monitoring algorithm proposed in the present work employs CART for failure classification based on residuals from faulty and healthy data in steady state condition. The first implementation step was choosing the failure modes, variables used for model seeded fault, list of sensors and operational data snapshots for analysis.

Eight types of faults were considered:

1. Increase in shaft torque extraction;
2. Increase in bleed;
3. Reduction in compressor efficiency;
4. Reduction in turbine efficiency;
5. Speed sensor bias;
6. EGT sensor bias;
7. Reduction in combustor efficiency;
8. Decrease in fuel flow.

The measured variables were assumed to be fuel flow, EGT and bleed pressure. Healthy data and faulty data were generated using the mathematical model of an APU derived from thermodynamic principles. The residuals used as inputs for CART were calculated as shown in Figure 4.

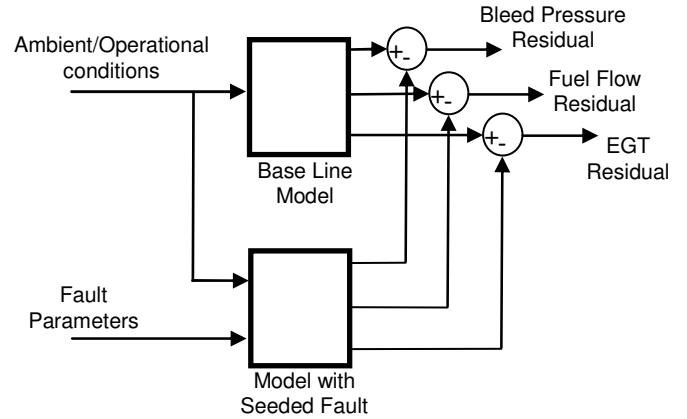


Figure 4: Calculation of the residuals employed in the proposed HM methodology.

4. MODEL DESCRIPTION

The thermodynamic model used in this work is represented schematically in Figure 5.

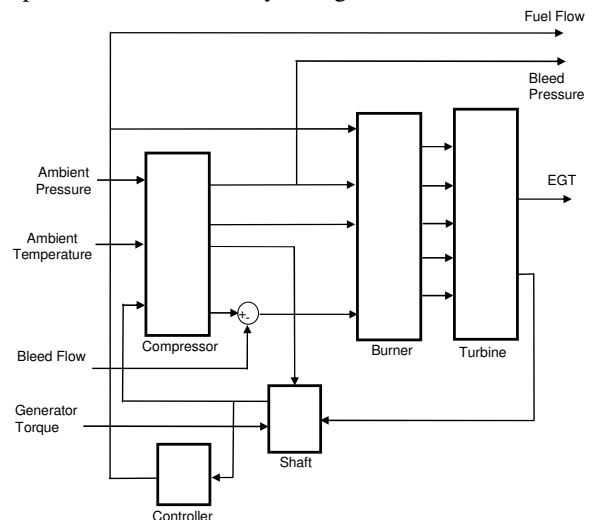


Figure 5: Block diagram of the APU model.

The model contains four inputs and three outputs, which represent the APU sensors. Three of the blocks model the thermodynamic behavior: the compressor, burner and turbine.

The inputs to the compressor block consist of ambient pressure and temperature, as well as shaft speed. The outputs are compressor torque, air flow, compressor pressure and temperature. The compressor behavior is based on a map which relates pressure ratio, airflow, speed and temperature as illustrated in Figure 6.

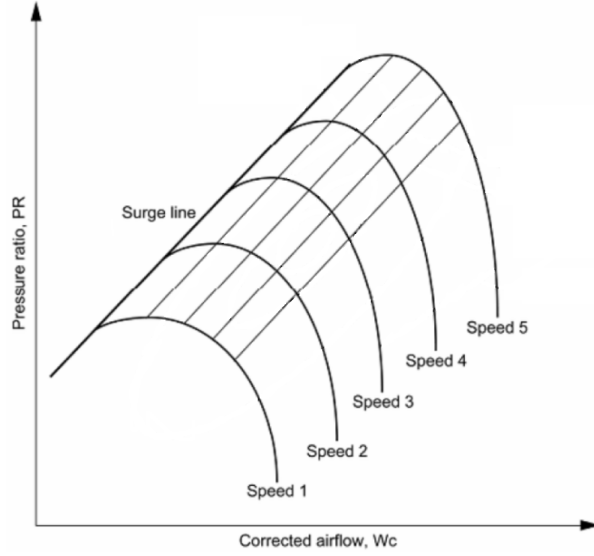


Figure 6: Compressor map.

The pressure ratio (PR) is defined as:

$$PR = \frac{P_{out}}{P_{in}} \quad (4)$$

where P_{out} is the outlet pressure and P_{in} is the inlet pressure.

The corrected air flow is defined as:

$$W_c = W \sqrt{\frac{\theta}{\delta}} \quad (5)$$

where W is the absolute air flow and δ and θ are given by:

$$\delta = \frac{P_{in}}{P_{ref}} \quad (6)$$

$$\theta = \frac{T_{in}}{T_{ref}} \quad (7)$$

where P_{ref} is the standard day pressure, T_{ref} is the standard day temperature and T_{in} is the inlet temperature.

The corrected speed (N_c) is defined as:

$$N_c = \frac{N}{\sqrt{\theta}} \quad (8)$$

where N is the absolute shaft speed.

The inputs to the burner block are air flow, compressor pressure and temperature, as well as fuel flow. The outputs are burner pressure and temperature,

air flow and Fuel Air Ratio (FAR). The input-output characteristic of this component is represented as:

$$h_{out} = \frac{W_{air} \cdot h_{in} + LHV \cdot W_{fuel}}{W_{air} + W_{fuel}} \quad (9)$$

where h_{in} and h_{out} are respectively the burner inlet and outlet enthalpies, W_{fuel} is fuel flow, W_{air} is the burner exhaust air flow and LHV is the fuel heating value.

The turbine is represented by a map as illustrated in Figure 7.

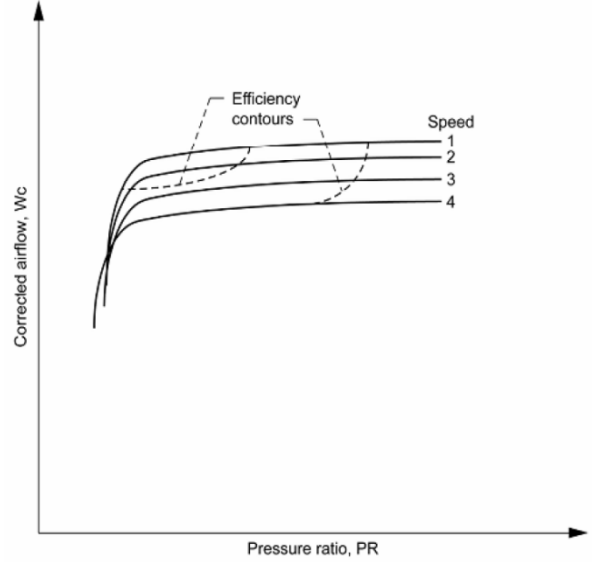


Figure 7: Turbine map.

Apart from the thermodynamic blocks, two other components are modeled in Fig. 5. The first one is the controller, which reproduces one of the main features of the FADEC, namely the control of shaft speed by manipulation of fuel flow. Here, a PID controller is used. The other block represents the energy balance of the shaft speed, which can be described by the following equation:

$$\sum \tau = I \cdot \dot{N} \Rightarrow N = \int \frac{\sum \tau}{I} \quad (10)$$

where I is the moment of inertia and $\sum \tau$ is the sum of compressor, turbine and generator torques. The latter represents the torque extracted from the APU to supply electrical components such as electrical pumps and lights.

5. RESULTS

For the construction of the initial classification tree, a set of 180 data vectors was generated. This dataset generation consisted of 20 simulations of APUs

startups for each of the failure modes and other 20 simulations for the APU operating without faults. Different loads, simulating pumps, engines and air cycle machine were used.

The data vectors collected comprised residual values of EGT, fuel flow and bleed pressure.

The resulting classification tree is presented in Figure 8.

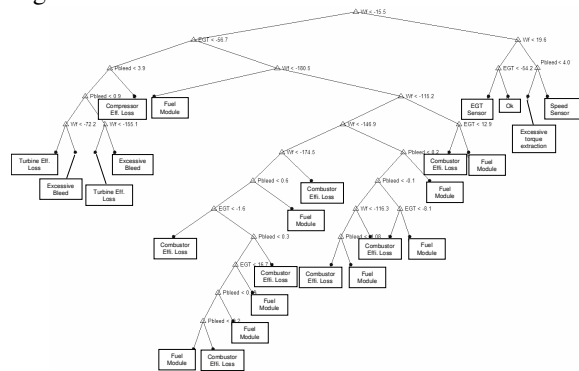


Figure 8: Initial classification tree.

Analyzing the initial classification tree it is possible to notice the great number of nodes possibly resulting in overfitting of the training data. This problem was solved by pruning some nodes of the tree using expert knowledge provided by an APU system specialist. Some of the failure modes were very similar and it would be a better strategy to group them into a reduced number of nodes. The resulting tree is presented in Figure 9.

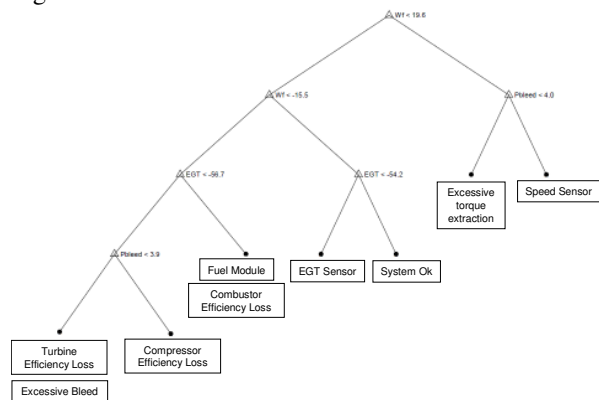


Figure 9: Classification tree resulting from the pruning procedure.

It is possible to observe in Figure 9 that some of the failure modes were grouped into the same node, indicating that they could not be separated based on the sensors that were used. Although the initial objective was to classify all failure modes, some ambiguities could not be resolved. However, the level of isolation provided by the proposed tree was considered adequate, as it helps to reduce significantly the troubleshoot time on an event of failure.

The proposed method was tested using actual data collected in the field. The dataset consisted of 18 data vectors comprising 6 healthy states and 12 failure events. The data were collected with the engine, the electric pump and the air cycle machine in either on or off state, as shown in Table 1.

Table 1: Field data.

	ACM	Pump	Engine
Healthy	off	off	off
Healthy	off	off	on
Healthy	off	off	on
Healthy	on	on	off
Healthy	on	on	on
Healthy	off	off	on
Excessive Bleed	off	off	off
Excessive Bleed	on	on	off
Excessive Bleed	off	off	off
50% Inlet Blockage	off	off	off
50% Inlet Blockage	on	on	off
50% Inlet Blockage	off	off	off
75% Inlet Blockage	off	off	off
75% Inlet Blockage	on	on	off
75% Inlet Blockage	off	off	off
Fuel Filter Blockage	off	off	off
Fuel Filter Blockage	on	on	off
Fuel Filter Blockage	off	off	off

Although the "Inlet Blockage" was neither modeled nor used for the training of the classification tree, the effects due to this type of failure are very similar to those of an "EGT sensor bias". Therefore, it is expected that these particular conditions should be classified as "EGT sensor bias" failures.

The results for the classification are presented in Table 2.

Table 2: Classification results for the field data.

Ground Truth	Classification
Healthy	No failure
Healthy	No failure
Healthy	No failure
Healthy	No failure
Healthy	No failure
Healthy	No failure
Excessive Bleed	Compressor Eff. loss
Excessive Bleed	Excessive Bleed
Excessive Bleed	Excessive Bleed
50% Inlet Blockage	EGT Sensor
50% Inlet Blockage	EGT Sensor
50% Inlet Blockage	No failure
75% Inlet Blockage	Excessive Bleed
75% Inlet Blockage	Excessive Bleed
75% Inlet Blockage	EGT Sensor
Fuel Filter Blockage	Fuel Module
Fuel Filter Blockage	Fuel Module
Fuel Filter Blockage	No failure

Observing the results presented in Table 2, one can notice that the classification algorithm was able to classify correctly all healthy states, that is, no healthy system was classified as faulty. On the other hand, the algorithm was not able to classify correctly all failure events.

In order to identify possible improvements on the method proposed, all classification errors were analyzed observing the raw data.

Looking at the data from “50% inlet blockage” and “fuel filter blockage” faults, both classified as “no failures”, no significant difference in the parameters were observed, as compared to a situation without fault. The conclusion is that the “Inlet Blockage” and the “Fuel Filter Blockage” were not sufficient to cause any modification on the monitored variables. One factor that could contribute to these errors is the difference in the behavior of the APU in hot and cold starts. This effect was not modeled in the present work.

Lack of precise calibration and modeling data, specifically compressor and turbine maps and lack of precise bleed flow test data lead to errors on “Excessive Bleed” being classified as “compressor efficiency loss” and the “75% Inlet Blockage” classified as “Excessive Bleed”.

6. CONCLUSION

This paper presented an APU health monitoring method using a Dynamic Model and a Classification and Regression Tree (CART). The CART was used to classify APU failure modes based on measurements of the exhaust gas temperature, the bleed pressure and the fuel flow.

After designing the CART, the method was tested using real APU data. Although the method was not capable to classify correctly all failure modes, it showed promising results.

ACKNOWLEDGMENT

The authors acknowledge the support of CNPq (research fellowships) and FAPESP (grant 06/58850-6).

REFERENCES

- Chen, W. (1991). *Nonlinear Analysis of Electronic Prognostics*. Ph. D. Thesis. The Technical University of Napoli.
- Jones S. M. (2007). An Introduction to Thermodynamic Performance Analysis of Aircraft Gas Turbine Engine Cycles Using the Numerical Propulsion System Simulation Code. NASA Glenn Research Center, Cleveland, Ohio NASA/TM—2007-214690.
- Li Y. G. (2003) A gas Turbine Approach with Transient Measurements *in Proceedings of the*

Institution of Mechanical Engineers, Part A: Journal of Power and Energy, Volume 217, Number 2, pp. 169-177

- Sabyasachi B., Farner S., Schimert J. and Wineland (2008) A Data Driven Method for Predicting Engine Wear from Exhaust Gas Temperature in Proceedings of International Conference on Prognostics and Health Management .
- SAE. (2006). E-32 AIR5317 - A guide to APU Health Management
- Saravanamuttoo H.I.H. and Maclsaac B.D. (1982). Thermodynamic Models for Pipeline Gas Turbine Diagnostics, *Journal of Engineering for Power*, Vol. 105, pp.875-884, October 1982
- Timofeev R. (2004), *Classification and Regression Trees (CART) Theory and Applications*, Master Thesis - CASE - Center of Applied Statistics and Economics Humboldt University, Berlin.
- Urban L. A.(1967). *Gas Turbine Engine Parameter Interrelationship*, HSD UTC, Windsor Locks, Ct., 1st edition.
- Vieira, F. M. and Bizarria C. O. (2009).Health Monitoring using Support Vector Classification on an Auxiliary Power Unit, *in Proceedings of IEEE Aerospace Conference*, Big Sky, MO.

Wlamir Olivares Loesch Vianna holds a bachelor’s degree on Mechanical Engineering (2005) from Universidade de São Paulo (USP), Brazil, and Master Degree on Aeronautical Engineering (2007) from Instituto Tecnológico de Aeronáutica (ITA), Brazil. He is with Empresa Brasileira de Aeronáutica S.A (EMBRAER), São José dos Campos, SP, Brazil, since 2007. He works as a Development Engineer of a R&T group at EMBRAER focused on PHM technology applications in aeronautical systems

João Paulo Pordeus Gomes holds a bachelor’s degree on Electrical Engineering (2004) from Universidade Federal do Ceará (UFC), Brazil, and Master Degree on Aeronautical Engineering (2006) from Instituto Tecnológico de Aeronáutica (ITA), Brazil. He is currently pursuing his Ph.D. from ITA. He is with Empresa Brasileira de Aeronáutica S.A (EMBRAER), São José dos Campos, SP, Brazil, since 2006. He works as a Development Engineer of a R&T group at EMBRAER focused on PHM technology applications in aeronautical systems

Roberto Kawakami Harrop Galvão is an Associate Professor of Systems and Control at the Electronic Engineering Department of ITA. He holds a bachelor’s degree in Electronic Engineering (Summa cum Laude, 1995) from Instituto Tecnológico de Aeronáutica (ITA), Brazil. He also obtained the master’s (1997) and

doctorate (1999) degrees in Systems and Control from the same institution. He is a Senior Member of the IEEE and an Associate Member of the Brazilian Academy of Sciences. He has published more than 150 papers in peer-reviewed journals and conference proceedings. His main areas of interest are fault diagnosis and prognosis, wavelet theory and applications, and model predictive control.

Takashi Yoneyama is a Professor of Control Theory with the Electronic Engineering Department of ITA. He received the bachelor's degree in electronic engineering from Instituto Tecnológico de Aeronáutica (ITA), Brazil, the M.D. degree in medicine from Universidade de Taubaté, Brazil, and the Ph.D. degree in electrical engineering from the University of London, U.K. (1983). He has more than 250 published papers, has written four books, and has supervised more than 50 theses. His research is concerned mainly with stochastic optimal control theory. He served as the President of the Brazilian Automatics Society in the period 2004-2006.

Jackson Paul Matsuura is an Associate Professor of Systems and Control at the Electronic Engineering Department of Instituto Tecnológico de Aeronáutica (ITA). He holds a bachelor's degree (1995) in Computer Engineering from ITA, Brazil. He also obtained the master's (2003) and doctorate (2006) degrees in Systems and Control from the same institution. His main areas of interest are fault tolerant control, robotics and augmented reality. He won the RoboChamps World Finals in 2008.

Identification of Correlated Damage Parameters under Noise and Bias Using Bayesian Inference

Dawn An¹, Joo-Ho Choi², and Nam H. Kim³

^{1,2}*Department of Aerospace & Mechanical Engineering, Korea Aerospace University, Goyang-City, Gyeonggi-do, 412-791, Korea*

*skal34@nate.com
jhchoi@kau.ac.kr*

³*Dept. of Mechanical & Aerospace Engineering, University of Florida, Gainesville, FL, 32611, U.S.A.
nkim@ufl.edu*

ABSTRACT

This paper presents statistical model parameter identification using Bayesian inference when parameters are correlated and observed data have noise and bias. The method is explained using the Paris model that describes crack growth in a plate under mode I loading. It is assumed the observed data are obtained through structural health monitoring systems, which may have random noise and deterministic bias. It was found that strong correlation exists (a) between two model parameters of the Paris model, and (b) between initially measured crack size and bias. As the level of noise increases, the Bayesian inference was not able to identify the correlated parameters. However, the remaining useful life was predicted accurately because the identification errors in correlated parameters were compensated by each other.

1. INTRODUCTION

Condition-based maintenance (CBM) provides a cost effective maintenance strategy by providing an accurate quantification of degradation and damage at an early stage without intrusive and time consuming inspections (Giurgiutiu, 2008). Structural health monitoring (SHM) has the potential to facilitate CBM. Most proposed SHM systems utilize on-board sensors/actuators to detect damage, to find the location of damage, and to estimate the significance of damage (Mohanty et al., 2011). Since the SHM systems can assess damage frequently, they can also be used to predict the future behavior of the system, which is critically important for maintenance scheduling and fleet management. SHM systems can have a significant impact

on increasing safety by allowing predictions of the structure's health status and remaining useful life (RUL), which is called prognostics.

In general, prognostics methods can be categorized into data-driven (Schwabacher, 2005), physical model-based (Luo et al., 2008), and hybrid (Yan and Lee, 2007) approaches, based on the usage of information. The data-driven method uses information from collected data to predict future status of the system without using any particular physical model. It includes least-square regression and Gaussian process regression, etc. The physical model-based method assumes that a physics model that describes the behavior of the system is available. This method combines the physics model with measured data to identify model parameters and predicts future behavior. Modeling the physical behavior can be accomplished at different levels, for example, micro- and macro-levels. Crack growth model (Paris and Erdogan, 1963) or fatigue life model (Yu and Harris, 2001) are often used for macro-level damage, and first principle models (Jaw et al., 1999) are used for micro-level damage. The hybrid method combines the abovementioned two methods, and includes particle filters (Orchard and Vachtsevanos, 2007; Orchard et al., 2008; Zio and Peloni, 2011) and Bayesian techniques (Sheppard et al., 2005; Saha and Goebel, 2008; Sankararaman et al., 2010; Ling et al., 2010). Since the data-driven method identifies abnormality based on the trend of data, it is powerful in predicting near-future behaviors, while the physical model-based method has advantages in predicting long-term behaviors of the system. It is noted that in the physical model-based method for fatigue applications, the history of load is required in addition to the measured crack data.

In this paper, a physics-based model for structural degradation due to damage is applied for prognostics since damage grows slowly and the physics governing its behavior is relatively well-known. The main purpose of

An et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

prognostics is to identify and repair those damages that threaten the system safety (condition-based maintenance) and to predict an appropriate maintenance schedule. Paris-family models are commonly used in describing the growth of cracks in aircraft panels under fatigue loading (Paris et al., 2008). In this paper, the original Paris model (Paris and Erdogan, 1963) is used because it has the least number of parameters. The main purpose of the paper is to present the usage of Bayesian inference in identifying model parameters and predicting the RUL—the remaining cycles before maintenance. The study focuses on crack growth in a fuselage panel under repeated pressurization loading, which can be considered regular loading cycles. In this type of application, the uncertainty in applied loading is small compared to other uncertainties. The improved accuracy in these model parameters allows more accurate prediction of the RUL of the monitored structural component.

Identifying the model parameters and predicting damage growth, however, is not a simple task due to the noise and bias of data from SHM systems and the correlation between parameters, which is prevalent in practical problems. The noise comes from variability of random environments, while the bias comes from systematic departures of measurement data, such as calibration error. However, there are not many research results for identifying model parameters under noise and bias, without mentioning correlated parameters (Orchard et al., 2008; Bechhofer, 2008).

The main objective of this paper is to demonstrate how Bayesian inference can be used to identify model parameters and to predict RUL using them, especially when the model parameters are correlated. In order to find the effects of noise and bias on the identified parameters, numerical studies utilize synthetic data; i.e., the measurement data are produced from the assumed model of noise and bias. The key interest is how the Bayesian inference identifies the correlated parameters under noise and bias in data.

The paper is organized as follows. In Section 2, a simple damage growth based on Paris model is presented in addition to the uncertainty model of noise and bias. In Section 3, parameter identification and RUL prediction using Bayesian inference and MCMC simulation method (Andrieu et al., 2003) is presented with different levels of noise and bias. Conclusions are presented in Section 4.

2. DAMAGE GROWTH AND MEASUREMENT UNCERTAINTY MODELS

2.1 Damage growth model

In this paper, a simple damage growth model is used to demonstrate the main idea of characterizing damage growth parameters. Although some experimental data on fatigue damage growth are available in the literature (Virkler et al., 1979), they are not measured using SHM systems.

Therefore, the level of noise and bias is much smaller than the actual data that will be available in SHM systems. In this paper, synthetic damage growth data are used in order to perform statistical study on the effect of various levels of noise and bias. It is assumed that a through-the-thickness center crack exists in an infinite plate under the mode I loading condition. In aircraft structure, this corresponds to a fuselage panel under repeated pressurization loadings (see Figure 1). In this approximation, the effect of finite plate size and the curvature of the plate are ignored. When the stress range due to the pressure differential is $\Delta\sigma$, the rate of damage growth can be written using the Paris model (Paris and Erdogan, 1963) as

$$\frac{da}{dN} = C(\Delta K)^m, \quad \Delta K = \Delta\sigma\sqrt{\pi a} \quad (1)$$

where a is the half crack size, N is the number of cycles, which is close to real time when the cycle is very short, ΔK is the range of stress intensity factor, and other parameters are shown in Table 1 for 7075-T651 aluminum alloy. Although the number of cycles, N , is an integer, it is treated as a real number in this model. The above model has two damage growth parameters, C and m , which are estimated to predict damage propagation and RUL. In Table 1, these two parameters are assumed to be uniformly distributed. The lower- and upper-bounds of these parameters were obtained from the scatter of experimental data (Newman et al., 1999). They can be considered as the damage growth parameters of generic 7075-T651 material. In general, it is well-known that the two Paris parameters are strongly correlated (Sinclair and Pierie, 1990, but it is assumed initially that they are uncorrelated because there is no prior knowledge on the level of correlation. Using measured data of crack sizes, the Bayesian inference will show the correlation structure between these two parameters. Since the scatter is so wide, the prediction of RUL using these distributions of parameters is meaningless. The specific panel being monitored using SHM systems may have much narrower distributions of the parameters, or even deterministic values.

The half crack size a_i after N_i cycles (flights) of fatigue loading can be obtained by integrating Eq. (1) and solving for a_i as

$$a_i = \left[N_i C \left(1 - \frac{m}{2} \right) (\Delta\sigma\sqrt{\pi})^m + a_0^{1-\frac{m}{2}} \right]^{\frac{2}{2-m}} \quad (2)$$

where a_0 is the initial half crack size. In SHM, the initial crack size does not have to be the micro-crack in the panel before applying any fatigue loading. This can be the crack size that is detected by SHM systems the first time. In such a case, N_i should be interpreted as the number of cycles

since detected. It is assumed that the panel fails when a_i reaches a critical half crack size, a_c . Here we assume that this critical crack size is when the stress intensity factor exceeds the fracture toughness K_{IC} . This leads to the following expression for the critical crack size:

$$a_c = \left(\frac{K_{IC}}{\Delta\sigma\sqrt{\pi}} \right)^2 \quad (3)$$

Even if the above crack growth model is the simplest form, it requires identifying various parameters. First, the damage growth parameters, C and m , need to be identified, which can be estimated from specimen-level fatigue tests⁰. However, due to material variability, these parameters show different values for different batches of panels. In addition, the initial crack size, a_0 , needs to be found. Liu and Mahadevan (2009) used an equivalent initial flaw size, but it is still challenging to find the initial crack size. In addition, the fracture toughness, K_{IC} , also shows randomness due to variability in manufacturing.

2.2 Measurement uncertainty model

In SHM-based inspection, the sensors installed on the panel are used to detect the location and size of damage. Even if the on-line inspection can be performed continuously, it would not be much different from on-ground inspection because the structural damage will not grow quickly. In addition, the on-ground inspection will have much smaller levels of noise than on-line. The on-ground inspection may provide a significant weight advantage because only sensors, not measurement equipment, are on-board. Our preliminary study showed that there is no need to inspect at every flight because the damage growth at each flight is extremely small.

A crack in the fuselage panel grows according to the applied loading, pressurizations in this case. Then the structural health monitoring (SHM) systems detect the crack. In general, the SHM system cannot detect a crack when it is small. Many SHM systems can detect a crack between the sizes of 5~10mm (Jerome and Kenneth, 2006). Therefore, the necessity of identifying the initial crack size becomes unimportant by setting a_0 to be the initially detected crack size. However, a_0 may still include noise and bias from the measurement. In addition, the fracture toughness, K_{IC} , is also unimportant because airliners may want to send the airplane for maintenance before the crack becomes critical.

The main objective of this paper is to show that the measured data can be used to identify crack growth parameters, and then, to predict the future behavior of the cracks. Since no airplanes are equipped with SHM systems

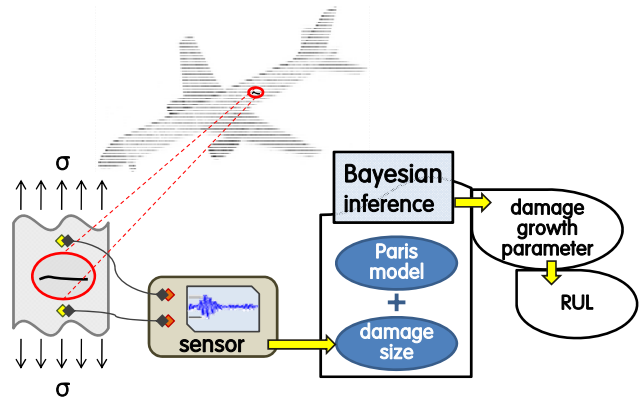


Figure 1. Through-the-thickness crack in a fuselage panel

yet, we simulate the measured crack sizes from SHM. In general, the measured damage includes the effect of bias and noise of the sensor measurement. The former is deterministic and represents a calibration error, while the latter is random and represents a noise in the measurement environment. The synthetic measurement data are useful for parameter study, that is, the different noise and bias levels show how the identification process is affected. In this context, bias is considered as two different levels, $\pm 2\text{mm}$, and noise is uniformly distributed between $-u$ mm and $+u$ mm. Four different levels of u are considered: 0mm, 0.1mm, 1mm, 5mm. The different levels of noise represent the quality of SHM systems.

The synthetic measurement data are generated by (a) assuming that the true parameters, m_{true} and C_{true} , and the initial half crack size, a_0 , are known; (b) calculating the true crack sizes according to Eq. (2) for a given N_i and $\Delta\sigma$; and (c) adding a deterministic bias and random noise to the true crack size data including the initial crack size. Once the synthetic data are obtained, the true values of crack sizes as well as the true values of parameters are not used in the prognostics process. In this paper, the following true values of parameters are used for all numerical examples: $m_{\text{true}} = 3.8$, $C_{\text{true}} = 1.5 \times 10^{-10}$, and $a_0 = 10\text{mm}$.

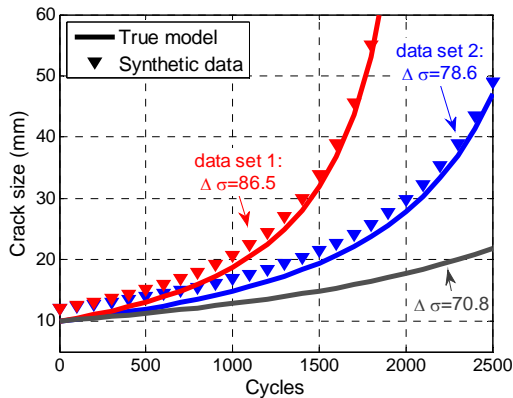
Table 1 shows three different levels of loading; the first two ($\Delta\sigma = 86.5$ and 78.6MPa) are used for estimating model

Property	Nominal stress $\Delta\sigma$ (MPa)	Fracture toughness K_{IC} (MPa $\sqrt{\text{m}}$)	Damage parameter m	Damage parameter $\log(C)$
Distribution type	case 1: 86.5 case 2: 78.6 case 3: 70.8	Deterministic 30	Uniform (3.3, 4.3)	Uniform ($\log(5E-11)$, $\log(5E-10)$)

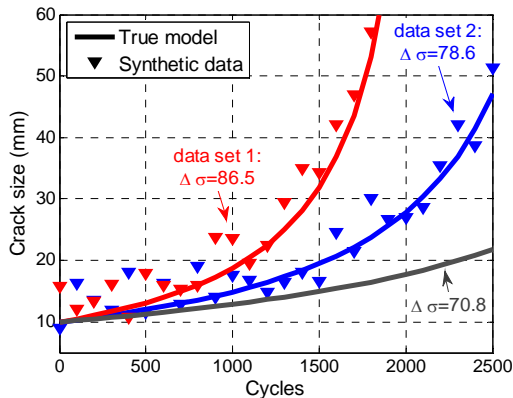
Table 1 Loading and fracture parameters of 7075-T651 Aluminum alloy

parameters, while the last ($\Delta\sigma = 70.8$) is used for validation purposes. The reason for using two sets of data for estimating damage growth parameters is to utilize more data having damage propagation information at an early stage. Theoretically, the true values of parameters can be identified using a single set of data because the Paris model is a nonlinear function of parameters. However, random noise can make the identification process slow, especially when parameters are correlated; i.e., many different combinations of correlated parameters can achieve the same crack size. This property delays the convergence of Bayesian process such that meaningful parameters can only be obtained toward the end of RUL. Based on preliminary study, two sets of data at different loadings can help the Bayesian process converge quickly.

Figure 2 shows the true crack growth curves for three different levels of loading (solid curves) and synthetic measurement data a_i^{meas} (triangles) that generated in two levels of loading including noise and bias. It is noted that the positive bias shifts the data above the true crack growth. On the other hand, the noises are randomly distributed between measurement cycles. It is assumed that the measurements are performed at every 100 cycles. Let there



(a) bias = +2mm and noise = 0mm



(b) bias = +2mm and noise = 5mm

Figure 2. Crack growth of three different loading conditions and two sets of synthetic data

be n measurement data. Then the measured crack sizes and corresponding cycles are represented by

$$\mathbf{a}^{meas} = \{a_0^{meas}, a_1^{meas}, a_2^{meas}, \dots, a_n^{meas}\} \quad (4)$$

$$\mathbf{N} = \{N_0 = 0, N_1 = 100, N_2 = 200, \dots, N_n\}$$

It is assumed that after N_n , the crack size becomes larger than the threshold and the crack is repaired.

3. BAYESIAN INFERENCE FOR CHARACTERIZATION OF DAMAGE PROPERTIES

3.1 Damage growth parameters estimation

Once the synthetic data (damage sizes vs. cycles) are generated, they can be used to identify unknown damage growth parameters. As mentioned before, m , C , and a_0 can be considered as unknown damage growth parameters. In addition, the bias and noise are used in generating the synthetic data are also unknown because they are only assumed to be known in generating crack size data. In the case of noise, the standard deviation, σ , of the noise is considered as an unknown parameter. The identification of σ will be important as the likelihood function depends on it. Therefore, the objective is to identify (or, improve) these five parameters using the measured crack size data. The vector of unknown parameters is defined by $\mathbf{y} = \{m, C, a_0, b, \sigma\}$.

Parameter identification can be done in various ways. The least-squares method is a traditional way of identifying deterministic parameters. For crack propagation, Coppe et al. (2010) used the least-square method to identify unknown damage growth parameter along with bias. However, in the least-squares method, it is non-trivial to estimate the uncertainty in the identified parameters. In this paper, Bayesian inference is used to identify the unknown parameters as well as the level of noise and bias. Coppe et al. (2010) used Bayesian inference in identifying damage growth parameter, C or m . They used the grid method to calculate the posterior distribution of one variable and discussed that updating multi-dimensional variables can be computationally expensive. The grid method computes the values of PDF at a grid of points after identifying the effective range, and calculates the value of the posterior distribution at each grid point. This method, however, has several drawbacks such as the difficulty in finding correct location and scale of the grid points, spacing of the grid, and so on. In addition, it becomes computationally expensive when the number of updating parameters increases. Markov Chain Monte Carlo (MCMC) simulation is a computationally efficient alternative to obtain the PDF by generating a chain of samples (Andrieu et al., 2003).

In Baye's theorem (Bayes, 1763), the knowledge of a system can be improved with additional observation of the system. More specifically, the joint probability density function (PDF) of \mathbf{y} will be improved using the measured crack sizes \mathbf{a}^{meas} . The joint posterior PDF is obtained by multiplying the prior PDF with the likelihood as

$$p_{\mathbf{Y}}(\mathbf{y} | \mathbf{a}^{\text{meas}}) = \frac{1}{K} p_A(\mathbf{a}^{\text{meas}} | \mathbf{Y} = \mathbf{y}) p_{\mathbf{Y}}(\mathbf{y}) \quad (5)$$

where $p_{\mathbf{Y}}(\mathbf{y})$ is the prior PDF of parameters, $p_A(\mathbf{a}^{\text{meas}} | \mathbf{Y} = \mathbf{y})$ is the likelihood or the PDF values of crack size at \mathbf{a}^{meas} given parameter value of \mathbf{y} , and K is a normalizing constant. It is noted that the likelihood is constructed using n measured crack size data. For prior distribution, the uniform distributions are used for the damage growth parameters, m and C , as described in Table 1. For other parameters, no prior distribution is used; i.e., non-informative. The likelihood is the probability of obtaining the observed crack sizes \mathbf{a}^{meas} given values of parameters. For the likelihood, it is assumed to be a normal distribution for given parameters:

$$p_A(\mathbf{a}^{\text{meas}} | \mathbf{Y} = \mathbf{y}) \propto \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left[-\frac{1}{2} \sum_{i=1}^n \frac{(a_i^{\text{meas}} - a_i(\mathbf{y}))^2}{\sigma^2} \right] \quad (6)$$

where

$$a_i(\mathbf{y}) = \left[N_i C \left(1 - \frac{m}{2} \right) \left(\Delta \sigma \sqrt{\pi} \right)^m + a_0^{1-\frac{m}{2}} \right]^{\frac{2}{2-m}} + b \quad (7)$$

is the crack size from the Paris model and a_i^{meas} is the measurement crack size at cycle N_i . In general, it is possible that the normal distribution in Eq. (6) may have a negative crack size, which is physically impossible; therefore, the normal distribution is truncated at zero.

A primitive way of computing the posterior PDF is to evaluate Eq. (5) at a grid of points after identifying the effective range. This method, however, has several drawbacks such as the difficulty in finding correct location and scale of the grid points, the spacing of the grid, and so on. Especially when a multi-variable joint PDF is required, which is the case in this paper, the computational cost is proportional to M^5 , where M is the number of grids in one-dimension. On the other hand, the MCMC simulation can be an effective solution as it is less sensitive to the number of variables (Andrieu et al., 2003). Using the expression of posterior PDF in Eq. (5), the samples of parameters are drawn by using MCMC simulation method.

The Metropolis-Hastings (M-H) algorithm is a typical method of MCMC and used in this paper.

3.2 The effect of correlation between parameters

Since the original data of crack sizes are generated from the assumed true values of parameters, the objective of Bayesian inference is to make the posterior joint PDF to converge to the true values. Therefore, it is expected that the PDF becomes narrower as n increases; i.e., more data are used. This process seems straightforward, but preliminary study shows that the posterior joint PDF may converge to values different from the true ones. It is found that this phenomenon is related to the correlation between parameters. For example, let the initially detected crack size be a_0^{meas} and let the measurement environment have no noise. This measured size is the outcome of the initial crack size and bias:

$$a_0^{\text{meas}} = a_0 + b \quad (8)$$

Therefore, there exist infinite possible combinations of a_0 and b to obtain the measured crack size. It is generally infeasible to identify the initial crack size and bias with a single measurement when the measured data is linearly dependent on multiple parameters. It was also well known that the two Paris model parameters, m and C , are strongly correlated (Carpinteri and Paggi, 2007). This can be viewed from the crack growth rate curve, as illustrated in Figure 3. In this graph, the parameter m is the slope of the curve, while C corresponds to the y-intercept at $\Delta K = 1$. If a specific value of crack growth rate da/dN is observed, this can be achieved by different combinations of these two parameters. However, in the case of Paris model parameters, it is feasible to identify them because the stress intensity factor gradually increases as the crack grows. However, the embedded noise can make it difficult to identify the two model parameters because the crack growth

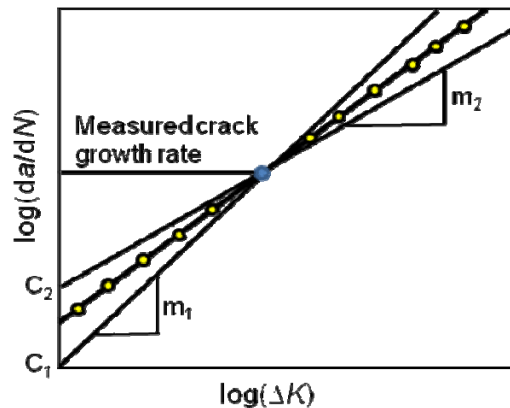


Figure 3. Illustration of showing the same crack growth rate with different combinations of parameters

rate may not be consistent with the noisy data. In addition, this can slow down the convergence in the posterior distribution because when the crack is small, there is no significant crack growth rate. The effect of noise becomes relatively diminished as the crack growth rate increases, which occurs toward the end of life.

In order to handle the abovementioned difficulty in identifying correlated parameters, the bias is removed from the Bayesian identification process using Eq. (8), assuming that they are perfectly correlated. Once the posterior PDF of a_0 are obtained, Eq. (8) is used to calculate the posterior PDF of bias. The two Paris model parameters are kept because they can be identified as the crack grows.

Figure 4 shows the posterior PDFs for the case of true bias of 2mm (a) when $n = 13$ ($N_{13} = 1,200$ cycles) and (b) when $n = 17$ ($N_{17} = 1,600$ cycles). The posterior joint PDFs are plotted separately by three groups for the plotting purpose. In this case, it is assumed that there is no noise in the crack size data. The true values of parameters are marked using a star symbol. Similar results were also obtained in the case with bias = -2mm. Firstly, it is clear that the two Paris model parameters are strongly correlated. The same is true

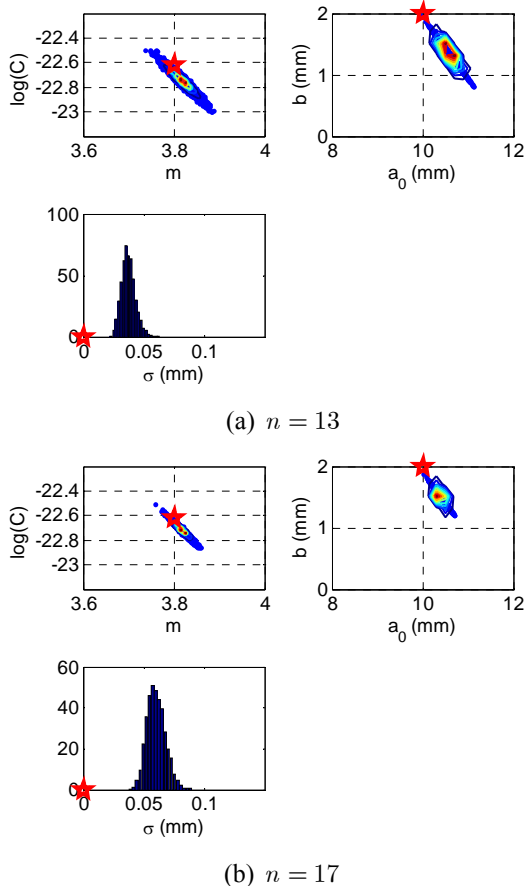


Figure 4. Posterior distributions of parameters with zero noise and true bias of 2mm

num. of data	parameter	true value	b=+2mm		b=-2mm	
			median	error (%)	median	error (%)
n = 13	m	3.8	3.82	0.49	3.78	0.40
	log(C)	-22.6	-22.8	0.57	-22.5	0.50
	a_0	10	10.6	5.67	9.50	4.96
	b	± 2	1.37	31.7	-1.44	28.0
n = 15	m	3.8	3.81	0.32	3.78	0.40
	log(C)	-22.6	-22.7	0.37	-22.5	0.48
	a_0	10	10.4	4.00	9.51	4.94
	b	± 2	1.53	23.6	-1.41	29.5
n = 17	m	3.8	3.82	0.47	3.78	0.55
	log(C)	-22.6	-22.7	0.44	-22.5	0.55
	a_0	10	10.4	3.84	9.49	5.11
	b	± 2	1.52	24.2	-1.35	32.7

Table 2 The median of identified parameters and the errors with the true values

for the initial crack size and bias—in fact the PDF of bias is calculated from that of initial crack size and Eq. (8). Secondly, it can be observed that the PDFs at $n = 17$ is narrower than that of $n = 13$, although the PDFs at $n = 13$ is quite narrow compared to the prior distribution. Lastly, the identified results look different from the true values due to the scale, but the errors between the true values and the median of identified results are at a maximum of around 5% except for bias. The error in bias looks large, but that is because the true value of bias is small. The error in bias is about 0.5mm. The same magnitude of error exists for the initial crack size due to the perfect correlation between them. Table 2 lists all six cases considered in this paper, and all of them show a similar level of errors. It is noted that the identified standard deviation of noise, σ , does not converge to its true value of zero. This occurred because the original data did not include any noise. Zero noise can cause a problem in the likelihood calculation as the denominator becomes zero in Eq. (6). However, this would not happen in practical cases in which noise always exists.

The next example is to investigate the effect of noise on the posterior PDFs of parameters. The results of identified posterior distributions with different levels of noise were shown in Figure 5 when the true bias is 2mm. Similar results were obtained when bias is -2mm. The black, blue and red colors, respectively, represent noise levels of 0.1mm, 1mm, and 5mm. The median location is denoted by a symbol (a circle for 0.1mm noise, a square for 1mm noise, and a star for 5mm noise). Each vertical line represents a 90% confidence interval (CI) of posterior PDF. The solid horizontal line is the true value of the parameter. In the case of noise level = 0.1mm, all parameters were identified accurately with very narrow CIs. In the case of noise level = 1mm, the initial crack size and bias were identified accurately as the number of data increased, whereas the CIs of two Paris parameters were not reduced. In addition, the

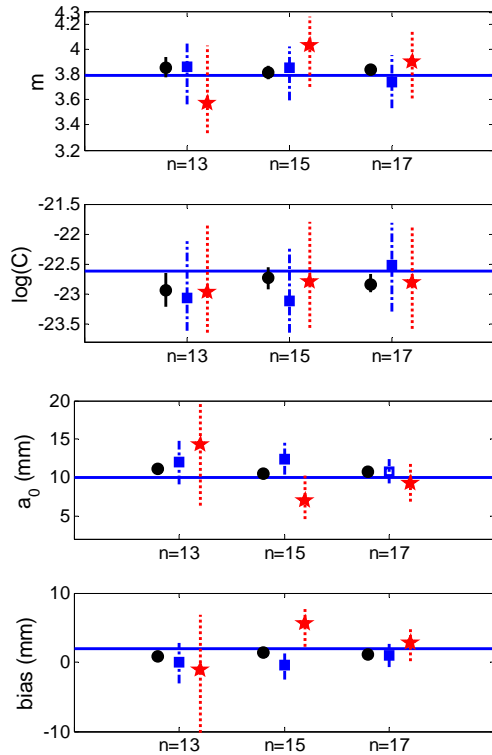
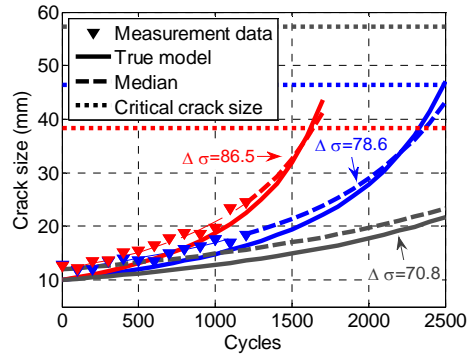


Figure 5. Posterior distributions with three different levels of noise (bias = 2mm)

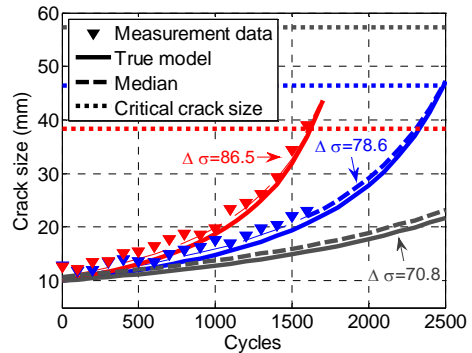
median values were somewhat different from the true parameter values. Worse results were observed in the case of noise level = 5mm. Therefore, it is concluded that the level of noise plays an important role in identifying correlated parameters using Bayesian inference. However, this does not mean that it is not able to predict RUL. Even if these parameters were not accurately identified because of correlation, the predicted RUL was relatively accurate, which will be discussed in detail next subsection.

3.3 Damage propagation and RUL prediction

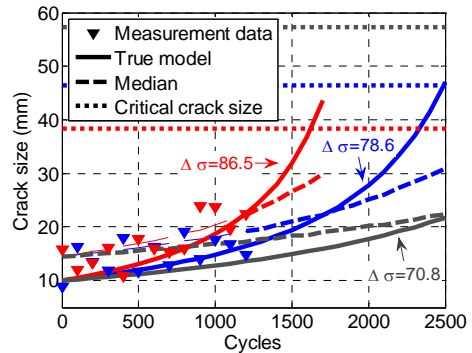
Once the parameters are identified, they can be used to predict the crack growth and estimate RUL. Since the parameters are available in terms of joint PDF, the crack growth and RUL will also be estimated probabilistically. Then the quality of prediction can be evaluated in terms of how close the median is to the true crack growth and how large the prediction interval (PI) is. First, the results of crack growth calculated by Eq. (2) are shown in Figure 6 when the true bias is 2mm. Different colors represent the three different loading conditions. The solid curves are true crack growth, while the dashed curves are medians of predicted crack growth distribution. The results are obtained as a distribution due to the uncertainty of parameters, but the medians of predicted crack growth are only shown in the figures for visibility. In addition, the critical crack sizes with different loadings are using horizontal lines.



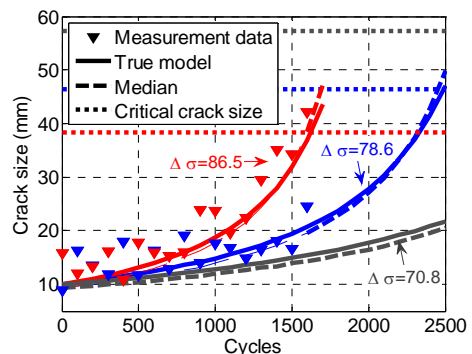
(a) noise = 1mm, $n = 13$



(b) noise = 1mm, $n = 17$



(c) noise = 5mm, $n = 13$



(d) noise = 5mm, $n = 17$

Figure 6. Prediction of crack growth with bias=+2mm

The figures show that the results closely predicted the true crack growth when noise is less than 1mm. Even if the level of noise is 5mm, the results of predicted crack growth become close to the true one as the number of data increases. This means that if there are many data (much information about crack growth), the future crack growth can be predicted accurately even if there is much noise. However, when the level of noise is large, the convergence is slow such that the accurate prediction happened almost at the end of life.

As can be seen from Figure 6, crack growth and RUL can be predicted with reasonable accuracy even though the true values of the parameters are not accurately identified. The reason is that the correlated parameters m and C work

together to predict crack growth in Eq.(2). For example, if m is underestimated, then the Bayesian process overestimates C to compensate for it. In addition, if there is large noise in the data, the distribution of estimated parameters becomes wider, which can cover the risk that comes from the inaccuracy of the identified parameters. Therefore it is possible to safely predict crack growth and RUL.

In order to see the effect of the noise level on the uncertainty of predicted RUL, Figure 7 plots the median and 90% prediction interval (PI) of the RUL and compared them with the true RUL. The RUL can be calculated by solving Eq. (2) for N when the crack size becomes the critical one:

$$N_f = \frac{a_C^{1-m/2} - a_i^{1-m/2}}{C(1 - \frac{m}{2})(\Delta\sigma\sqrt{\pi})^m} \quad (9)$$

The RUL is also expressed as a distribution due to the uncertainty of the parameters. In Figure 7, the solid diagonal lines are the true RULs at different loading conditions ($\Delta\sigma = 86.5, 78.6, 70.8$). The precision and accuracy are fairly good when the noise is less than 1mm, which is consistent with the crack growth results. In the case of a large noise, 5mm, the medians are close to the true RUL, and the wide intervals are gradually reduced as more data are used. That is, the accuracy and precision can be better as more data are used in spite of large noise and bias in data. In the case that there are not as much data as covering large noise, the results also can be used to define the acceptable limits on system noise for useful RUL prediction. Therefore, it is concluded that the RULs are predicted reasonably in spite of noise and bias in data.

4. CONCLUSIONS

In this paper, Bayesian inference and the Markov Chain Monte Carlo (MCMC) method are used for identifying the Paris model parameters that govern the crack growth in an aircraft panel using structural health monitoring (SHM) systems that measure crack sizes with noise and bias. Focuses have been given to the effect of correlated parameters and the effect of noise and bias levels. The correlation between the initial crack size and bias was explicitly imposed using analytical expression, while the correlation between two Paris parameters was identified through the Bayesian inference. It is observed that the correlated parameter identification is sensitive to the level of noise, while predicting the remaining useful life is relatively insensitive to the level of noise. It is found that greater numbers of data are required to narrow the distribution of parameters when the level of noise is large. When parameters are correlated, it is difficult to identify the true values of the parameters, but the correlated parameters work together to predict accurate crack growth and RUL.

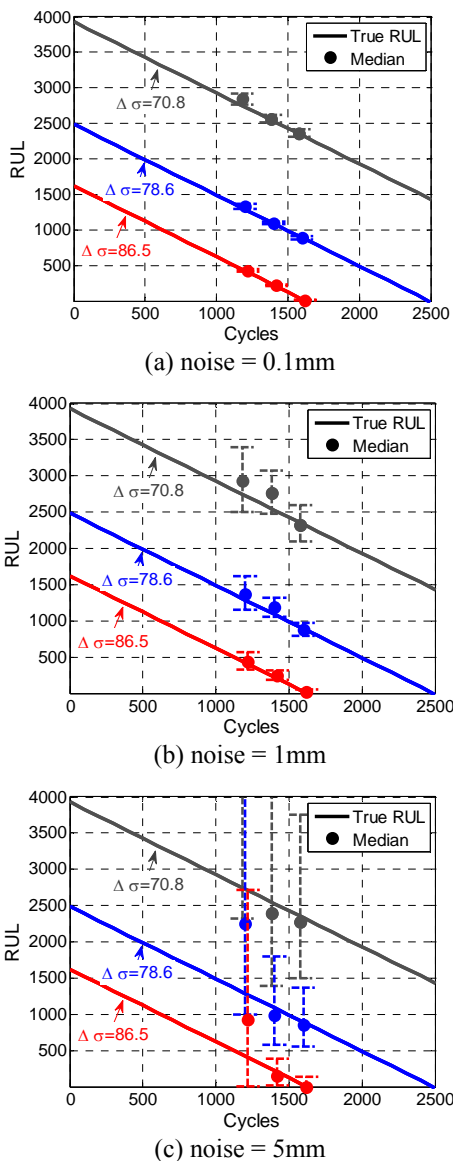


Figure 7. Median and 90% of prediction interval of the predicted RUL (bias = 2mm)

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2009-0081438).

REFERENCES

- Andrieu, C., Freitas, de N., Doucet, A., & Jordan, M. (2003). An introduction to MCMC for Machine Learning. *Machine Learning*, vol. 50(1), pp. 5-43.
- Bayes, T. (1763). An Essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, vol. 53, pp. 370-418.
- Bechhoefer, E. (2008). A Method for Generalized Prognostics of a Component Using Paris Law. *Proceedings of the American Helicopter Society 64th Annual Forum*, Montreal, CA.
- Carpinteri, A., & Paggi, M. (2007). Are the Paris' law parameters dependent on each other?. *Frattura ed Integrità Strutturale*, vol. 2, pp. 10-16.
- Coppe, A., Haftka, R. T., & Kim, N. H. (2010). Least Squares-Filtered Bayesian Updating for Remaining Useful Life Estimation. *12th AIAA Non-Deterministic Approaches Conference*, Orlando, FL.
- Coppe, A., Haftka, R. T., Kim, N. H., & Yuan, F. G. (2010). Uncertainty reduction of damage growth properties using structural health monitoring. *Journal of Aircraft*, in press.
- Giurgiutiu, V. (2008). *Structural Health Monitoring: with Piezoelectric Wafer Active Sensors*, Academic Press (an Imprint of Elsevier), Burlington, MA.
- Jaw, L. C., Inc, S. M., & Tempe, A. Z. (1999). Neural networks for model-based prognostics. *IEEE Aerospace Conference*.
- Jerome P. L., & Kenneth J. L. (2006). A Summary Review of Wireless Sensors and Sensor Networks for Structural Health Monitoring. *The Shock and vibration digest*, vol. 38(2), pp. 91-128.
- Ling, Y., Shantz, C., Sankararaman, S., & Mahadevan, S. (2010). Stochastic Characterization and Update of Fatigue Loading for Mechanical Damage Prognosis. *Annual Conference of the Prognostics and Health Management Society*.
- Liu, Y., & Mahadevan, S. (2009). Probabilistic fatigue life prediction using an equivalent initial flaw size distribution. *International Journal of Fatigue*. vol. 31(3), pp. 476-487.
- Luo, J., Pattipati, K. R., Qiao, L., & Chigusa, S. (2008). Model-based Prognostic Techniques Applied to a Suspension System. *IEEE Transactions on System, Man and Cybernetics*, vol. 38(5), pp. 1156-1168.
- Mohanty, S., Chattopadhyay, A., Peralta, P., & Das, S. (2011). Bayesian Statistic Based Multivariate Gaussian Process Approach for Offline/Online Fatigue Crack Growth Prediction. *Experimental Mechanics*, vol. 51, pp. 833-843.
- Newman Jr, J. C., Phillips, E. P., & Swain, M. H. (1999). Fatigue-life prediction methodology using small-crack theory. *International Journal of Fatigue*, vol. 21, pp. 109-119.
- Orchard, M., & Vachtsevanos, G. (2007). A Particle Filtering Approach for On-Line Failure Prognosis in a Planetary Carrier Plate. *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 7(4), pp. 221-227.
- Orchard, M., Kacprzyński, G., Goebel, K., Saha, B., & Vachtsevanos, G. (2008). Advances in Uncertainty Representation and Management for Particle Filtering Applied to Prognostics. *International Conference on Prognostics and Health Management*, Denver, CO.
- Paris, P.C., & Erdogan, F. (1963). A Critical Analysis of Crack Propagation Laws. *ASME Journal of Basic Engineerin* , vol. 85, pp. 528-534.
- Paris, P. C., Lados D., & Tada, H. (2008). Reflections on identifying the real $\Delta K_{\text{effective}}$ in the threshold region and beyond. *International Journal of fatigue*, vol.75, pp. 299-305.
- Saha, B., & Goebel, K. (2008) Uncertainty Management for Diagnostics and Prognostics of Batteries using Bayesian Techniques. *IEEE Aerospace Conference*.
- Sankararaman, S., Ling, Y., & Mahadevan, S. (2010). Confidence Assessment in Fatigue Damage Prognosis. *Annual Conference of the Prognostics and Health Management Society*.
- Schwabacher, M. A. (2005). A survey of data-driven prognostics, in *AIAA Infotech@Aerospace Conference*. Reston,VA.
- Sheppard, J. W., Kaufman, M. A., Inc, A., & Annapolis, M. D. (2005). Bayesian diagnosis and prognosis using instrument uncertainty. *IEEE Autotestcon*, pp. 417-423.
- Sinclair, G. B., & Pierie, R. V. (1990). On obtaining fatigue crack growth parameters from the literature. *International Journal of Fatigue*, vol. 12(1), pp. 57-62.
- Virkler, D. A., Hillberry, B. M., & Goel, P. K. (1979). The statistical nature of fatigue crack propagation. *ASME Journal of Engineering Materials and Technology*, vol. 101, pp.148-153.
- Yan, J., & Lee, J. (2007) A Hybrid Method for On-line Performance Assessment and Life Prediction in Drilling Operations. *IEEE International Conference on Automation and Logistics*.
- Yu, W. K., & Harris, T. A. (2001). A new stress-based fatigue life model for ball bearings. *Tribology Transactions*, vol. 44, pp. 11-18. doi: 10.1080/10402000108982420.
- Zio, E., & Peloni, G. (2011). Particle filtering prognostic estimation of the remaining useful life of nonlinear components. *Reliability Engineering & System Safety*, vol. 96(3), pp. 403-409.

Dawn An received the B.S. degree and M.S. degree of mechanical engineering from Korea Aerospace University in 2008 and 2010, respectively. She is now a joint Ph.D.

student at Korea Aerospace University and the University of Florida. Her current research is focused on the Bayesian inference, correlated parameter identification and the methodology for prognostics and health management and structural health monitoring.

Joo-Ho Choi received the B.S. degree of mechanical engineering from Hanyang University in 1981, the M.S. degree and Ph.D. degree of mechanical engineering from Korea Advanced Institute of Science and Technology (KAIST) in 1983 and 1987, respectively. During the year 1988, he worked as a Postdoctoral Fellow at the University of Iowa. He joined the School of Aerospace and Mechanical Engineering at Korea Aerospace University, Korea, in 1997 and is now Professor. His current research is focused on the

reliability analysis, design for life-time reliability, and prognostics and health management.

Nam H. Kim received the B.S. degree of mechanical engineering from Seoul National University in 1989, the M.S. degree and Ph.D. degree of mechanical engineering from Korea Advanced Institute of Science and Technology (KAIST) in 1991 and the University of Iowa in 1999, respectively. He worked as a Postdoctoral Associate at the University of Iowa from 1999 to 2001. He joined the Dept. of Mechanical & Aerospace Engineering at the University of Florida, in 2002 and is now Associate Professor. His current research is focused on design under uncertainty, design optimization of automotive NVH problem, shape DSA of transient dynamics (implicit/explicit) and structural health monitoring.

Integrating Probabilistic Reasoning and Statistical Quality Control Techniques for Fault Diagnosis in Hybrid Domains

Brian Ricks¹, Craig Harrison², Ole J. Mengshoel³

¹ *University of Texas at Dallas, Richardson, TX, 75080, USA*

bwr031000@utdallas.edu

² *University of Maine, Orono, ME, 04469, USA*

craig.harrison@umit.maine.edu

³ *Carnegie Mellon University, NASA Ames Research Center, Moffett Field, CA, 80523, USA*

ole.mengshoel@sv.cmu.edu

ABSTRACT

Bayesian networks, which may be compiled to arithmetic circuits in the interest of speed and predictability, provide a probabilistic method for system fault diagnosis. Currently, there is a limitation in arithmetic circuits in that they can only represent discrete random variables, while important fault types such as drift and offset faults are continuous and induce continuous sensor data. In this paper, we investigate how to handle continuous behavior while using discrete random variables with a small number of states. Central in our approach is the novel integration of a method from statistical quality control, known as cumulative sum (CUSUM), with probabilistic reasoning using static arithmetic circuits compiled from static Bayesian networks. Experimentally, an arithmetic circuit model of the ADAPT Electrical Power System (EPS), a real-world EPS located at the NASA Ames Research Center, is considered. We report on the validation of this approach using PRODIAGNOSE, which had the best performance in three of four industrial track competitions at the International Workshop on Principles of Diagnosis in 2009 and 2010 (DXC-09 and DXC-10). We demonstrate that PRODIAGNOSE, augmented with the CUSUM technique, is successful in diagnosing faults that are small in magnitude (offset faults) or drift linearly from a nominal state (drift faults). In one of these experiments, detection accuracy dramatically improved when CUSUM was used, jumping from 46.15% (CUSUM disabled) to 92.31% (CUSUM enabled).

First Author et.al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

1. INTRODUCTION

Arithmetic circuits (ACs) (Darwiche, 2003; Chavira & Darwiche, 2007), which may be compiled from Bayesian networks (BNs) (Lauritzen & Spiegelhalter, 1988; Pearl, 1988) to achieve speed and predictability, provide a powerful probabilistic method for system fault diagnosis. While arithmetic circuits represent a significant advance in many ways, they can currently only represent discrete random variables (Darwiche, 2003; Chavira & Darwiche, 2007; Darwiche, 2009). At the same time, systems that one wants to diagnose are often hybrid—both discrete and continuous (Lerner, Parr, Koller, & Biswas, 2000; Poll et al., 2007; Langseth, Nielsen, Rumi, & Salmeron, 2009). For example, important fault types such as drift and offset faults are continuous and also induce continuous evidence (Poll et al., 2007; Kurtoglu et al., 2010).

The literature describes two main approaches to handling hybrid behavior in a discrete probabilistic setting: discretization (Kozlov & Koller, 1997; Langseth et al., 2009) and uncertain evidence, in particular soft evidence (Pearl, 1988; Chan & Darwiche, 2005; Darwiche, 2009). Naive discretization performed off-line leads to an excessive number of states, which is problematic both from the point of view of BN construction and fast on-line computation (Langseth et al., 2009). Discretization can be performed dynamically on-line (Kozlov & Koller, 1997; Langseth et al., 2009), however this is inconsistent with this paper's focus on off-line compilation into arithmetic circuits and fast on-line inference. Uncertain evidence in the form of soft (or virtual) evidence (Pearl, 1988) can be used to handle, in a limited way, continuous random variables (Darwiche, 2009). Typically, soft evidence is limited to continuous children of discrete random variables with two discrete states (0/1, low/high, etc.). In addition, the soft evidence approach requires changing the probability parameters

on-line, in the AC or BN, and is thus more complicated from a systems engineering or verification and validation (V&V) point of view.

In this paper, we describe an approach to handle continuous behavior using discrete random variables with a small number of states. We integrate a method from statistical quality control, known as cumulative sum (CUSUM) (Page, 1954), with probabilistic reasoning using arithmetic circuits (Darwiche, 2003; Chavira & Darwiche, 2007). We carefully and formally define our approach, and demonstrate that it can diagnose faults that are small in magnitude (continuous offset faults) or drift linearly from a nominal state (continuous drift faults).

Experimentally, we show the power of integrating CUSUM calculations into our diagnostic algorithm PRODIAGNOSE (Ricks & Mengshoel, 2009a, 2009b), which uses arithmetic circuits. We consider the challenge of diagnosing a broad range of faults in electrical power systems (EPSs), focusing on our development of an arithmetic circuit model of the Advanced Diagnostics and Prognostics Testbed (ADAPT), a real-world EPS located at the NASA Ames Research Center (Poll et al., 2007). The experimental data are mainly from the 2010 diagnostic competition (DXC-10) (Kurtoglu et al., 2010). In addition to the challenge of hybrid behavior, this data is sampled at varying sampling frequency, may contain multiple faults, and may contain sensor noise and other behavioral characteristics that are considered nominal behavior of ADAPT. Using the DXC data, we perform a comparison between experiments (i) with CUSUM and (ii) without CUSUM, and find significant improvements in diagnostic performance when CUSUM is used. In fact, PRODIAGNOSE was the best performer in the two competitions making up the industrial track of DXC-09, and a winner of one of the competitions in the industrial track of DXC-10.¹

In this paper, we extend previous research on the use of CUSUM with BNs and ACs (Ricks & Mengshoel, 2009b) in several ways. Specifically, we use CUSUM for drift faults (previously it was used for offset faults only); provide a more detailed discussion and analysis; and report on experimental results for a new dataset (DXC-10) that contains a broader range of faults, including abrupt faults, intermittent faults, and drift faults. Our CUSUM approach is crucial for handling two types of continuous faults, namely offset faults and drift faults.

The rest of this paper is structured as follows. In Section 2. we introduce concepts related to Bayesian networks, arithmetic circuits, CUSUM, and the fault types we investigate. Section 3. presents integration of CUSUM into our fault diagnosis approach, the PRODIAGNOSE algorithm, and discuss both the modeling and diagnostic perspectives. We present strong experimental results for electrical power system data in Section 4., and conclude in Section 5..

¹Please see <http://www.dx-competition.org/> for details.

2. PRELIMINARIES

2.1 Bayesian networks and Arithmetic Circuits

Diagnostic problems can be solved using *Bayesian networks* (BNs) (Lauritzen & Spiegelhalter, 1988; Pearl, 1988; Darwiche, 2009; Choi, Darwiche, Zheng, & Mengshoel, 2011). A Bayesian network is a directed acyclic graph (DAG) where each *node* in the BN represents a *discrete* random variable,² and each edge typically represents a causal dependency between nodes. Distributions for each node are represented as *conditional probability tables* (CPTs). Let \mathbf{X} represent the set of all nodes in a BN, $\Omega(X) = \{x_1, \dots, x_m\}$ the states of a node $X \in \mathbf{X}$, and $|X| = |\Omega(X)| = m$ the cardinality (number of states). The size of a node's CPT is dependent on its cardinality and the cardinality of each parent node. By taking a subset $\mathbf{E} \subseteq \mathbf{X}$, denoted the evidence nodes, and *clamping* each of these nodes to a specific state, the answers to various probabilistic queries can be computed. Formally, we are providing *evidence* \mathbf{e} to all nodes $E \in \mathbf{E}$, in which $\mathbf{E} = \{E_1, E_2, \dots, E_n\}$, $\mathbf{e} = \{(E_1, e_1), (E_2, e_2), \dots, (E_n, e_n)\}$, and $e_i \in \Omega(E_i)$ for $1 \leq i \leq n$ and $n \leq m$. Probabilistic queries for BNs include the marginal posterior distribution over *one* node $X \in \mathbf{X}$, denoted $\text{BEL}(X, \mathbf{e})$, over a *set* of nodes \mathbf{X} , denoted $\text{BEL}(\mathbf{X}, \mathbf{e})$, and most probable explanations over nodes $\mathbf{X} - \mathbf{E}$, denoted $\text{MPE}(\mathbf{e})$.

While Bayesian networks can be used directly for inference, we compile them to *arithmetic circuits* (ACs) (Chavira & Darwiche, 2007; Darwiche, 2003), which are then used to answer BEL and MPE queries. Key advantages to using ACs are speed and predictability, which are important for resource-bounded real-time computing systems including those found in aircraft and spacecraft (Mengshoel, 2007; Ricks & Mengshoel, 2009b, 2010; Mengshoel et al., 2010). The benefits of using arithmetic circuits are derived from the fact that BEL computations, for example, amount to simple addition and multiplication operations over numbers structured in a DAG. Compared to alternative approaches to computation using BNs, for example join tree clustering (Lauritzen & Spiegelhalter, 1988) and variable elimination (Dechter, 1999), AC computation has substantial advantages in terms of speed and predictability, even when implemented in software as done in this paper and previously (Chavira & Darwiche, 2007; Darwiche, 2003; Ricks & Mengshoel, 2009b; Mengshoel et al., 2010). The fundamental limitation of ACs is that they may, in the general case, grow to the point where memory is exhausted. In particular, this is a problem in highly connected BNs. The BNs investigated in this paper, as well as in similar fault diagnosis applications, are typically sparse (Mengshoel, Poll, & Kurtoglu, 2009; Ricks & Mengshoel, 2009a, 2009b; Mengshoel et al., 2010; Ricks & Mengshoel, 2010), and memory consumption turns out not to be a problem.

²Continuous random variables can also be represented in BNs, however in this article we focus on the discrete case.

2.2 Cumulative sum (CUSUM)

Cumulative sum (*CUSUM*) is a sequential analysis technique used in the field of statistical quality control (Page, 1954). CUSUMs can be used for monitoring changes in a continuous process' samples, such as a sensor's readings, over time. Let $\delta_p(t)$ represent the CUSUM of process p at time t . Then, taking $s_p(t)$ to be the unweighted sample (or sensor reading) from process p at time t , we formally define CUSUM as:

$$\delta_p(t) = [s_p(t) - s_p(t-1)] + \delta_p(t-1). \quad (1)$$

If $\delta_p(t)$ crosses a *threshold*, denoted as $v(i)$, a change in process p 's samples can be recorded. These thresholds represent points at which an interval change occurs in p —a transition from one interval to another (indicating a change from nominal to faulty in a closely related health node, for example). In other words, thresholds provide discretization points for our continuous CUSUM values. It is assumed that a process p starts out with $\delta_p(t)$ such that $v(i-1) \leq \delta_p(t) < v(i)$ for some i . Formally, an interval change for process p occurs when, at any time t , $\delta_p(t) < v(i-1)$ or $\delta_p(t) \geq v(i)$, in which $v(i-1)$ and $v(i)$ are a pair of thresholds at levels $i-1$ and i , respectively. Thresholds themselves are independent of time, in that they can be crossed at any time t . Values of thresholds are configurable, and obtained through experimentation.

Often, a set of thresholds for a sensor will only contain two thresholds. For these cases, we refer to $v(0)$ as the *lower threshold* and $v(1)$ as the *upper threshold*. This implies that the interval set of p has a cardinality of 3. The initial CUSUM value ($\delta_p(0)$) with respect to $v(0)$ and $v(1)$ will be $v(0) \leq \delta_p(0) < v(1)$. A lower and upper threshold are thus used to trigger an interval change if $\delta_p(t)$ ventures outside a nominal range bounded by the interval $[v(0), v(1))$ of the real line \mathfrak{R} .

2.3 Continuous Offset and Drift Faults

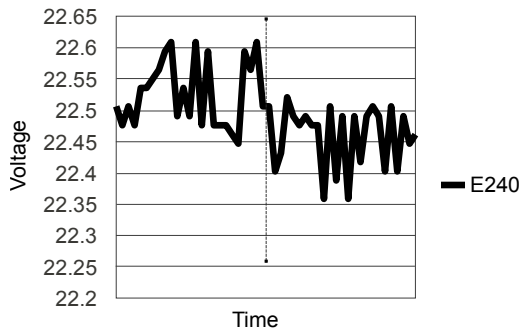


Figure 1. Graph illustrating an abrupt battery degradation over a span of 10 seconds, manifested as a voltage drop for sensor E240. The vertical dotted line on the graph indicates when the fault occurred.

We consider a system consisting of parts.³ For example, a system might be an electrical power system, and parts might

³A “part” is either a “component” or a “sensor” according to the termi-

be a battery, a wire, a voltage sensor, a temperature sensor, etc. Let $p(t)$ denote a measurable property of a part at time t . We now consider how a persistent fault, which is the emphasis of this paper, takes place.⁴ Let $p_n(t)$ denote the value of the property *before* fault injection, and $p_f(t)$ denote the value of the property *after* fault injection. More formally, let t^* be the time of fault injection. We then have:

$$p(t) = \begin{cases} p_n(t) & \text{for } t < t^* \\ p_f(t) & \text{for } t \geq t^* \end{cases}.$$

We can now formally introduce continuous offset and drift faults. A simple model for a *continuous (abrupt) offset fault* at time t is defined as follows (Kurtoglu et al., 2010):

$$p_f(t) = p_n(t) + \Delta p, \quad (2)$$

where Δp is an arbitrary positive or negative constant representing the offset magnitude. In other words, we do not know the values of Δp or t^* ahead of time, however once Δp is injected at time t^* , it does not change.

A key challenge is that Δp will be small for small-magnitude offset faults. For example, degradation of a battery (Figure 1), is often very small in magnitude (low voltage drop). When discussing the change in a sensor reading of a property, the notation Δs_p (sensed offset) rather than Δp (actual offset) is used. Sensor noise, while not reflected in (2), can mask the fault almost completely. In Figure 1, the magnitude of sensor noise would make diagnosis very difficult without first filtering data. An orthogonal issue is the large number of states needed in a discrete random variable, if a naive approach is used by representing a large number of offset intervals directly, which we would like to avoid.

A simple model for a *continuous drift fault* for a process p at time t is defined as follows (Kurtoglu et al., 2010):

$$p_f(t) = p_n(t) + m(t - t^*), \quad (3)$$

in which m is the slope. For example, a drifting sensor could output values that start gradually increasing at a roughly linear rate (Figure 2). As seen in Figure 2, drift faults may not be so obvious at first, due to sensor and other noise not reflected in (3). In a static Bayesian environment, the lack of time-measurement may make these faults appear as continuous offset faults. Not only would that diagnosis be incorrect, but the time of diagnosis may be quite a while after the initial fault, depending on the time t from t^* the drifting value crossed a threshold $v(i)$.

A major goal of our research is to correctly and quickly diagnose continuous offset and drift faults while minimizing the number of discretization levels of random variables.

nology used in this paper. In the DXC-10 documentation (Kurtoglu et al., 2010), the term “component” rather than “part” is used.

⁴The case of intermittent faults has been discussed previously (Ricks & Mengshoel, 2010).

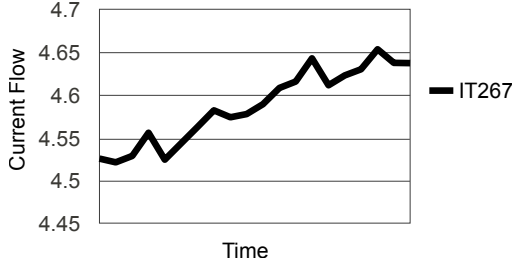


Figure 2. Graph of continuous drift fault behavior for a current sensor IT267 during a 10 second time span.

3. PROBABILISTIC DIAGNOSIS ALGORITHM

The PRODIAGNOSE diagnostic algorithm takes input from the environment, translates it into evidence, and computes a posterior distribution. The posterior distribution is then used to compute a diagnosis (Ricks & Mengshoel, 2009a, 2009b; Mengshoel et al., 2010; Ricks & Mengshoel, 2010). In this section we first summarize how PRODIAGNOSE computes diagnoses (Section 3.1) and the types of BN nodes it uses (Section 3.2). We then discuss how sensor readings, CUSUM, and Bayesian discretization fit together (see Section 3.3), and finally how PRODIAGNOSE handles continuous offset and drift faults by means of CUSUM techniques (Section 3.4 and Section 3.5).

3.1 The PRODIAGNOSE Diagnostic Algorithm

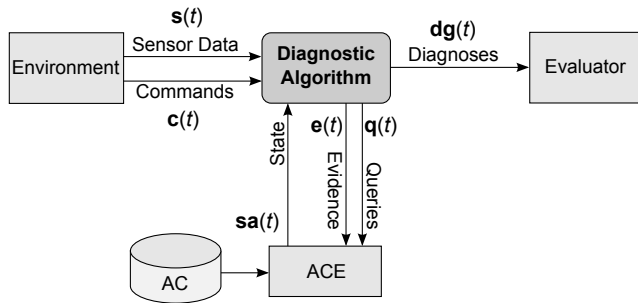


Figure 3. The PRODIAGNOSE DA architecture.

The PRODIAGNOSE diagnostic algorithm (DA) integrates the processing of environmental data (Figure 3) with an inference engine and post-processing of query results. At the highest level, the goal of PRODIAGNOSE is to compute a diagnosis $\mathbf{dg}(t)$ from sensor data $s(t)$ and commands $c(t)$. Input from the environment takes the form of sensor data, $s(t)$, and commands, $c(t)$. PRODIAGNOSE performs diagnosis when it receives sets or subsets of sensor data, which is expected at regular sample rate(s). It uses $s(t)$ and $c(t)$ to generate evidence, $e(t)$, reflecting the state of the system, which is then passed to the AC inference engine (ACE). The generation of $e(t)$ from $s(t)$ and $c(t)$ is non-trivial, and includes the use of CUSUM as discussed below. The health state of the system, $\mathbf{sa}(t)$,

is returned in response to a probabilistic query $\mathbf{q}(t)$, which is either an MPE or BEL query, $\text{MPE}(e(t))$ or $\text{BEL}(\mathbf{H}, e(t))$ respectively. PRODIAGNOSE then uses $\mathbf{sa}(t)$ to generate a diagnosis of the system, $\mathbf{dg}(t)$, by extracting the faulty states (if any) of the BN nodes \mathbf{H} that represent the health of the system being diagnosed. The algorithms used to compute CUSUM and perform offset and drift diagnosis (see Equation 5 and Algorithm 1) are called from within PRODIAGNOSE (Figure 3) as further discussed in the following subsections.

One strength of PRODIAGNOSE is its configurability. The BN representation of each part (a physical component or sensor) in an environment is configured individually, and this data is used by PRODIAGNOSE to initially calibrate itself to the environment and guides its behavior when receiving $s(t)$ or $c(t)$. In addition, PRODIAGNOSE is controlled by several global parameters, including:

- **Diagnosis Delay, t_{DD} :** Measured in milliseconds, this parameter gives the delay to start diagnosis output $\mathbf{dg}(t)$. In other words, $\mathbf{dg}(t)$ is empty for $t < t_{DD}$. Diagnosis delay is used at the beginning of environment monitoring, and is useful to filter out false positives (often due to transients) during the initial calibration process.
- **Fault Delay, t_{FD} :** This parameter delays the output of a new diagnosis (for a short while). Suppose that $\mathbf{sa}(t_i)$ contains, for the first time, a fault state f for a health node H . Then we hold off until time t_j , such that $t_j - t_i > t_{FD}$, to include f in $\mathbf{dg}(t_j)$. In many environments, one can get spurious diagnoses because of system transients (perhaps due to mode switching, perhaps due to faults), and this is a way to filter them out.

3.2 Bayesian Network Structures

A Bayesian network model of a system, for example an EPS, consists of structures modeled after physical components of the EPS (Ricks & Mengshoel, 2009a, 2009b). We discuss the following BN node types:

- $S \in \mathbf{S}$: All sensors utilize a *sensor* node S to clamp a discretized value of the physical sensor's reading $s(t)$. S nodes typically consist of three states, and thus have lower and upper thresholds.
- $H \in \mathbf{H}$: Consider $H \in \mathbf{H}$, namely a *health* node H . A health state $h \in \Omega(H)$ is output in $\mathbf{dg}(t)$, based on the result of BEL or MPE computation, to describe the health state of the component represented by H . We assume that $\Omega(H)$ is partitioned into nominal (or normal) states $\Omega_n(H)$ and faulty (or abnormal) states $\Omega_f(H)$. Only a faulty state $h \in \Omega_f(H)$ is output in $\mathbf{dg}(t)$. This is done for all $H \in \mathbf{H}$.
- $CH \in \mathbf{CH}$: A *change* node CH is used to clamp evidence for the purpose of change detection. CH nodes have a varying number of thresholds, depending on the purpose of the node (see Section 3.4).

- $DR \in \mathbf{DR}$: The *drift* node DR is used to clamp evidence about drift-type behavior. Typically, DR nodes use four thresholds, though this is configurable (see Section 3.5).
- $A \in \mathbf{A}$: An *attribute* node A is used to represent a hidden state, such as voltage or current flow.

Note that there are other nodes types besides S , CH , and DR nodes used to input evidence $e(t)$ (Ricks & Mengshoel, 2009a, 2009b; Mengshoel et al., 2010; Ricks & Mengshoel, 2010). Since they are not the focus in this paper, we represent these by nodes E_1, \dots, E_n in Figures 6 and 9.

3.3 Sensors, CUSUM, and Bayesian Network Discretization

To help filter out sensor noise, we take a weighted average of the raw sensor readings, and these weighted averages are then the basis for CUSUM computation. Let $\bar{s}_p(t)$ be the weighted average of readings $\{s_p(t-n), \dots, s_p(t)\}$ for sensor p at time t . Specifically, we have:

$$\bar{s}_p(t) = \sum_{i=0}^n s_p(t-i)w(t-i), \quad (4)$$

in which $s_p(t-i)$ is the raw sensor reading and $w(t-i)$ is the weight at time $t-i$. The summation is over all sensor readings from time t to time $t-n$. In other words, we keep a history of the past n sensor readings. The values of n and of all weights $\{w(t-n), \dots, w(t)\}$ are configurable and set based on experimentation.

The weighted sensor averages in (4) can be used when calculating CUSUM (Ricks & Mengshoel, 2009b), and (1) can be modified accordingly:

$$\bar{\delta}_p(t) = [\bar{s}_p(t) - \bar{s}_p(t-1)] + \bar{\delta}_p(t-1), \quad (5)$$

in which $\bar{\delta}_p(t)$ is the weighted average CUSUM of sensor p at time t . Weighted averages help to smooth spikes in sensor readings that could otherwise lead to false positives or negatives, and (5) is the CUSUM variant used in PRODIAGNOSE.

Figure 4 shows, for voltage sensor E240, the weighted CUSUM values $\bar{\delta}_p(t)$ overlaid with the raw sensor readings for the same time period. A downward trend after the offset fault occurrence is visually seen when looking at the CUSUM values, which makes setting appropriate thresholds to catch the fault possible. After the lower threshold, $v(0)$, is reached, the voltage sensor's CH node changes state. The CUSUM values in Figure 4 are calculated by keeping a history of the past $n = 6$ sensor readings, for any time t .

In this paper, CUSUM intervals are mapped to BN node states in a natural way. If we have k CUSUM intervals defined on the real line \mathbb{R} (so $k-1$ thresholds), then the corresponding BN evidence node E has k states $\Omega(E) = \{e_1, \dots, e_k\}$ also. If a CUSUM $\bar{\delta}_p(t)$ crosses a threshold into an interval $[v(i),$

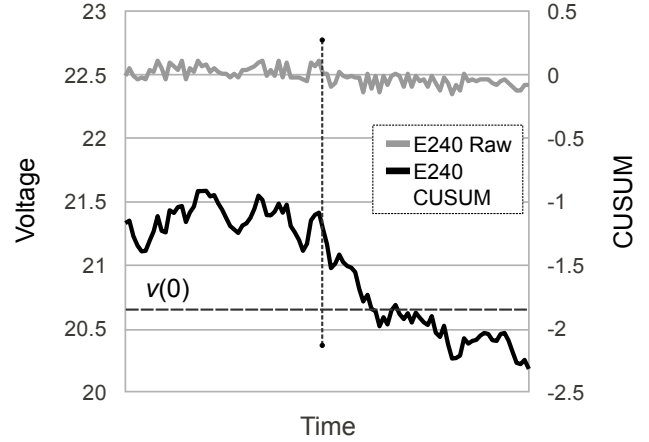


Figure 4. Graph illustrating raw voltage sensor readings, for sensor E240, and corresponding CUSUM values over a time span of 10 seconds. The vertical dotted line on the graph indicates when a very small offset fault occurred, and the horizontal dashed line represents the lower threshold, $v(0)$. The E240 CUSUM shows a very distinct downward trend after the fault.

$v(i+1)$), the corresponding BN evidence node E_p has a corresponding transition into state e_{i+1} .

We now describe CUSUM's characteristics and benefits, taking (5), the discretization, and Figure 4 as starting points. First, CUSUM *amplifies* a small offset, along the y -axis, by making it larger such that it becomes easier to create a threshold for and detect. Second, CUSUM *normalizes* by shifting offsets that can take place at different y -values to a normalized y -value, such that offsets can be detected using thresholds that apply to many y -values. Please see Figure 5 for an example. There is a clear impact on IT240, which can easily be detected with upper and lower thresholds, while the impact on IT267 is minimal since this sensor is downstream of a compensating inverter.

CUSUM's normalization works in combination with weighting the sensor values to give better discretization points. A key point here is that our algorithm calibrates in the beginning of a scenario. The algorithm computes a zero (or nominal) line based on initial sensor readings, and does not flag diagnoses. To the left in Figure 5, both IT240 and IT267 are close to this zero line. This zero line can help compensate for early transients, which may trigger diagnoses, but more importantly, it makes the nominal value of a sensor something we do not need to know ahead of time. Without CUSUM, we would have to know this nominal value ahead of time, which becomes difficult as a system naturally ages. With CUSUM, we use the first time period t_{DD} of a scenario to figure this out (calibration). After t_{DD} , CUSUM is relative number to the nominal reading, with zero being the nominal (weighted) value of the sensor. After the fault injection, we see in Figure 5 that IT267 CUSUM stays well above the $v(0)$ lower thresh-

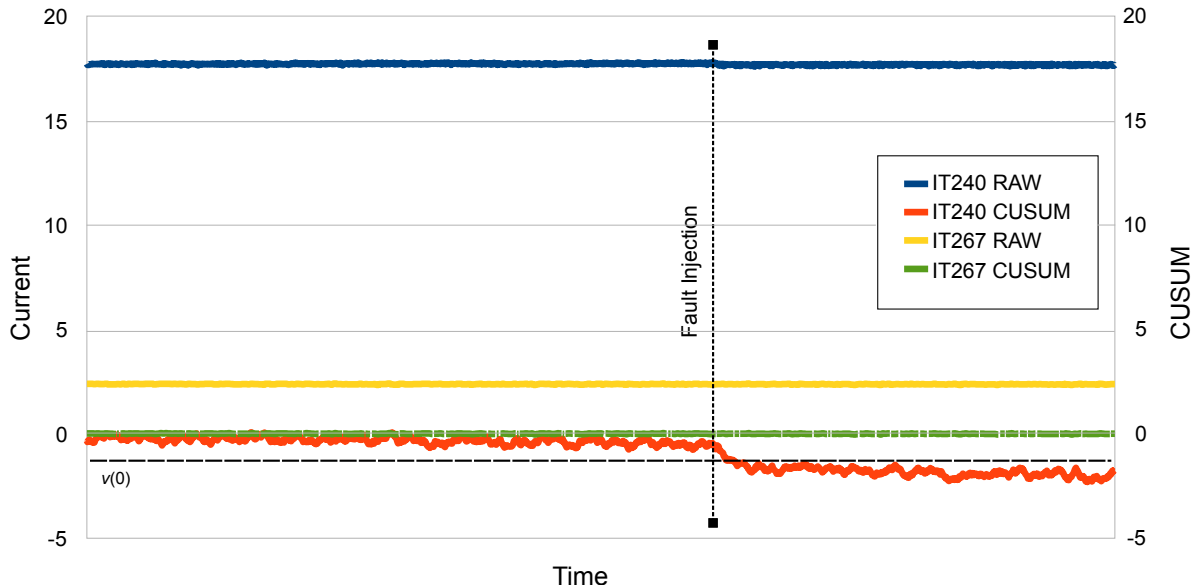


Figure 5. Graph illustrating the impact of an abrupt battery degradation offset fault on two current sensors. One sensor is immediately downstream from the battery (IT240), while the other sensor is farther away, downstream of an inverter (IT267). The vertical dotted line on the graph indicates when the fault occurred during the 3 minute time interval shown.

old, while IT240 CUSUM drops below $v(0)$.

3.4 Handling Continuous Offset Faults: Change Nodes and CUSUM

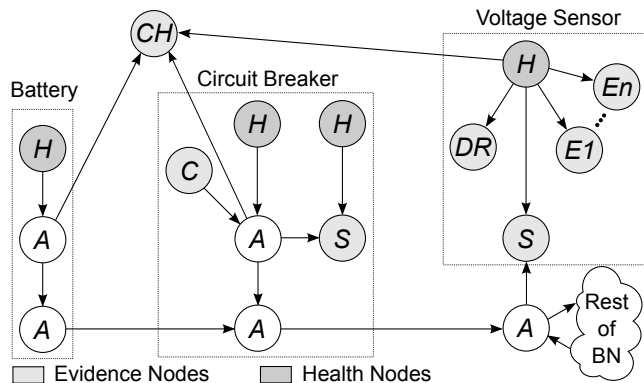


Figure 6. Bayesian network structure for a battery. The CH node provides additional evidence to the health state of the battery.

Suppose that we have a battery, a circuit breaker, a voltage sensor, and a load (say, a light bulb or a fan) connected in series. We assume that all these components and loads may fail, in different ways, and a realistic EPS will contain additional components and sensors that may also fail. Now consider the case of a continuous offset fault in the battery. There is the challenge of diagnosing an offset fault, which might be very small, using a discrete BN with relatively few states per node. In addition, there is the challenge of detecting an offset fault

Change_voltage.e140				
CH		H		
nominal	drop	A (Battery)	A (Breaker)	H (Sensor)
0.95	0.05	enabledHigh	closed	healthy
0.05	0.95	enabledLow		
0.5	0.5	disabled	open	
0.5	0.5	enabledHigh		
0.5	0.5	enabledLow		
0.5	0.5	disabled	closed	(not healthy)
0.5	0.5	enabledHigh		
0.5	0.5	enabledLow		
0.5	0.5	disabled		
0.5	0.5	enabledHigh		
0.5	0.5	enabledLow		
0.5	0.5	disabled		

Table 1. Conditional probability table (CPT) of the CH node from Figure 6. Each row shows the probabilities for the CH node's *nominal* and *drop* states (first two columns) for each combination of states from the battery's A node, circuit breaker's A node, and voltage sensor's H node (columns 3-5). Since the probabilities remain identical for all rows when the H node is unhealthy, we simplified the table by combining all these unhealthy states into (*not healthy*).

in an upstream component (the battery) using a downstream sensor (the voltage sensor). Clearly, we want to retain the capability of diagnosing other types of faults (see (Ricks & Mengshoel, 2009a, 2009b; Mengshoel et al., 2010; Ricks & Mengshoel, 2010)) in the battery, the circuit breaker, and the voltage sensor.

Figure 6 illustrates how we meet these challenges using BNs

and CUSUM computation. Specifically, Figure 6 shows the BN representation of a battery, a circuit breaker, and a voltage sensor connected in series. Under nominal conditions, power flows from the battery, through the circuit breaker and voltage sensor, and to one or more loads (not depicted). Traditionally, voltage sensor readings $s_p(t)$ for a sensor p lie on the real line \mathfrak{R} , but can be discretized so they end up being in one of the BN states $\Omega(S) = \{voltageLo, voltageHi, voltageMax\}$ for BN node S . The resulting set of thresholds will contain two levels for the S node in the BN (the lower and upper threshold). In the case in which the current state is $S = voltageHi$, the transition from this state to another occurs when $s_p(t) < v(0)$ or $s_p(t) \geq v(1)$ (see Section 2.2). If the current state is $S = voltageLo$, a state change would occur when $s_p(t) \geq v(0)$.

While the simple discretization approach discussed above is sufficient in many cases, in other cases it does not help, and this is the case for example for offset faults. If Δp (Equation 2) is small such that no state change among $\Omega(S)$ occurs, then it is possible that an offset fault may go undiagnosed.

To handle this problem, we use CH nodes and CUSUM. Intuitively, the purpose of the CH node in Figure 6 is to provide additional information about the battery, and specifically continuous offsets in sensor readings, while at the same time be influenced by the circuit breaker and voltage sensor. Offsets in sensor readings (Δs_p) that are too small in magnitude to cross a threshold are handled using CH nodes and CUSUM techniques. Evidence $e(t)$ that is clamped to a CH node is derived from the readings $s_p(t)$ of a sensor assigned to it, called the *source*. The readings from this source sensor are converted to CUSUM values, which are then discretized and clamped to the CH node.

In Figure 6, it is the battery, and specifically its health node H , that the CH node should influence when clamped with evidence $e(t)$ that indicates a continuous offset fault. We will call the battery the *target*. To understand in more detail how this BN design works, consider in Figure 6 the parent nodes of the CH node. Both the circuit breaker and voltage sensor parent nodes have evidence nodes as children or parents. This is not true, however, for the battery's A node, which is also a parent node of the CH node. This design makes sure that evidence $e(t)$ clamped to CH has less influence on the health nodes of the circuit breaker and voltage sensor compared to the influence on the battery's health node. And it is after all battery health, specifically offset faults, that we are targeting with this CH node.

More generally, any component between a source sensor and a target could affect the relevance of the CH node's evidence $e(t)$ for the target. For these intermediary components, it is the physical state that we are concerned about. For a circuit breaker, this is either *open* or *closed*. An *open* state should increase the probability that any voltage drop downstream is the result of this circuit breaker state and not due to a degra-

ation of the battery. The physical state of these components are usually represented as the state of an A node that belongs to the component structure, as in Figure 6. These A nodes will have a parent-child relationship to the CH node that is similar to the relationship of the source sensor's H node.

Table 1 shows the CPT for the CH node depicted in Figure 6. Notice that for all rows in which the voltage sensor (column 5) is *not healthy*, the probabilities for all CH node states are equal. Therefore, despite the discretized CUSUM value that gets clamped to the CH node, the impact on the posterior distribution of the battery's health node H (from the additional evidence provided by the CH node) will be very small in this case.

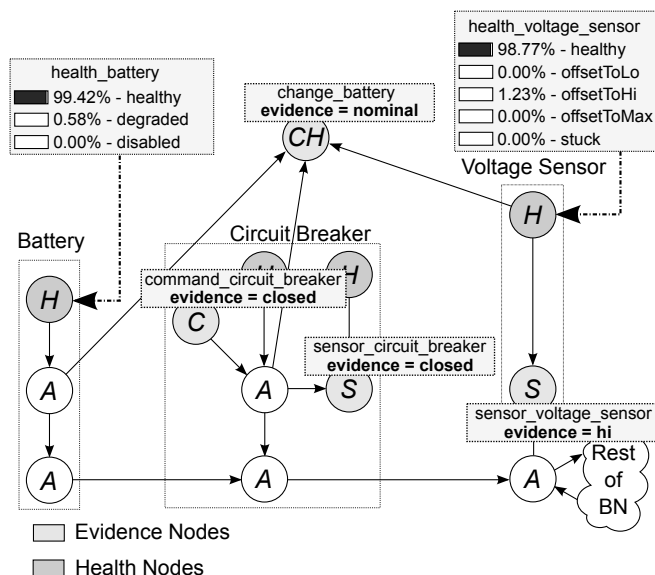


Figure 7. The marginal posterior distributions for the H nodes of the voltage sensor and battery from Figure 6 when no fault is present.

A CH node used for offset fault diagnosis is typically discretized into the same number of states as its source. For instance, the CH node for Figure 6 only has three states, $\Omega(CH) = \{drop, nominal, rise\}$, to indicate a downward change, no change, or upward change, respectively. Thus, as its source sensor, the CH node will utilize a lower and upper threshold. In the case of a battery degradation, the state of the voltage sensor's S node (Figure 6) may not change at all, but the CH node's state will become $CH = drop$ after crossing $v(0)$ to indicate the slight voltage drop due to the degradation.

Figure 7 shows the marginal posterior distributions for the H nodes of the source sensor and battery (Figure 6) under nominal conditions. In this example, the circuit breaker's state is *closed*, and the source sensor is deemed to be *healthy*. The state of the CH node is *nominal*. Thus, the state of the battery's health node is *healthy* with very high probability.

Now suppose the source sensor's value drops, but the mag-

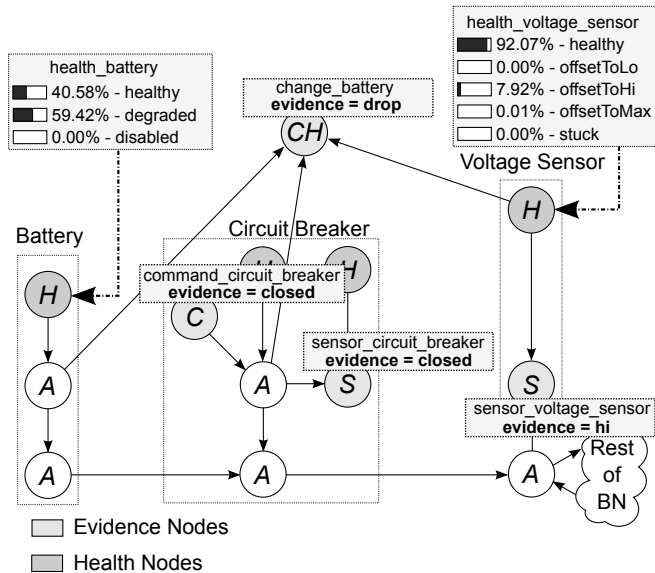


Figure 8. The marginal posterior distributions for the H nodes of the voltage sensor and battery from Figure 6, after a slight voltage drop.

nitude of the drop is too small to cross a threshold $v(i)$ in the source sensor. In such a scenario, the state of the source sensor would not change, and assuming all other evidence clamped in the BN stays the same, this slight drop would not cause the target’s health state to change in the absence of the CH node. With the CH node present however (Figure 6), this voltage drop could be detected. In Figure 8, the state of the CH node changes from *nominal* to *drop*, and thus, the probability of the *degraded* state for the battery’s H node increases to become the most probable explanation.

Figure 9 shows another use for CH nodes and CUSUM. Here, we have a bank with many loads, but very few sensors to provide evidence concerning their state. Some of these loads have no sensors at all, and hence diagnosing these loads becomes difficult. Here, we create additional CUSUM evidence $e(t)$, and clamp it to a single CH node. The CUSUM values are derived from the single current flow sensor (the source) that measures current flow into the load bank (the target). The CH node provides discretization of this CUSUM at a higher resolution (with more states) compared to the CH node from our previous example (Figure 6).

In a configuration such as this, the CH node will have many thresholds, $\{v(0), v(1), \dots, v(n)\}$, that correspond to $n + 1$ states. Taking $v(i - 1)$ and $v(i)$ to be the bounds on a nominal range (nominal state) for the CH node’s CUSUM values, each sequential crossing of a threshold $\{v(i), v(i + 1), \dots, v(n)\}$ or $\{v(i - 1), v(i - 2), \dots, v(0)\}$ represents an offset of increasing magnitude, and typically corresponds to a specific fault in a component in the bank. These faults are usually offset faults or component failures.

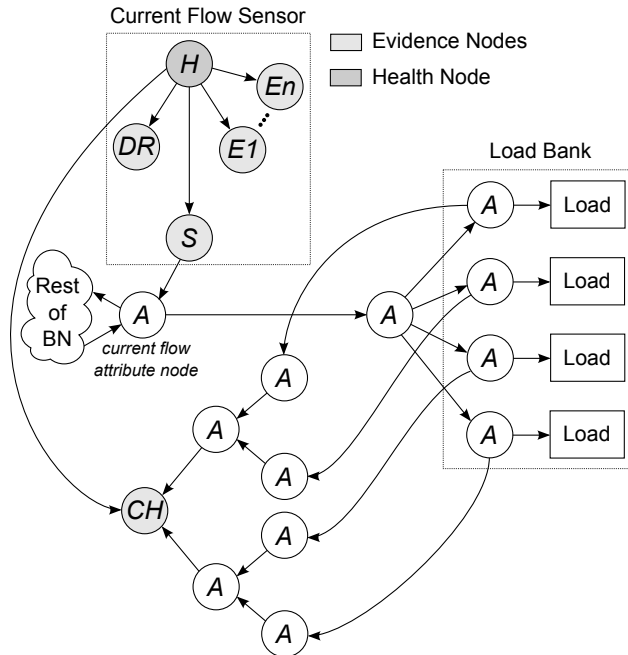


Figure 9. A simplified model of a BN model for a load bank. Evidence $e(t)$ clamped to the CH node is derived from the current sensor’s readings. The CH node forms the leaf of a tree, which is structured to limit the CPT size of CH .

Figure 9 contains a tree-like structure that connects the components in the bank to the CH node. This serves to sum the current flows (providing the source sensor measures current flow) of each component, so that the CPT size of the CH node is minimized (Table 2). Note that this does not affect the number of states, $|\Omega(CH)|$, but rather the number of parents of a CH node, and hence the size of its CPT. This technique serves to combine similar states along the tree that would otherwise be present in the CH node’s CPT if each load bank component was a parent of the CH node. For example, all components in the load bank have a state that corresponds to no current draw. In Figure 9, this would equal a total of four *no current* states. The two parent A nodes of the CH node (Figure 9) themselves have two parents. Each of these parents has a state which corresponds to *no current* and thus the CH node’s CPT only has to have probabilities for *no current* states pertaining to its two parents (see Table 2) rather than all four parents if all load bank component A nodes were parents of the CH node.

Using Figure 9 as an example, consider a simple situation in which all loads in the bank are healthy with a state set of $\Omega(X) = \{healthy, failed\}$. Assuming this corresponds to an A state of $w60$ for each load and the source sensor is healthy, we would see the most probable state of the CH node as $w240$. This state would be based on the CUSUM generated from the source sensor’s values. Now suppose one of the loads failed. Since the source sensor is the sole sensor

Change_current_it167								
CH						A	A	H
w0	w30	w60	...	w420	w450	A	A	H
0.95	0.05	0.0	...	0.0	0.0	w0	w0	healthy
0.0	0.05	0.9	...	0.0	0.0	w60		
0.0	0.0	0.05	...	0.0	0.0	w90		
...		
0.0	0.0	0.0	..	0.0	0.0	w240		
0.0	0.0	0.0	...	0.0	0.0	w270		
0.05	0.9	0.05	...	0.0	0.0	w0		
0.0	0.0	0.05	...	0.0	0.0	w60		
0.0	0.0	0.0	...	0.0	0.0	w90		
...		
0.0	0.0	0.0	...	0.0	0.0	w240		
0.0	0.0	0.0	...	0.0	0.0	w270		
...		
0.0	0.0	0.0	...	0.0	0.0	w0		
0.0	0.0	0.0	...	0.0	0.0	w60		
0.0	0.0	0.0	...	0.0	0.0	w90		
...		
0.0	0.0	0.0	...	0.05	0.0	w240		
0.0	0.0	0.0	...	0.9	0.05	w270		
0.06	0.06	0.06	...	0.06	0.06	w0		
...		
0.06	0.06	0.06	...	0.06	0.06	w270		
...		
0.06	0.06	0.06	...	0.06	0.06	w0		
...		
0.06	0.06	0.06	...	0.06	0.06	w270		

Table 2. Conditional probability table (CPT) of the CH node from Figure 9. This table is laid out in the same format as Table 1. Some simplifications were made to the CPT so it would fit, including taking out some intermediary states (represented by '...') from the CH node and its parents.

for the entire bank, and considering it only has three states, this load failure may not cause enough of a current drop to cause a change of state in the source sensor. If this were the case, the fault would be missed completely. Fortunately, the CH node would detect this drop and its state would change to $w180$.⁵

3.5 Handling Continuous Drift Faults: Drift Nodes and CUSUM

Drift faults are characterized by a gradual, approximately linear change of behavior (see Section 2.3), though sensor noise may disrupt strict linearity. While abrupt offset faults are diagnosed as soon as a threshold is crossed, due to their near vertical slope at the moment the fault occurs, drift faults usually do not cross these same offset thresholds immediately, and our diagnosis of them utilizes time t alongside thresholds that are specific for drift faults. Utilizing time and new thresholds help to differentiate an abrupt fault that trips a threshold, from a drift fault that would also trip this same threshold after

⁵The additional evidence from the CH node would be enough to determine that a failure had occurred, but not for a specific load. We assume that a few more sensors would be available to provide additional evidence to the load bank.

a certain period of time.

A DR node (see Figure 6, voltage sensor) is used to clamp a boolean drift state, $\Omega(DR) = \{nominal, faulty\}$ for a component structure. By default, it is clamped to $DR = nominal$, or no drift present.

Drift tracking uses threshold values and times on multiple levels, defined as $i \in \{0, \dots, n\}$. For the i -th level we have:

$$\lambda(i) = [v(i), t_{\min}(i), t_{\max}(i)], \quad (6)$$

where $v(i)$ is a threshold value (see Section 2.2), $t_{\min}(i)$ is a threshold minimum time, and $t_{\max}(i)$ is a threshold maximum time. Here, $v(i)$ represents the threshold that must be reached to move to the next level. Level $i = 0$ is the initial level and has these thresholds associated with it: $\lambda(0) = (v(0), 0, \infty)$. Once $v(i)$ is reached, PRODIAGNOSE moves to level $i + 1$ only if $t_{\min}(i) \leq t(i) < t_{\max}(i)$, in which $t(i)$ is the time elapsed since the last threshold was reached.⁶ If $v(i)$ is reached but $t(i) < t_{\min}(i)$ or $t(i) \geq t_{\max}(i)$, drift tracking resets to $i = 0$. Once the maximal level n has been reached, there is no reset. Pseudo-code for the DRIFTTRACKER is provided in Algorithm 1.

The number of levels is configurable. Currently, $i \in \{0, 1, 2, 3\}$; thus $\lambda(3)$ represents evidence $e(t)$ of a drift fault, and $DR = faulty$ at this point. In other words, DR is clamped to $faulty$ once level $i = 3$ has been reached.

Algorithm 1 DRIFTTRACKER($\lambda(i), t(i), i, n$)

```

th ← v(i) ∈ λ(i)
min ← tmin(i) ∈ λ(i)
max ← tmax(i) ∈ λ(i)
if i = n then
    return faulty
end if
if |δp(t)| > th and min ≤ t(i) < max then
    if i < n then
        i ← i + 1
    else
        return faulty
    end if
else
    i ← 0
end if
return nominal

```

To bring out the underlying behavioral patterns in a sensor's readings and help filter out noise, a threshold $v(i)$ is compared against the sensor's CUSUM value, $\bar{\delta}_p(t)$, at each timestep t , and not its raw reading $s_p(t)$.

Algorithm 1 is integrated into PRODIAGNOSE for DR nodes similar to how CUSUM is integrated for CH nodes (Ricks & Mengshoel, 2009b, 2010). The raw sensor data for S nodes are handled first, which includes assigning the raw sensor values and discretization of these values (Ricks & Mengshoel,

⁶To handle both upward and downward drift faults, we take the absolute value of $\bar{\delta}_p(t)$. We can safely do that here, under the assumption that drift faults are consistently one or the other.

2009b). We then process CUSUM values for all *CH* nodes and call the DriftTracker (Algorithm 1) for all *DR* nodes (which also uses CUSUM values and thus shares much code with that for *CH* node processing). *CH* and *DR* node types take input from their *S* node source sensors. Similar to *S* node processing, the CUSUM value for each *CH* and *DR* node is discretized after being calculated. These discretized values are then clamped to their respective evidence nodes in the BN before inference is performed.

Drift_voltage.e140		
<i>DR</i>		
nominal	faulty	<i>H</i> (Sensor)
0.99	0.01	healthy
0.99	0.01	offsetToLo
0.99	0.01	offsetToHi
0.99	0.01	offsetToMax
0.01	0.99	drift
0.99	0.01	stuck

Table 3. Conditional probability table (CPT) of the *DR* node from Figure 6. If the *DR* node is clamped to *nominal*, then the *H* node (column 3) has a high probability of being in any state except for *drift*. Conversely, if *faulty* is clamped to the *DR* node, the *H* node has a high probability of being in the *drift* state.

Using Figure 6 as an example, consider a situation in which no drift fault is present. This would result in a *DR* state of *nominal*. According to the *DR* node’s CPT (Table 3), with a clamped state of *nominal*, the parent *H* node has equal probability of being in any state except for the *drift* state. Now if a drift fault were to be detected for the voltage sensor, the *DR* clamped state would become *faulty*. This would greatly increase the probability of the voltage sensor’s health state changing to *drift* (Table 3). Note that since *drift* is an unhealthy state, the *CH* node from Figure 6 would no longer have much influence over the battery under this condition (see Section 3.4).

4. EXPERIMENTAL RESULTS

For experimentation, the ADAPT EPS was used. ADAPT is a real-world EPS, located at the NASA Ames Research Center, that reflects the characteristics of a true aerospace vehicle EPS, while providing a controlled environment for injecting faults and recording data (Poll et al., 2007). Data from each ADAPT run is stored in a scenario file, which can later be ingested by the diagnostic software. This design means the diagnostic algorithm can be repeatably run on any scenario file, supporting iterative improvement of the BN and diagnostic system during the development process.

In this section, we report on experimental results for PRODIAGNOSE using two ADAPT data sets, namely DXC-09 and DXC-10 data used for competitions arranged as part of the

DX Workshop Series.⁷

4.1 Experiments Using DXC-10 Training Data

The Diagnosis Workshop’s 2010 competition (DXC-10)⁸ was divided into two tiers: Diagnostic Problem 1 (DP1) and Diagnostic Problem 2 (DP2). A main difference, compared to the 2009 competition (DXC-09), was the inclusion of drift (or incipient) and intermittent faults in DXC-10. Abrupt faults (including abrupt offset faults) were included in DXC-10, as in DXC-09. Consequently, these data sets test the performance of PRODIAGNOSE on drift and abrupt offset faults, which is where our CUSUM-based technique are intended to help. These experimental results were obtained by running training set scenarios provided to all DXC-10 competitors.

4.1.1 Diagnostic Problem 1

DP1 uses a subset of the ADAPT EPS. This subset consists of one battery, connected to a DC load, and an inverter with two AC loads. ADAPT is in a fully powered-up state throughout a scenario. Scenarios generated from this configuration of the EPS are either single-fault or nominal. DP1 contains both offset faults and drift faults, both of which test our CUSUM-based diagnosis technique.

DP1 consists of 39 scenarios in its training set. Of these, 5 are nominal (no fault injection), 12 involve sensor faults, and 22 involve component faults. Of these 39 scenarios, 7 contain offset faults, and 7 contain drift faults. Note that 9 other scenarios in the DP1 training set contain intermittent offset faults. While PRODIAGNOSE handles these similar to the abrupt case, details have been discussed previously (Ricks & Mengshoel, 2010) and are beyond this paper’s scope.

The DP1 Bayesian network currently has a total of 148 nodes, 176 edges, and a cardinality range of [2, 10]. The DP1 BN has the same overall structure as the DP2 BN (see Section 4.1.2). Some notable differences are the inclusion, in DP1, of additional evidence nodes (such as *DR* nodes) for fault types that are not present in DP2, specifically intermittent and drift faults, and additional *CH* nodes to aid in load fault diagnosis of fault types such as drift faults.

The metrics in Table 4 are briefly summarized here to aid interpretation of the results. *Mean Time To Isolate* refers to the time from when a fault is injected until that fault is diagnosed. *Mean Time To Detect* refers to the time from when a fault is injected until any fault is detected. *False Positives* occur when PRODIAGNOSE diagnoses a fault that is not actually present. *False Negatives* occur when PRODIAGNOSE fails to diagnose a fault that is present. Low *False Positive Rates* are important because it is undesirable to perform corrective action when the system is operating correctly. A low *False Negatives Rate*

⁷More information about the diagnostic competitions, including these data sets, can be found here: <http://www.dx-competition.org/>.

⁸More information on DXC-10, including scenario files, can be found here: <https://www.phmsociety.org/competition/dxc/10>.

Metric	CUSUM	
	Enabled	Disabled
Detection Accuracy	92.31%	46.15%
False Positives Rate	0%	0%
False Negatives Rate	8.82%	61.76%
Mean Time To Detect	17.97 s	28.36 s
Mean Time To Isolate	72.27 s	51.14 s

Table 4. Experimental results with CUSUM enabled and disabled using electrical power system scenarios for DP1.

indicates that few system faults will remain undetected.

In these experiments, PRODIAGNOSE achieved an impressive *False Positives Rate* of 0% and a *False Negatives Rate* of 8.82% when CUSUM was enabled. When CUSUM was disabled, on the other hand, detection accuracy plummeted to 46% with a false negative rate of almost 62%. Detection times also increased with CUSUM disabled, due to increased detection time of certain offset faults that now must rely solely on an *S* node state change. Note that when CUSUM is disabled, drift faults are difficult to diagnose correctly (they will appear as abrupt offset faults) due to drift tracking's dependence on CUSUM. However, this actually lowers isolation times due to no isolation time being recorded for a mis-diagnosis.⁹

4.1.2 Diagnostic Problem 2

DP2 represents the entire ADAPT EPS. ADAPT consists of three batteries as the source, connected to two DC load banks, and two inverters each connected to an AC load bank. Scenarios generated from the full ADAPT EPS can be single, double, or triple-fault; or nominal. ADAPT is initially in a powered-down state, and various relays are closed and opened through a scenario to provide power to various components of the EPS. DP2 contains offset faults, but no drift faults, and thus our CUSUM-based diagnosis approach is not as extensively tested as in DP1.

DP2's training set contains 34 scenarios in total: 7 nominal, 9 with sensor faults, and 21 with component faults (some scenarios have both sensor and component faults). Among DP2 scenarios, 6 contain offset faults.

Since DP2 does not contain scenarios with drift and intermittent faults, the DP2 Bayesian network does not implement support for all the fault types seen in DP1. Thus, additional evidence nodes (such as *DR* nodes) for these fault types are omitted from the DP2 BN. The DP2 BN currently has a total of 493 nodes, 599 edges, and a cardinality range of [2, 16].

Experimental results for DP2 are summarized in Table 5. Compared to DP1, the DP2 data set did not have as many scenarios that might benefit from CUSUM (though it is worth

⁹This is according to the DXC definition of *Mean Time To Isolate*; one could certainly make the argument that a mis-diagnosis should be punished more harshly.

Metric	CUSUM	
	Enabled	Disabled
Detection Accuracy	90.91%	87.88%
False Positives Rate	3.03%	3.03%
False Negatives Rate	7.69%	11.54%
Mean Time To Detect	5.74 s	10.56 s
Mean Time To Isolate	36.78 s	39.97 s

Table 5. Experimental results with CUSUM enabled and disabled using electrical power system scenarios for DP2.

noting that 13 scenarios which involved component faults are diagnosed by catching offsets in sensor readings). Consequently, DP2's increase in accuracy when using CUSUM is not as pronounced, although it does improve from 87.88% to 90.91%. Most of DP2's faults could be diagnosed without needing the additional evidence provided by *CH* nodes. DP2 also does not contain drift faults, for which PRODIAGNOSE is dependent on CUSUM techniques to diagnose. In addition, with CUSUM enabled, the mean detection time decreased by almost half, due to the role it plays in load bank component diagnosis (see Section 3.4). This is significant, as quick diagnosis is very important in aircraft and spacecraft. Finally, using CUSUM should not adversely impact the overall diagnostic performance of PRODIAGNOSE, and we see that all metrics in Table 5 are equally good or better when CUSUM is enabled compared to when it is disabled.

4.2 DXC-09 and DXC-10 Competition Results

PRODIAGNOSE had the best performance in three of four of the Diagnosis Workshop's industrial track competitions in 2009 and 2010 (DXC-09 and DXC-10). In both DXC-09 and DXC-10 the CUSUM techniques discussed in this paper played a crucial role. DXC-10 competition data indicate strong performance of PRODIAGNOSE (Kurtoglu et al., 2010), implementing the CUSUM approach for diagnosis of offset and drift faults, against algorithms relying on alternate techniques. In the official competition, PRODIAGNOSE achieved an overall scenario detection accuracy of 82.5% in DP1 and 89.2% in DP2, surpassing the second-best DP2 entrant by 19%. In the DP2 category, PRODIAGNOSE also had the fewest fault classification errors and the quickest fault detection time. Data from the 2009 competition (DXC-09) indicate PRODIAGNOSE as the top performer with detection accuracies of 96.7% and 88.3% in Tier 1 and Tier 2, respectively (Kurtoglu et al., 2009).

5. CONCLUSION

For fault diagnosis in complex and resource-constrained environments, we would like a diagnosis algorithm to be exact, fast, predictable, able to handle hybrid (discrete and continuous) as well as dynamic behavior, and easy to verify and validate (V&V).

Fulfilling all these requirements is certainly a tall order. However, we have in this paper extended previous work on the PRODIAGNOSE diagnostic algorithm (Ricks & Mengshoel, 2009a, 2009b; Mengshoel et al., 2010; Ricks & Mengshoel, 2010), and discussed the promise of using static arithmetic circuits, compiled from static Bayesian networks. In particular, we have shown how fault diagnosis using static arithmetic circuits can be augmented with a cumulative sum (CUSUM) technique, resulting in dramatically improved performance in situations with continuous fault dynamics. In experiments with data from a real-world electrical power system, we have observed that our CUSUM-based technique leads to significantly improved performance in situations with continuous offset and drift faults. In addition, the CUSUM techniques discussed in this paper played a crucial role in the strong performance of PRODIAGNOSE in the Diagnosis Workshop's industrial track competitions in 2009 and 2010 (DXC-09 and DXC-10). In DXC-09 and DXC-10, PRODIAGNOSE achieved the best performance in three of four industrial track competitions.

Acknowledgments

This material is based, in part, upon work supported by NSF awards CCF0937044 and ECCS0931978.

REFERENCES

- Chan, H., & Darwiche, A. (2005). On the revision of probabilistic beliefs using uncertain evidence. *Artificial Intelligence*, 163(1), 67–90.
- Chavira, M., & Darwiche, A. (2007). Compiling Bayesian Networks Using Variable Elimination. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07)* (p. 2443-2449). Hyderabad, India.
- Choi, A., Darwiche, A., Zheng, L., & Mengshoel, O. J. (2011). Data Mining in Systems Health Management: Detection, Diagnostics, and Prognostics. In A. Srivastava & J. Han (Eds.), (chap. A Tutorial on Bayesian Networks for System Health Management). Chapman and Hall/CRC Press.
- Darwiche, A. (2003). A Differential Approach to Inference in Bayesian Networks. *Journal of the ACM*, 50(3), 280–305.
- Darwiche, A. (2009). *Modeling and Reasoning with Bayesian Networks*. Cambridge, UK: Cambridge University Press.
- Dechter, R. (1999). Bucket Elimination: A Unifying Framework for Reasoning. *Artificial Intelligence*, 113(1-2), 41-85. Available from citeseer.nj.nec.com/article/dechter99bucket.html
- Kozlov, A., & Koller, D. (1997). Nonuniform Dynamic Discretization in Hybrid Networks. In *In Proc. UAI* (pp. 314–325). Morgan Kaufmann.
- Kurtoglu, T., Feldman, A., Poll, S., deKleer, J., Narasimhan, S., Garcia, D., et al. (2010). *Second International Diagnostic Competition (DXC10): Industrial Track Diagnostic Problem Descriptions* (Tech. Rep.). NASA ARC and PARC. Available from <http://www.phmsociety.org/competition/dxc/10/files>
- Kurtoglu, T., Narasimhan, S., Poll, S., Garcia, D., Kuhn, L., deKleer, J., et al. (2009, June). First International Diagnosis Competition - DXC'09. In *Proc. of the Twentieth International Workshop on Principles of Diagnosis (DX'09)* (pp. 383–396). Stockholm, Sweden.
- Langseth, H., Nielsen, T. D., Rumí, R., & Salmeron, A. (2009). Inference in hybrid Bayesian networks. *Reliability Engineering & System Safety*, 94(10), 1499–1509.
- Lauritzen, S., & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society series B*, 50(2), 157–224.
- Lerner, U., Parr, R., Koller, D., & Biswas, G. (2000). Bayesian Fault Detection and Diagnosis in Dynamic Systems. In *Proceedings of the Seventeenth national Conference on Artificial Intelligence (AAAI-00)* (p. 531-537). Available from citeseer.ist.psu.edu/lerner00bayesian.html
- Mengshoel, O. J. (2007). Designing Resource-Bounded Reasoners using Bayesian Networks: System Health Monitoring and Diagnosis. In *Proceedings of the 18th International Workshop on Principles of Diagnosis (DX-07)* (pp. 330–337). Nashville, TN.
- Mengshoel, O. J., Chavira, M., Cascio, K., Poll, S., Darwiche, A., & Uckun, S. (2010). Probabilistic Model-Based Diagnosis: An Electrical Power System Case Study. *IEEE Trans. on Systems, Man, and Cybernetics*, 40(5), 874–885.
- Mengshoel, O. J., Poll, S., & Kurtoglu, T. (2009). Developing Large-Scale Bayesian Networks by Composition: Fault Diagnosis of Electrical Power Systems in Aircraft and Spacecraft. In *Proc. of the IJCAI-09 Workshop on Self- and Autonomous Systems (SAS): Reasoning and Integration Challenges*.
- Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41, 100 - 115.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.
- Poll, S., Patterson-Hine, A., Camisa, J., Garcia, D., Hall, D., Lee, C., et al. (2007). Advanced Diagnostics and Prognostics Testbed. In *Proceedings of the 18th International Workshop on Principles of Diagnosis (DX-07)* (pp. 178–185). Nashville, TN.
- Ricks, B. W., & Mengshoel, O. J. (2009a). The Diagnostic

Challenge Competition: Probabilistic Techniques for Fault Diagnosis in Electrical Power Systems. In *Proc. of the 20th International Workshop on Principles of Diagnosis (DX-09)*. Stockholm, Sweden.

Ricks, B. W., & Mengshoel, O. J. (2009b). Methods for Probabilistic Fault Diagnosis: An Electrical Power System Case Study. In *Proc. of Annual Conference of the PHM Society, 2009 (PHM-09)*. San Diego, CA.

Ricks, B. W., & Mengshoel, O. J. (2010). Diagnosing Intermittent and Persistent Faults using Static Bayesian Networks. In *Proc. of the 21st International Workshop on Principles of Diagnosis (DX-10)*. Portland, OR.



Brian Ricks received the Bachelor's of Science degree in Computer Science from the University of Texas at Dallas, in 2010.

He is currently a graduate student at the University of Texas at Dallas, majoring in Computer Science. He previously worked at the NASA Ames Research Center, Intelligent Systems Division, as an intern with

Carnegie Mellon University - Silicon Valley, and also at the NASA Ames Research Center as an intern with the Universities Space Research Program. He will graduate in Spring of 2012 with a Master's in Computer Science. Mr. Ricks performed part of the research reported here during both prior internships, under the leadership of Dr. Ole J. Mengshoel.



Ole J. Mengshoel received the B.S. degree from the Norwegian Institute of Technology, Trondheim, Norway, in 1989, and the Ph.D. degree from the University of Illinois at Urbana-Champaign, Illinois, in 1999, both in computer science.

He is currently a Senior Systems Scientist with Carnegie Mellon University (CMU), Silicon Valley, CA, and affiliated with the Intelligent Systems Division, National Aeronautics and Space Administration (NASA) Ames Research Center, Moffett Field, CA. Prior to joining CMU, he was a Senior Scientist and Research Area Lead with USRA/RIACS and a Research Scientist with the Decision Sciences Group, Rockwell Scientific, and Knowledge-Based Systems, SINTEF, Norway. His current research focuses on reasoning, machine learning, diagnosis, prognosis, and decision support under uncertainty – often using Bayesian networks and with aerospace applications of interest to NASA. He has published more than 50 papers and papers in journals, conference proceedings, and workshops. He is the holder of four U.S. patents.

Dr. Mengshoel is a member of the Association for the Advancement of Artificial Intelligence, the Association for Computer Machinery, and IEEE.

Investigating the Effect of Damage Progression Model Choice on Prognostics Performance

Matthew Daigle¹ Indranil Roychoudhury² Sriram Narasimhan¹ Sankalita Saha³ Bhaskar Saha³ and Kai Goebel⁴

¹ *University of California, Santa Cruz, NASA Ames Research Center, Moffett Field, CA, 94035, USA*
matthew.j.daigle@nasa.gov, sriram.narasimhan-1@nasa.gov

² *SGT, Inc., NASA Ames Research Center, Moffett Field, CA, 94035, USA*
indranil.roychoudhury@nasa.gov

³ *MCT, Inc., NASA Ames Research Center, Moffett Field, CA, 94035, USA*
bhaskar.saha@nasa.gov, sankalita.saha-1@nasa.gov

⁴ *NASA Ames Research Center, Moffett Field, CA, 94035, USA*
kai.goebel@nasa.gov

ABSTRACT

The success of model-based approaches to systems health management depends largely on the quality of the underlying models. In model-based prognostics, it is especially the quality of the damage progression models, i.e., the models describing how damage evolves as the system operates, that determines the accuracy and precision of remaining useful life predictions. Several common forms of these models are generally assumed in the literature, but are often not supported by physical evidence or physics-based analysis. In this paper, using a centrifugal pump as a case study, we develop different damage progression models. In simulation, we investigate how model changes influence prognostics performance. Results demonstrate that, in some cases, simple damage progression models are sufficient. But, in general, the results show a clear need for damage progression models that are accurate over long time horizons under varied loading conditions.

1. INTRODUCTION

Model-based prognostics is rooted in the use of models that describe the behavior of systems and components and how that behavior changes as wear and damage processes occur (Luo, Pattipati, Qiao, & Chigusa, 2008; Saha & Goebel, 2009; Daigle & Goebel, 2011). The problem of model-based prognostics fundamentally consists of two sequential problems, (i) a joint state-parameter estimation problem, in which, using the model, the health of a system or component is determined based on its observations; and (ii) a prediction problem, in which, using the model, the state-parameter distribution is simulated forward in time to compute *end of life* (EOL) and *remaining useful life* (RUL). The model must describe both how damage manifests in the system observations,

and how damage progresses in time. Clearly, the prognostics performance inherently depends on the quality of the models used by the algorithms.

In modeling the complex engineering systems targeted by prognostics algorithms, many modeling choices must be made. In particular, one must decide on the appropriate level of abstraction at which to model the system in order to estimate system health and predict remaining life. The choice is mainly one of model granularity, i.e., the extent to which the model is broken down into parts, either structural or behavioral. The selected models must then provide enough fidelity to meet the prognostics performance requirements. But, model development cost, available level of expertise, model validation effort, and computational complexity all constrain the models that may be developed. For example, finer-grained models may result in increased model fidelity and thus increased prognostics performance, but may take more effort to construct and increase computational complexity. Therefore, a clear need exists to investigate the impact of such modeling choices on prognostics performance.

In this paper, we use a centrifugal pump as a case study with which to explore the impact of model quality on prognostics performance. Typically, developing a reliable model of nominal system operation is relatively straightforward, as the dynamics are usually well-understood in terms of first principles or physics equations, and, most importantly, there is typically sufficient data available with which to validate this model. The major difficulty lies in developing models of damage progression, because these models are often component-dependent, and so the understanding of these processes is generally lacking. Further, the data necessary to properly validate these models are, in practice, rarely available. Using the pump model, we develop several damage progression models and evaluate their effect on prognostics performance using simulation-based experiments. To the best of our knowledge, this, along with a companion paper exploring these issues

Daigle et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

with application to battery health management (Saha, Quach, & Goebel, 2011), is the first time this type of analysis has been performed within the context of prognostics.

The paper is organized as follows. Section 2 describes the model-based prognostics framework. Section 3 presents the modeling methodology and develops the centrifugal pump model with several damage progression models. Section 4 generalizes the different models within the framework of model abstraction. Section 5 describes the particle filter-based damage estimation method, and Section 6 discusses the prediction methodology. Section 7 provides results from a number of simulation-based experiments and evaluates the effect of the different damage progression models on prognostics performance. Section 8 concludes the paper.

2. MODEL-BASED PROGNOSTICS

We assume the system model may be described using

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{f}(t, \mathbf{x}(t), \boldsymbol{\theta}(t), \mathbf{u}(t), \mathbf{v}(t)) \\ \mathbf{y}(t) &= \mathbf{h}(t, \mathbf{x}(t), \boldsymbol{\theta}(t), \mathbf{u}(t), \mathbf{n}(t)),\end{aligned}$$

where $\mathbf{x}(t) \in \mathbb{R}^{n_x}$ is the state vector, $\boldsymbol{\theta}(t) \in \mathbb{R}^{n_\theta}$ is the parameter vector, $\mathbf{u}(t) \in \mathbb{R}^{n_u}$ is the input vector, $\mathbf{v}(t) \in \mathbb{R}^{n_v}$ is the process noise vector, \mathbf{f} is the state equation, $\mathbf{y}(t) \in \mathbb{R}^{n_y}$ is the output vector, $\mathbf{n}(t) \in \mathbb{R}^{n_n}$ is the measurement noise vector, and \mathbf{h} is the output equation. The model may be nonlinear with no restrictions on the functional forms of \mathbf{f} or \mathbf{h} , and the noise terms may be nonlinearly coupled with the states and parameters. The parameters $\boldsymbol{\theta}(t)$ evolve in an unknown way.

The goal of prognostics is to predict EOL (and/or RUL) at a given time point t_P using the discrete sequence of observations up to time t_P , denoted as $\mathbf{y}_{0:t_P}$. EOL is defined as the time point at which the component no longer meets a functional or performance requirement. In general, these requirements do not need to be directly tied to permanent failure, rather, they refer to a state of the system that is undesirable. The system can leave this state through repair or other actions, and sometimes no action is needed and the component needs only to rest (e.g., with power electronics, or self-recharge of batteries). These functional requirements may be expressed through a threshold, beyond which the component is considered to have failed. In general, we may express this threshold as a function of the system state and parameters, $T_{EOL}(\mathbf{x}(t), \boldsymbol{\theta}(t))$, where $T_{EOL}(\mathbf{x}(t), \boldsymbol{\theta}(t)) = 1$ if a requirement is violated, and 0 otherwise.

So, EOL may be defined as

$$EOL(t_P) \triangleq \inf\{t \in \mathbb{R} : t \geq t_P \wedge T_{EOL}(\mathbf{x}(t), \boldsymbol{\theta}(t)) = 1\},$$

i.e., EOL is the earliest time point at which the threshold is reached. RUL may then be defined with

$$RUL(t_P) \triangleq EOL(t_P) - t_P.$$

Due to various sources of uncertainty, including uncertainty in the model, the goal is to compute a probability distribution of

the EOL or RUL. We compute, at time t_P , $p(EOL(t_P)|\mathbf{y}_{0:t_P})$ or $p(RUL(t_P)|\mathbf{y}_{0:t_P})$.

In model-based prognostics, there are two fundamental problems: (i) joint state-parameter estimation, and (ii) prediction. In discrete time k , we estimate \mathbf{x}_k and $\boldsymbol{\theta}_k$, and use these estimates to predict EOL and RUL at desired time points. The model-based prognostics architecture is shown in Fig. 1 (Daigle & Goebel, 2011). Given inputs \mathbf{u}_k , the system provides measured outputs \mathbf{y}_k . If available, a fault detection, isolation, and identification (FDII) module may be used to determine which damage mechanisms are active, represented as a fault set \mathbf{F} . The damage estimation module may use this result to limit the dimension of the estimation problem. It determines estimates of the states and unknown parameters, represented as a probability distribution $p(\mathbf{x}_k, \boldsymbol{\theta}_k|\mathbf{y}_{0:k})$. The prediction module uses the joint state-parameter distribution, along with hypothesized future inputs, to compute EOL and RUL as probability distributions $p(EOL_{k_P}|\mathbf{y}_{0:k_P})$ and $p(RUL_{k_P}|\mathbf{y}_{0:k_P})$ at given prediction times k_P . In this paper, we assume a solution to FDII that provides us with the single active damage mechanism, initiating prognostics.

Prognostics performance is evaluated based on the accuracy and precision of the predictions. We use the relative accuracy (RA) metric (Saxena, Celaya, Saha, Saha, & Goebel, 2010) to characterize prediction accuracy. For a given prediction time k_P , RA is defined as

$$RA_{k_P} = 100 \left(1 - \frac{|RUL_{k_P}^* - \widehat{RUL}_{k_P}|}{RUL_{k_P}^*} \right),$$

where $RUL_{k_P}^*$ is the true RUL at time k_P , and \widehat{RUL}_{k_P} is the mean of the prediction. The prognostic horizon (PH) refers to the time between EOL and the first prediction that meets some accuracy requirement RA^* (e.g., 90%):

$$PH = 100 \frac{EOL^* - \min\{k_P : RA_{k_P} \geq RA^*\}}{EOL^*},$$

where EOL^* denotes the true EOL. A larger value means an accurate prediction is available earlier. This is a version of the PH metric given in (Saxena et al., 2010) normalized to EOL. Prediction spread is computed using relative median absolute deviation (RMAD):

$$RMAD(X) = 100 \frac{\text{Median}_i (|X_i - \text{Median}_j (X_j)|)}{\text{Median}_j (X_j)},$$

where X is a data set and X_i is an element of that set.

3. PUMP MODELING

In our modeling methodology, we first describe a nominal model of system behavior. We then extend the model by including *damage progression functions* within the state equation \mathbf{f} that describe how *damage variables* $\mathbf{d}(t) \subseteq \mathbf{x}(t)$ evolve over time. The damage progression functions are parameterized by unknown *wear parameters* $\mathbf{w}(t) \subseteq \boldsymbol{\theta}(t)$. We use

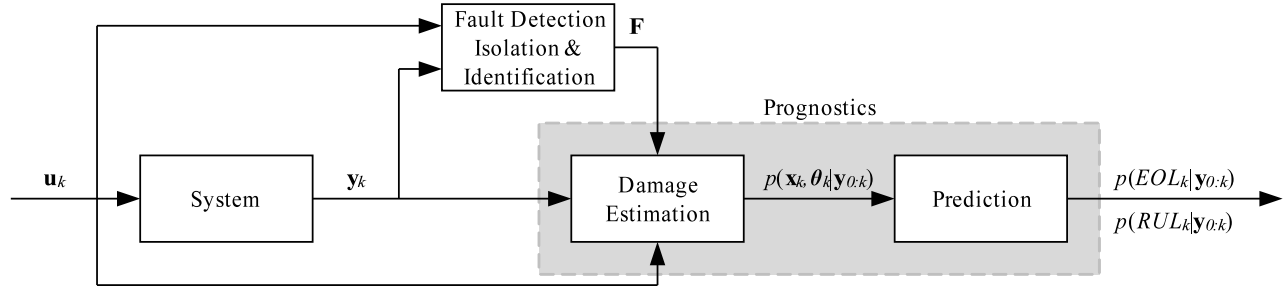


Figure 1. Prognostics architecture.

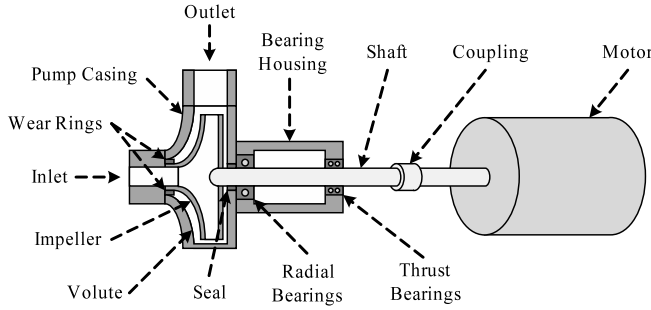


Figure 2. Centrifugal pump.

a centrifugal pump as a case study. In this section, we first describe the nominal model of the pump, and then describe common damage progression models.

3.1 Nominal Model

A schematic of a typical centrifugal pump is shown in Fig. 2. Fluid enters the inlet, and the rotation of the impeller, driven by an electric motor, forces fluid through the outlet. The radial and thrust bearings help to minimize friction along the pump shaft. The bearing housing contains oil which lubricates the bearings. A seal prevents fluid flow into the bearing housing. Wear rings prevent internal pump leakage from the outlet to the inlet side of the impeller, but a small clearance is typically allowed to minimize friction. The nominal pump model has been described previously in (Daigle & Goebel, 2011), and we review it here for completeness.

The state of the pump is given by

$$\mathbf{x}(t) = [\omega(t) \quad T_t(t) \quad T_r(t) \quad T_o(t)]^T,$$

where $\omega(t)$ is the rotational velocity of the pump, $T_t(t)$ is the thrust bearing temperature, $T_r(t)$ is the radial bearing temperature, and $T_o(t)$ is the oil temperature.

The rotational velocity of the pump is described using a torque balance,

$$\dot{\omega} = \frac{1}{J} (\tau_e(t) - r\omega(t) - \tau_L(t)),$$

where J is the lumped motor/pump inertia, τ_e is the electromagnetic torque provided by the motor, r is the lumped fric-

tion parameter, and τ_L is the load torque. In an induction motor, a voltage is applied to the stator, which creates a current through the stator coils. A polyphase voltage applied to the stator creates a rotating magnetic field that induces a current in the rotor, causing it to turn. The torque produced on the rotor is nonzero only when there is a difference between the synchronous speed of the supply voltage, ω_s and the mechanical rotation, ω . This *slip* is defined as

$$s = \frac{\omega_s - \omega}{\omega_s}.$$

The expression for the torque τ_e is derived from an equivalent circuit representation for the three-phase induction motor based on rotor and stator resistances and inductances, and the slip s (Lyshevski, 1999):

$$\tau_e = \frac{npR_2}{s\omega_s} \frac{V_{rms}^2}{(R_1 + R_2/s)^2 + (\omega_s L_1 + \omega_s L_2)^2},$$

where R_1 is the stator resistance, L_1 is the stator inductance, R_2 is the rotor resistance, L_2 is the rotor inductance, n is the number of phases (typically 3), and p is the number of magnetic pole pairs. The dependence of torque on slip creates a feedback loop that causes the rotor to follow the rotation of the magnetic field. The rotor speed may be controlled by changing the input frequency ω_s .

The load torque τ_L is a polynomial function of the pump flow rate and the impeller rotational velocity (Wolfram, Fussel, Brune, & Isermann, 2001; Kallesøe, 2005):

$$\tau_L = a_0\omega^2 + a_1\omega Q - a_2Q^2,$$

where Q is the flow, and a_0 , a_1 , and a_2 are coefficients derived from the pump geometry (Kallesøe, 2005).

The rotation of the impeller creates a pressure difference from the inlet to the outlet of the pump, which drives the pump flow, Q . The pump pressure is computed as

$$p_p = A\omega^2 + b_1\omega Q - b_2Q^2,$$

where A is the impeller area, and b_1 and b_2 are coefficients derived from the pump geometry. The discharge flow, Q , is comprised of the flow through the impeller, Q_i , and a leakage flow, Q_l :

$$Q = Q_i - Q_l.$$

The flow through the impeller is computed using the pressure differences:

$$Q_i = c\sqrt{|p_s + p_p - p_d| \text{sign}(p_s + p_p - p_d)},$$

where c is a flow coefficient, p_s is the suction pressure, and p_d is the discharge pressure. The small (normal) leakage flow from the discharge end to the suction end due to the clearance between the wear rings and the impeller is described by

$$Q_l = c_l\sqrt{|p_d - p_s| \text{sign}(p_d - p_s)},$$

where c_l is a flow coefficient.

Pump temperatures are often monitored as indicators of pump condition. The oil heats up due to the radial and thrust bearings and cools to the environment:

$$\dot{T}_o = \frac{1}{J_o} (H_{o,1}(T_t - T_o) + H_{o,2}(T_r - T_o) - H_{o,3}(T_o - T_a)),$$

where J_o is the thermal inertia of the oil, and the $H_{o,i}$ terms are heat transfer coefficients. The thrust bearings heat up due to the friction between the pump shaft and the bearings, and cool to the oil and the environment:

$$\dot{T}_t = \frac{1}{J_t} (r_t\omega^2 - H_{t,1}(T_t - T_o) - H_{t,2}(T_t - T_a)),$$

where J_t is the thermal inertia of the thrust bearings, r_t is the friction coefficient for the thrust bearings, and the $H_{t,i}$ terms are heat transfer coefficients. The radial bearings behave similarly:

$$\dot{T}_r = \frac{1}{J_r} (r_r\omega^2 - H_{r,1}(T_r - T_o) - H_{r,2}(T_r - T_a))$$

where J_r is the thermal inertia of the radial bearings, r_r is the friction coefficient for the radial bearings, and the $H_{r,i}$ terms are heat transfer coefficients.

The overall input vector \mathbf{u} is given by

$$\mathbf{u}(t) = [p_s(t) \quad p_d(t) \quad T_a(t) \quad V(t) \quad \omega_s(t)]^T.$$

The measurement vector \mathbf{y} is given by

$$\mathbf{y}(t) = [\omega(t) \quad Q(t) \quad T_t(t) \quad T_r(t) \quad T_o(t)]^T.$$

Fig. 3 shows nominal pump operation. Input voltage and line frequency are varied to control the pump speed. Initially, slip is 1, and this produces an electromagnetic torque that causes the rotation of the motor to match the rotation of the magnetic field, with a small amount of slip remaining (depending on the load). Fluid flows through the pump due to the impeller rotation. The bearings heat and cool as the pump rotation increases and decreases.

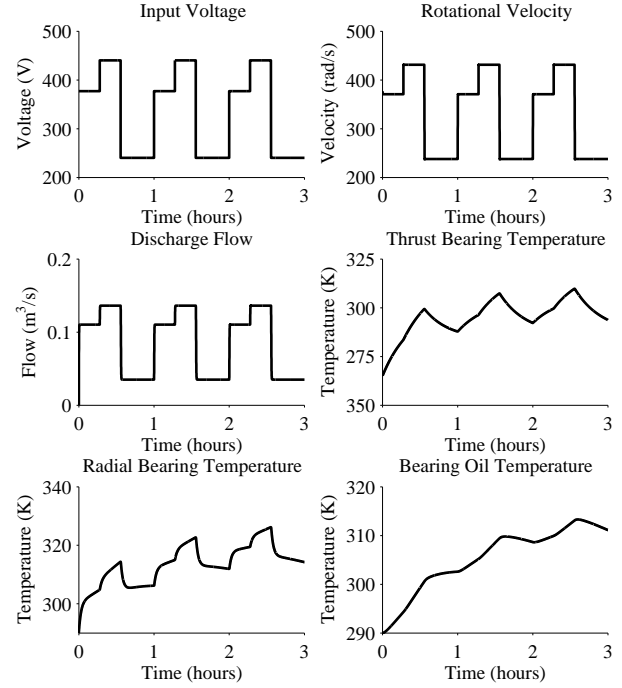


Figure 3. Nominal pump operation.

3.2 Damage Modeling

The most significant forms of damage for pumps are impeller wear, caused by cavitation and erosion by the flow, and bearing failure, caused by friction-induced wear of the bearings. In each case, we map the damage to a particular parameter in the nominal model, and this parameter becomes a damage variable in $\mathbf{d}(t)$ that evolves by a damage progression function. Several types of damage progression models have been explored in literature. In this paper, we focus on macro-level, lumped-parameter models. Within this modeling style, damage evolves as a function of dynamic energy-related variables. Several common forms may be assumed here, including linear, polynomial, and exponential, as these forms have been observed in practice. We derive these forms for the considered damage modes as well as wear-based models based on physics analysis.

Impeller wear is represented as a decrease in impeller area A (Biswas & Mahadevan, 2007; Tu et al., 2007; Daigle & Goebel, 2011). Impeller wear can only progress when flow through the impeller, Q_i , is nonzero. So, the rate of change of impeller area, \dot{A} , must be a function of Q_i . We consider the following damage progression models based on the common observed forms:

$$\dot{A} = -w_A Q_i \quad (1)$$

$$\dot{A} = -w_A Q_i^2 \quad (2)$$

$$\dot{A} = -w_{A1} Q_i - w_{A2} Q_i^2 \quad (3)$$

$$\dot{A} = -w_{A1} \exp(w_{A2} Q_i), \quad (4)$$

where w_A , w_{A1} , and w_{A2} are unknown wear parameters.

From a physics analysis, we see that the erosive wear equation applies here (Hutchings, 1992). The erosive wear rate is proportional to fluid velocity times friction force. Fluid velocity is proportional to volumetric flow rate, and friction force is proportional to fluid velocity, so, lumping the proportionality constants into the wear coefficient w_A , we obtain

$$\dot{A} = -w_A Q_i^2. \quad (5)$$

Note that this agrees with one of the commonly assumed damage forms, equation 2, above.

A decrease in the impeller area will decrease the pump pressure, which, in turn, reduces the delivered flow, and, therefore, pump efficiency. The pump must operate at a certain minimal efficiency. This requirement defines an EOL criteria. We define A^- as the minimum value of the impeller area at which this requirement is met, hence, $T_{EOL} = 1$ if $A(t) < A^-$.

The damage progression up to EOL for impeller wear is shown in Fig. 4a for equation 5, for the rotational velocity alternating between 3600 RPM for the first half of every hour of usage and 4300 RPM for the second half, causing the pump flow to alternate as well. Within a given cycle, shown in the inset of Fig. 4a, the damage progresses at two different rates, but over a long time horizon, the damage progression appears fairly linear. This suggests that a linear approximation may suffice for accurate long-term predictions if the future inputs cycle in the same way. The damage progression rate actually decreases slightly over time, because as impeller area decreases, flow will decrease, and therefore \dot{A} will diminish.

Bearing wear is captured as an increase in the corresponding friction coefficient (Daigle & Goebel, 2011). Bearing wear can only occur when the pump is rotating, i.e., ω is nonzero. So, the rate of change of the bearing friction coefficient, \dot{r}_t for the thrust bearing, and \dot{r}_r for the radial bearing, must be a function of ω . For the thrust bearing wear, we consider the following damage progression models based on the common observed forms:

$$\dot{r}_t(t) = w_t \omega \quad (6)$$

$$\dot{r}_t(t) = w_t \omega^2 \quad (7)$$

$$\dot{r}_t(t) = w_{t1} \omega + w_{t2} \omega^2 \quad (8)$$

$$\dot{r}_t(t) = w_{t1} \exp(w_{t2} \omega), \quad (9)$$

where w_t , w_{t1} , and w_{t2} are unknown wear parameters. For the radial bearing, the equations are the same, but with the t subscript replaced by an r subscript:

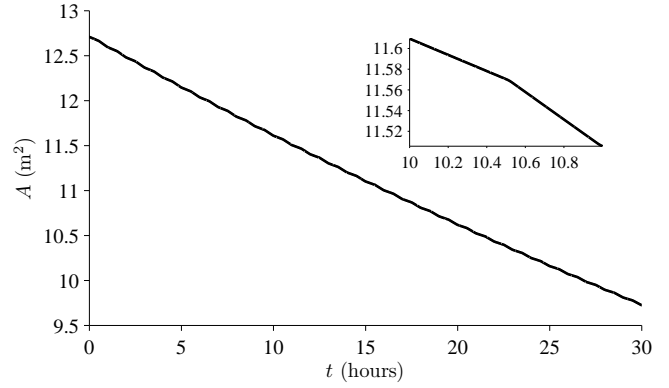
$$\dot{r}_r(t) = w_r \omega \quad (10)$$

$$\dot{r}_r(t) = w_r \omega^2 \quad (11)$$

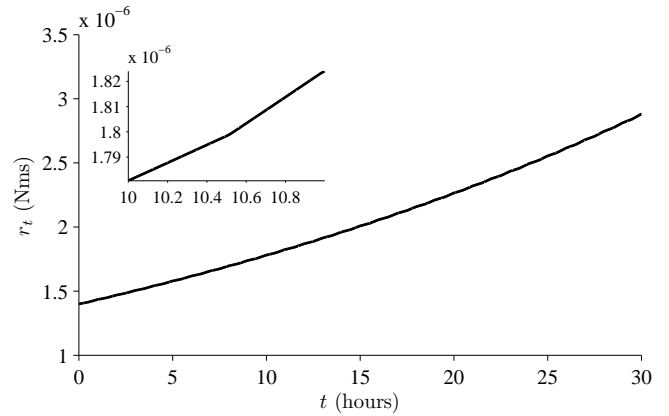
$$\dot{r}_r(t) = w_{r1} \omega + w_{r2} \omega^2 \quad (12)$$

$$\dot{r}_r(t) = w_{r1} \exp(w_{r2} \omega). \quad (13)$$

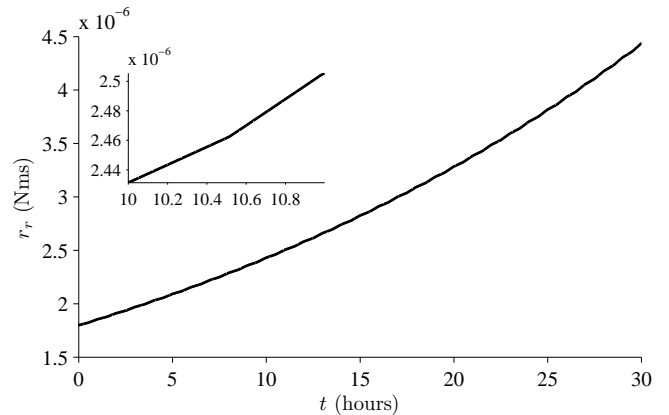
From a physics analysis, we observe that sliding and rolling



(a) Damage progression for impeller wear.



(b) Damage progression for thrust bearing wear.



(c) Damage progression for radial bearing wear.

Figure 4. Damage progression for the pump.

friction generate wear of material which increases the coefficient of friction (Hutchings, 1992; Daigle & Goebel, 2010):

$$\dot{r}_t(t) = w_t r_t \omega^2 \quad (14)$$

$$\dot{r}_r(t) = w_r r_r \omega^2, \quad (15)$$

where w_t and w_r are the wear parameters. Note that equations 6–9 neglect the direct relationship between \dot{r}_t and r_t .

Changes in bearing friction can be observed by means of the bearing temperatures. Limits on the maximum values of these temperatures define EOL for bearing wear. We define r_t^+ and r_r^+ as the maximum permissible values of the friction coefficients, before the temperature limits are exceeded over a typical usage cycle. So, $T_{EOL} = 1$ if $r_t(t) > r_t^+$ or $r_r(t) > r_r^+$. Damage progression up to EOL for bearing wear is shown in Figs. 4b and 4c, for equations 14 and 15, with the rotational velocity again alternating between 3600 RPM and 4300 RPM. In this case, the rate of damage progression increases over time. Therefore, a simple linear approximation would not be accurate. This behavior occurs because $\dot{r}_t(t)$ increases with $r_t(t)$, and $\dot{r}_r(t)$ increases with $r_r(t)$.

4. MODEL ABSTRACTION

The previous section presented a number of different models. In general, these differences may be captured by the idea of *model abstraction* (Frantz, 1995; Lee & Fishwick, 1996; Zeigler, Praehofer, & Kim, 2000). Abstraction is driven by the questions that the model must address. For prognostics, the models must address the question of the EOL/RUL of a system. In order to do this, the models must (i) describe how damage manifests in the system outputs (i.e., measured variables or computed features), so that damage estimation can be performed; and (ii) describe how damage evolves in time as a function of the system loading, so that prediction can be performed. The chosen level of model abstraction must be such that these tasks can be accomplished at the desired level of performance.

Abstraction is a process of simplification. Common abstractions include aggregation, omission, linearization, deterministic/stochastic replacement, and formalism transformation (e.g., differential equations to discrete-event systems) (Zeigler et al., 2000). These abstractions may manifest as *structural abstraction*, in which the model is abstracted by its structure, or *behavioral abstraction*, in which the model is abstracted by its behaviors (Lee & Fishwick, 1996). For example, a structural abstraction might ignore the individual circuit elements of an electric motor and aggregate them into a lumped component. A behavioral abstraction might omit the individual processes and effects comprising a damage progression process and instead consider their lumped effects. Or, perhaps a given process might really take on an exponential form, but is abstracted to a linear form. The linear form consists of a simpler relationship that is described by fewer free parameters.

Model granularity is a particular measure of model abstraction. The *granularity* of a model is the extent to which it is divided into smaller parts. The concept of granularity does not address the degree of complexity of the specific functional relationships within a part of the model. Granularity can manifest both structurally and behaviorally. For example, a lumped parameter model is coarser-grained than a fi-

nite element model. In the context of physics-based prognostics models, a model with fine granularity may include more lower-level physical processes (e.g., micro-level effects rather than macro-level effects), or model processes at a greater level of detail, than a model with coarse granularity.

In quantifiable terms, granularity may be expressed using the number of state variables, the number of relationships between them, and the number of free (unknown) parameters. By definition, the state variables are the minimal set of variables needed to describe the state of the system as it progresses through time. So a finer-grained model may entail an additional number of state variables because aspects of the physical description that were not captured before are now described. With the same state variables, a model may also become more granular by adding functional relationships between the state variables. In a linear system, with $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$, this would correspond to zeros in the \mathbf{A} matrix becoming nonzero. Note that this is only a fair comparison between two models capturing the same process.

The different damage models developed in Section 3.2 can be viewed within this framework. For a particular damage mode, the different damage models each capture the same physical process, i.e., the damage progression, but make different assumptions about the complexity of the process. Thus, these models capture damage progression at different levels of behavioral abstraction. For example, for the impeller wear, the polynomial form (equation 3) may be viewed as less abstract than both the linear (equation 1) and squared forms (equation 2), because it is a sum of these individual processes. For the bearing wear, equations 6–9 are all coarser-grained models than 14, because they neglect the direct relationship between \dot{r}_t and r_t .

One may describe the system behavior in very low-level physical relationships, but, of course, there are trade-offs to be made among the modeling constraints. A finer-grained model takes more effort to develop and validate, and may result in an increased computational cost. It also may result in an increase in the number of free parameters, which increases the complexity of the joint state-parameter estimation problem. The increase in model development cost to create models with finer granularity is justified only when it results in an appropriate increase in fidelity (i.e., the extent to which a model reproduces the observable behaviors of the system being modeled) and a corresponding increase in prognostics performance. Also, higher levels of abstraction make sense when the computation associated with lower levels of abstraction becomes too complicated for practical implementation. Requirements on prognostics performance and constraints on model size, development cost, level of modeling expertise, and computational complexity all drive the model development process.

5. DAMAGE ESTIMATION

Damage estimation is fundamentally a joint state-parameter estimation problem, i.e., computation of $p(\mathbf{x}_k, \boldsymbol{\theta}_k | \mathbf{y}_{0:k})$. The damage states and wear parameters must be estimated along with the other state variables and unknown parameters of the system. We use the *particle filter* (Arulampalam, Maskell, Gordon, & Clapp, 2002) as a general solution to this problem.

In a particle filter, the state distribution is approximated by a set of discrete weighted samples, or *particles*:

$$\{(\mathbf{x}_k^i, \boldsymbol{\theta}_k^i), w_k^i\}_{i=1}^N,$$

where N denotes the number of particles, and for particle i , \mathbf{x}_k^i denotes the state vector estimate, $\boldsymbol{\theta}_k^i$ denotes the parameter vector estimate, and w_k^i denotes the weight. The posterior density is approximated by

$$p(\mathbf{x}_k, \boldsymbol{\theta}_k | \mathbf{y}_{0:k}) \approx \sum_{i=1}^N w_k^i \delta_{(\mathbf{x}_k^i, \boldsymbol{\theta}_k^i)}(d\mathbf{x}_k d\boldsymbol{\theta}_k),$$

where $\delta_{(\mathbf{x}_k^i, \boldsymbol{\theta}_k^i)}(d\mathbf{x}_k d\boldsymbol{\theta}_k)$ denotes the Dirac delta function located at $(\mathbf{x}_k^i, \boldsymbol{\theta}_k^i)$.

We use the sampling importance resampling (SIR) particle filter. Each particle is propagated forward to time k by first sampling new parameter values, and then sampling new states using the model. The particle weight is assigned using \mathbf{y}_k . The weights are then normalized, followed by the resampling step. Pseudocode is given in (Arulampalam et al., 2002; Daigle & Goebel, 2011).

Parameter values are sampled using a random walk, i.e., for parameter θ , $\theta_k = \theta_{k-1} + \xi_{k-1}$, where ξ_{k-1} is sampled from some distribution. Particles generated with parameter values closest to the true values should be assigned higher weight and allow the particle filter to converge to the true values. The random walk variance is modified dynamically online to maintain a user-specified relative spread of the unknown wear parameters using the variance control algorithm presented in (Daigle & Goebel, 2011). The algorithm increases or decreases the random walk variance proportional to the difference between the desired spread and the actual spread, computed with relative median absolute deviation (RMAD). The algorithm behavior is specified using four parameters: the desired spread during the initial convergence period, v_0^* (e.g., 50%), the threshold that specifies the end of the convergence period, T (e.g., 60%), the final desired spread v_∞^* (e.g., 10%), and the proportional gain P (e.g. 1×10^{-3}). The spread is first controlled to v_0^* until the spread reaches T , at which point it is controlled to v_∞^* .

6. PREDICTION

Given the current joint state-parameter estimate at a desired prediction time k_P , $p(\mathbf{x}_{k_P}, \boldsymbol{\theta}_{k_P} | \mathbf{y}_{0:k_P})$, the prediction step

computes $p(EOL_{k_P} | \mathbf{y}_{0:k_P})$ and $p(RUL_{k_P} | \mathbf{y}_{0:k_P})$. The particle filter provides

$$p(\mathbf{x}_{k_P}, \boldsymbol{\theta}_{k_P} | \mathbf{y}_{0:k_P}) \approx \sum_{i=1}^N w_{k_P}^i \delta_{(\mathbf{x}_{k_P}^i, \boldsymbol{\theta}_{k_P}^i)}(d\mathbf{x}_{k_P} d\boldsymbol{\theta}_{k_P}).$$

We approximate a prediction distribution n steps forward as (Doucet, Godsill, & Andrieu, 2000)

$$p(\mathbf{x}_{k_P+n}, \boldsymbol{\theta}_{k_P+n} | \mathbf{y}_{0:k_P}) \approx \sum_{i=1}^N w_{k_P}^i \delta_{(\mathbf{x}_{k_P+n}^i, \boldsymbol{\theta}_{k_P+n}^i)}(d\mathbf{x}_{k_P+n} d\boldsymbol{\theta}_{k_P+n}).$$

Similarly, we approximate the EOL as

$$p(EOL_{k_P} | \mathbf{y}_{0:k_P}) \approx \sum_{i=1}^N w_{k_P}^i \delta_{EOL_{k_P}^i}(dEOL_{k_P}).$$

To compute EOL, then, we propagate each particle forward to its own EOL and use that particle's weight at k_P for the weight of its EOL prediction. The prediction is made using hypothesized future inputs of the system. In this work, we assume these inputs are known in advance. Pseudocode for the prediction algorithm is given in (Daigle & Goebel, 2011).

7. RESULTS

We ran a number of simulation experiments for the different pump models in order to evaluate the relative performance. We took the damage models using the physics-based wear equations as the reference models that generated the measurement data. The model used by the prognostics algorithm was either the reference model \mathcal{M} (using equations 5, 14, and 15), the linear model \mathcal{M}_{Linear} (using equations 1, 6, and 10), the squared model $\mathcal{M}_{Squared}$ (using equations 2, 7, and 11), the second order polynomial model \mathcal{M}_{Poly} (using equations 3, 8, and 12), or the exponential model \mathcal{M}_{Exp} (using equations 4, 9, and 13). In each experiment, the pump speed cycled from 3600 RPM for the first half of every hour of usage to 4300 RPM for the second half hour.

In order to analyze results on a per-damage mode basis, in each experiment we assumed only a single damage mode was active. We selected the reference model's wear parameter values randomly in each experiment, within $[0.5 \times 10^{-3}, 4 \times 10^{-3}]$ for w_A , in $[0.5 \times 10^{-11}, 7 \times 10^{-11}]$ for w_t and w_r , such that the maximum wear rates corresponded to a minimum EOL of 20 hours. The particle filters had to estimate the states and the wear parameters associated with their assumed damage progression models. We considered the case where the future input was known in order to focus on the differences in performance based on the different assumed damage models. We also varied the process noise variance from 0, to nominal, and 10 times nominal, in order to artificially represent the nominal model at various levels of granularity.

Model	\mathbf{v}	$\overline{\text{RA}}$	$\overline{\text{RMAD}}_{RUL}$
\mathcal{M}	0	97.87	10.33
	1	97.42	10.30
	10	97.63	10.41
\mathcal{M}_{Linear}	0	94.12	10.42
	1	92.28	10.91
	10	83.68	12.42
\mathcal{M}_{Poly}	0	97.55	3.35
	1	96.97	6.62
	10	89.98	10.55
\mathcal{M}_{Exp}	0	87.27	12.87
	1	88.83	13.01
	10	81.78	12.90

Table 1. Prognostics Performance for Impeller Wear

The assumption here is that the process noise represents finer-grained unmodeled processes that are not incorporated into the model and therefore look like noise.

Prognostics performance is dependent on both the underlying models used and on the prognostics algorithm. In order to focus on the dependence on modeling, we fix the algorithm and its parameters. The particle filter used $N = 500$ in all cases. The variance control algorithm used $v_0^* = 50\%$, $T = 60\%$, $v_\infty^* = 10\%$ in all cases, and used $P = 1 \times 10^{-3}$ for the damage models with one unknown wear parameter and $P = 1 \times 10^{-4}$ for those with two unknown wear parameters.

The prognostics performance results for impeller wear using different damage models and different levels of process noise variance are shown in Table 1. The process noise variance multiplier is shown in the second column of the table. We average RA over all prediction points to summarize the accuracy, denoted using $\overline{\text{RA}}$, and we average RMAD over all prediction points to summarize the spread, denoted using $\overline{\text{RMAD}}_{RUL}$. Multiple experiments were run for each case, and the table presents the averaged results. We can see that the linear damage model actually does fairly well. Its performance decreases as process noise increases, but for small amounts of process noise the accuracy is over 90%. The polynomial model also does well, which is expected since the second term by itself is the reference damage model. The particle filter still estimates a linear component which tracks damage progression over a short term fairly well, and it is the presence of this linear term that causes the accuracy to decrease. The exponential model does not do as well, partly because the behavior is very sensitive to the wear parameter inside the exponential function, w_{A2} , and so estimating both wear parameters simultaneously is more difficult for the particle filter.

The estimation performance using the reference model and the linear model is compared in Fig. 5. In both cases, the

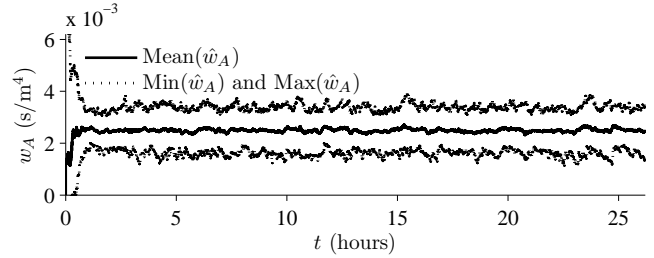
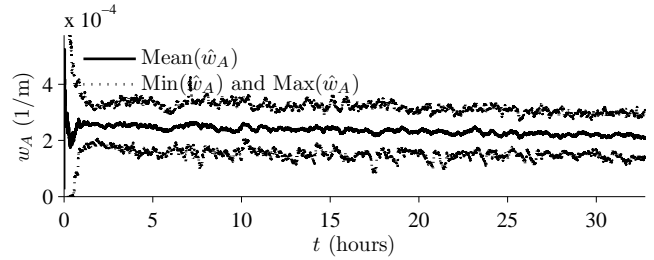

 (a) w_A estimation performance for \mathcal{M} .

 (b) w_A estimation performance for \mathcal{M}_{Linear} .

Figure 5. Impeller wear parameter estimation.

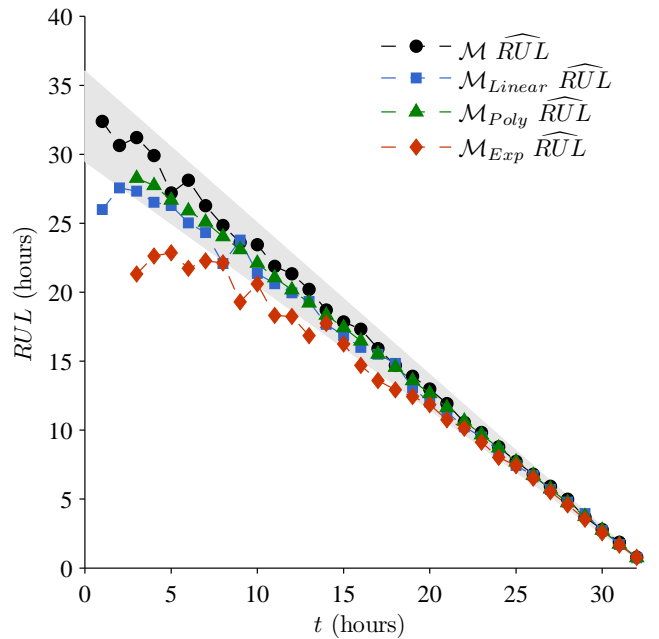


Figure 6. Impeller wear RUL prediction performance.

damage variable, A , was tracked well. When using the same damage model as in the reference model, the wear parameter is tracked easily and after convergence remains fairly constant. As a result, the predictions, shown in Fig. 6, using the mean, denoted by \widehat{RUL} , are very accurate and appear within 10% of the true value at all prediction points (shown using the gray cone in the figure). Because the rate of damage progres-

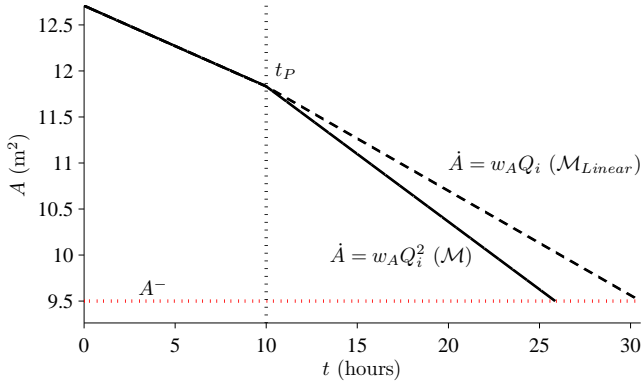


Figure 7. Impeller wear damage progression prediction, where at t_P , Q_i increases by 30%.

sion in the reference model decreases slowly over time, and the linear model does not accurately capture that behavior, its wear parameter estimate decreases over time in order to keep tracking the short-term damage progression. This is reflected also in the RUL predictions. Although the RUL accuracy is also very good, it is clear that it consistently underestimates the true RUL, because at any point in time it is overestimating the rate of damage progression that would occur in the future. However, the prognostic horizon is still very high. As shown in Fig. 6, by the second or third prediction, the predictions are all within the desired accuracy cone, except for the exponential model, which has PH of around 60%, meaning that at 60% life remaining, the exponential model is making accurate predictions. In many practical situations that may, in fact, be enough time for decision-making.

For impeller wear, the linear model does well in this case because the future loading is the same as the current loading. If Q_i is held constant, then the reference damage model $\dot{A} = w_A Q_i^2$, which equals $(w_A Q_i) Q_i$, looks exactly like the linear form because the product $w_A Q_i$ is constant. So the particle filter would estimate a wear parameter for the linear model that is the product of the wear parameter for the reference model multiplied by Q_i . So under constant loading, the linear model, or any other damage model that predicts a constant \dot{A} under uniform loading, will produce accurate predictions. But, if the future loading is different than the current loading, then the product $w_A Q_i$ will change and the wear parameter estimated for the linear model will no longer be valid. This is illustrated in Fig. 7. At t_P , Q_i increases by 30%. The algorithm using the reference damage model captures the relationship between \dot{A} and Q_i consistently with the simulation, and predicts EOL to be a little over 25 hours. In contrast, the linear model overestimates the RUL, because its wear parameter was tuned to the previous value of Q_i , and results in a RA of only around 80%. So for complex loading situations, it is important to correctly capture the relationship between loading and damage progression.

Model	\mathbf{v}	$\overline{\text{RA}}$	$\overline{\text{RMAD}}_{\text{RUL}}$
\mathcal{M}	0	97.80	11.61
	1	97.57	11.43
	10	97.50	11.18
$\mathcal{M}_{\text{Linear}}$	0	79.93	10.72
	1	83.93	10.79
	10	82.45	9.41
$\mathcal{M}_{\text{Squared}}$	0	78.05	11.59
	1	79.68	12.15
	10	74.59	11.17
$\mathcal{M}_{\text{Poly}}$	0	78.43	6.07
	1	78.94	9.09
	10	76.48	11.76
\mathcal{M}_{Exp}	0	82.34	9.23
	1	79.87	12.43
	10	69.37	21.32

Table 2. Prognostics Performance for Thrust Bearing Wear

The prognostics performance results for thrust bearing wear using different damage models and different levels of process noise variance are shown in Table 2. Results for radial bearing wear are similar, since the same damage models were used, and are omitted here. For the thrust bearing wear, only the case using the correct damage model obtains reasonable accuracy. The estimation results for some of the damage models are shown in Fig. 8. In all cases, the damage variable, r_t , was tracked well. With the algorithm using the reference damage model, the wear parameter is tracked well and after convergence remains approximately constant. In contrast, the linear model does not capture the relationship with ω correctly (i.e., in the reference model it is really a function of ω^2), so as ω changes between the two RPM levels, the estimate of the wear parameter must constantly increase and decrease to correctly track the damage progression. Further, because the rate of damage progression in the reference model increases over time (since it is a function of r_t), and the linear model does not capture that behavior, its wear parameter estimate must increase over time. With the polynomial model also, the parameter estimates do not take on constant values. This is also due partly to the fact that a wide number of pairs of w_{t1} and w_{t2} , i.e., multiple solutions to the damage progression equation, can track the short-term damage progression well. Hence, the wear parameter estimates can change over the long-term while still tracking short-term, leading also to an increased variability in the prediction accuracy.

The prediction performance is compared in Fig. 9. The algorithm using the reference model obtains accurate predictions. On the other hand, the other models consistently overestimate the RUL, because at any point in time they are underesti-

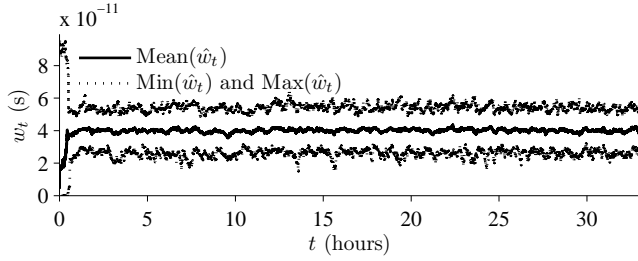
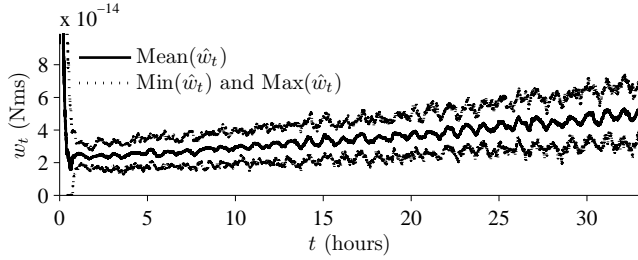
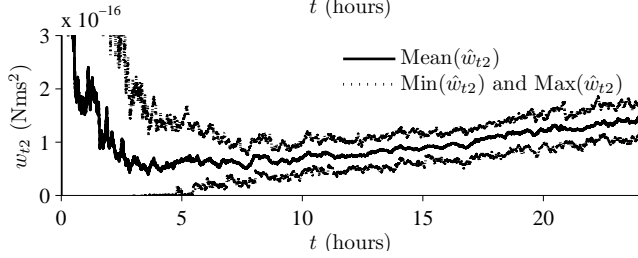
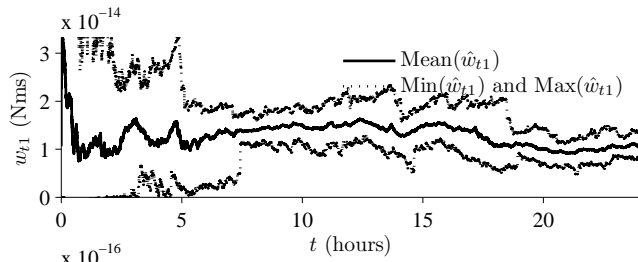

 (a) w_t estimation performance for \mathcal{M} .

 (b) w_t estimation performance for \mathcal{M}_{Linear} .

 (c) w_{t1} and w_{t2} estimation performance for \mathcal{M}_{Poly} .

Figure 8. Thrust bearing wear parameter estimation.

mating the rate of damage progression that would occur in the future. So, early on, the predictions are overly optimistic and could result in poor decisions based on that information. These models also produce very similar predictions. For the reference model \mathcal{M} , PH is around 95%, but for the remaining models, PH is around 30% or worse, so, for these models, accurate predictions are only being obtained with less than 30% life remaining, as observed in Fig. 9.

Note also that as the process noise increased, the algorithm using the reference model had only small decreases in performance, whereas for the other models, performance decreased quite significantly. In this case it was more difficult for the particle filters using these models to track damage over the

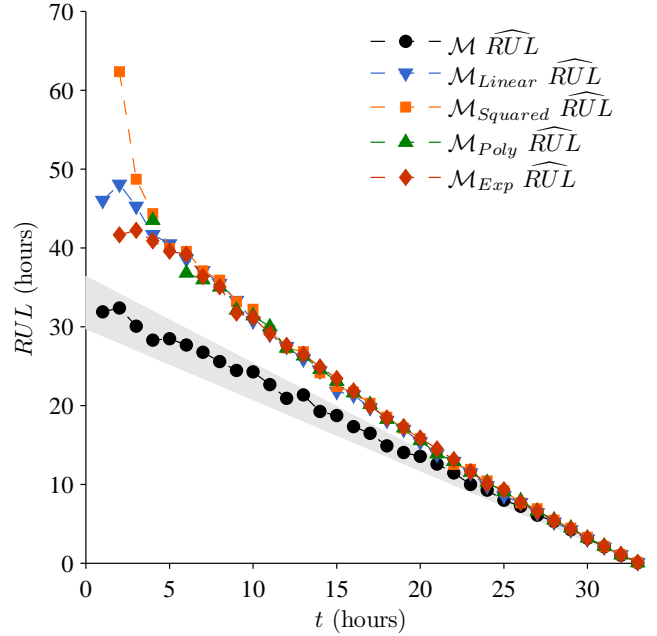


Figure 9. Thrust bearing RUL prediction performance.

short term, which resulted in a greater variation in the wear parameter estimates, leading to large decreases in accuracy.

Overall, this analysis illustrates the trade-off in the development of models of damage progression. In some cases, simple, more abstract or less granular models may suffice, especially if the system load remains constant. But with more complex operational scenarios, the need for a damage model that accurately captures the relationship with the load is necessary. In the case of the thrust bearing wear, even though the current and future inputs were the same, the fact that all of the less granular models did not account for the relationship between \dot{r}_t and r_t , which caused the damage progression rate to increase over time, resulted in poor prognostics performance, even for the more complex models. The more complex models, i.e., those with more unknown wear parameters, allowed more flexibility to correctly approximate the correct damage progression function, but this also increased the dimension of the joint state-parameter space and made estimation more difficult.

8. CONCLUSIONS

We presented a model-based prognostics methodology, and investigated the effect of the choice of damage progression models on prognostics performance. In prognostics modeling, accurate damage progression models are crucial to achieving useful predictions. Using a centrifugal pump as a simulation-based case study, we developed several different damage progression models, and, assuming some physics-based wear equations as the reference form, compared the performance of the prognostics algorithm using the different

models. In some cases, such as under cyclic or constant loading, it was shown that simple linear models may suffice. Some models also performed poorly early on but achieved accurate predictions before 50% life remaining. But, omitting additional interactions within the damage progression models may cause inaccurate results, even under simple loading scenarios. Further, even though the prognostics algorithm was robust enough to track the damage with all the different models, this did not translate to accurate predictions when a different damage progression model was used relative to the reference model.

In future work, we will extend this analysis to other domains such as electrochemical systems and electrical devices, in order to establish general design guidelines for prognostics models. For a desired level of prognostics performance, we want to be able to determine what level of model granularity is necessary. These ideas also apply to data-driven models, and models for diagnosis, which will be addressed in future work as well.

ACKNOWLEDGMENTS

The funding for this work was provided by the NASA System-wide Safety and Assurance Technologies Project (SSAT) project.

REFERENCES

- Arulampalam, M. S., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for on-line nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2), 174–188.
- Biswas, G., & Mahadevan, S. (2007, March). A Hierarchical Model-based approach to Systems Health Management. In *Proceedings of the 2007 IEEE Aerospace Conference*.
- Daigle, M., & Goebel, K. (2010, March). Model-based prognostics under limited sensing. In *Proceedings of the 2010 IEEE Aerospace Conference*.
- Daigle, M., & Goebel, K. (2011, March). Multiple damage progression paths in model-based prognostics. In *Proceedings of the 2011 IEEE Aerospace Conference*.
- Doucet, A., Godsill, S., & Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10, 197–208.
- Frantz, F. (1995). A taxonomy of model abstraction techniques. In *Proceedings of the 27th conference on Winter Simulation* (pp. 1413–1420).
- Hutchings, I. M. (1992). *Tribology: friction and wear of engineering materials*. CRC Press.
- Kallesøe, C. (2005). *Fault detection and isolation in centrifugal pumps*. Unpublished doctoral dissertation, Aalborg University.
- Lee, K., & Fishwick, P. A. (1996). Dynamic model abstraction. In *Proceedings of the 28th conference on Winter Simulation* (pp. 764–771).
- Luo, J., Pattipati, K. R., Qiao, L., & Chigusa, S. (2008, September). Model-based prognostic techniques applied to a suspension system. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 38(5), 1156–1168.
- Lyshevski, S. E. (1999). *Electromechanical Systems, Electric Machines, and Applied Mechatronics*. CRC.
- Saha, B., & Goebel, K. (2009, September). Modeling Li-ion battery capacity depletion in a particle filtering framework. In *Proceedings of the Annual Conference of the Prognostics and Health Management Society 2009*.
- Saha, B., Quach, P., & Goebel, K. (2011, September). Exploring the model design space for battery health management. In *Proceedings of the Annual Conference of the Prognostics and Health Management Society 2011*.
- Saxena, A., Celaya, J., Saha, B., Saha, S., & Goebel, K. (2010). Metrics for offline evaluation of prognostic performance. *International Journal of Prognostics and Health Management*.
- Tu, F., Ghoshal, S., Luo, J., Biswas, G., Mahadevan, S., Jaw, L., et al. (2007, March). PHM integration with maintenance and inventory management systems. In *Proceedings of the 2007 IEEE Aerospace Conference*.
- Wolfram, A., Fussel, D., Brune, T., & Isermann, R. (2001). Component-based multi-model approach for fault detection and diagnosis of a centrifugal pump. In *Proceedings of the 2001 American Control Conference* (Vol. 6, pp. 4443–4448).
- Zeigler, B., Praehofer, H., & Kim, T. (2000). *Theory of modeling and simulation* (2nd ed.). Academic Press.

Investigation on the opportunity to introduce prognostic techniques in railways axles maintenance

Mattia Vismara¹

¹*Hupac SA, Chiasso, Switzerland*
mvismara@hupac.ch

ABSTRACT

In this study the opportunity to introduce PHM (prognostic and health monitoring) concepts into a cracked railway axle management is investigated.

The performances of two different prognostic algorithm are assessed on the basis of their RUL (remaining useful life) predictions accuracy: a prognostic model based on the Bayesian theory and a physical prognostic model. Both models rely on the measured crack size. The measured crack growth measure has been built from simulated probabilistic crack growth path by adding measurements errors. The effect of monitoring frequency and the measurement HW and SW infrastructure size error to RUL predictions' accuracy is assessed as well, trying to evaluate the hypothetical measuring infrastructure capabilities' (sensors + layout) effect on the overall PHM predictions.

Furthermore the PHM approach is compared to the classical preventive maintenance approach to railway axle maintenance management based on expensive and regular NDT.

1. INTRODUCTION

Railway axles are designed to have an infinite lifetime (EN13103, 2001). However occasional failures have been and are observed in service. The typical failure positions are the press-fits for wheels, gears, and brakes or the axle body close to notches and transitions. Such failures always occur as fatigue crack propagations whose nucleation can be due to different causes (U. Zerbst M. V., 2005). In the case of railway axles, the presence of widespread corrosion (Hoddinot, 2004) (C.P. Lonsdale, 2004) or the possible damage due to the ballast impacts (M. Carboni, 2007) may constitute such causes.

This kind of failures is usually tackled by employing the 'damage tolerance' methodology, whose philosophy consists (U. Zerbst M. V., 2005) (U. Zerbst K. M., 2005) in determining the most opportune inspection interval given the 'probability of detection' (PoD) of the adopted non-

destructive testing (NDT) technique or, alternatively, in defining the needed NDT specifications given a programmed inspection interval.

The negligible number of axle failures is reached thanks to role played by inspections carried out with the aim of keeping developing fatigue problems at bay. As reported by (R.A. Smith, 2004) in the United Kingdom there have been about 1.6 axle failures per year over the last 25 years, out of a population of about 180,000 axles. (A similar number of new axles are introduced every year in PR China, where some 2.5×10^6 wheelsets are in fleet service.) These large numbers of axles are subjected to inspections in order to try to identify cracks before failures occur. In general, the examinations are expensive, time consuming and not particularly effective in finding cracks. Furthermore, the dismantling needed to examine axles, such as the drawing-off of bearings, can cause scratching damage that is sufficiently severe to cause an axle to be retired. The rationale behind the frequency of testing is that the largest crack that would not be detected in an inspection should not grow to failure during the service interval to the next inspection. This implies that crack propagation calculations should be performed with sufficient accuracy to set the inspection interval. However, as stated by (R.A. Smith, 2004) some difficulties arises:

- Due to the difficulty in determining the reliability and sensitivity of the inspection techniques, the initial crack length chosen for the life calculation must be set larger, leading to shorter intervals between inspections than are really necessary.
- The service loads are much more stochastic in nature than the well-defined hypothetical loads used for the initial design rule suggest. In many cases, in the absence of experimental measurement, the magnitudes and frequencies of these events are unknown, thus making cycle-by-cycle crack growth predictions unreliable.
- Important inputs to fatigue calculations are material properties such as crack growth data, fatigue limits and fatigue thresholds, which are very sensitive to material condition, manufacturing route, surface treatment, orientation and load sequence. In many cases these data

M.Vismara. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

are lacking, particularly from large specimens representative of axles.

- Abnormal conditions may arise in service. There is debate about the best means of protecting axles from corrosion and the extent to which coatings may hinder inspection. The interactions between fatigue and corrosion mechanisms in extending defects are still inadequately understood. Higher speeds have led to increased examples of damage of axles from flying ballast, which may be of the form of crack-like indentations on axle surfaces that initiate premature failure.

These considerations can lead to think that maybe, instead of using a preventive maintenance approach a predictive maintenance approach based on prognostics could be convenient. Several aspects has to be considered in order to assess the technical and economical feasibility of this approach. The first and the most important is the assessment of the prognostic algorithm predictions accuracy and its sensibility to the goodness of the diagnostic and monitoring equipment used.

This section constitute the first attempt to answer to this question through an explanatory assessment of two prognostic algorithms. The first one is based on statistical method, the second one exploit the good understanding of the crack propagation physical process to estimate the time to fail of a cracked axle. Moreover, the predictive maintenance approach is qualitatively compared to the classical preventive approach.

2. PROBLEM FORMULATION

2.1 Simulation of the crack growth paths – The stochastic crack growth algorithm

In this paragraph the stochastic crack growth model used in this work is presented. The non-powered railway axle considered in the present study is manufactured in A1N steel and used in Y25 bogie with a diameter D equal to 160 mm.

Service loads acting on railway axles are the result of vertical and lateral forces (EN13103, 2001) due to their normal functioning, and the maximum bending moments can be found in the area of the wheels press-fit (U. Zerbst M. V., 2005)(M. Carboni, 2007). On the basis of these considerations, fatigue crack growth has here been analyzed at the typical T-transition between the axle body and the press-fits.

Different algorithms for simulating the crack growth of cracked components are available in literature. Some of them consider the crack growth modeling as stochastic process, see for example (K.Ortiza, 1988),(D.A. Virkler, 1979).(J.L Bogdanoff, 1985). However, the likelihood of lifetime calculations depends on the adopted FCG algorithm

and only the most complex algorithms are able to adequately describe crack propagation under variable amplitude loading in railway axles (S. Beretta M. C., 2006).

In the present work the NASGRO algorithm (Anonymus, 2006) will be considered. This FCG model has been chosen because it is the reference algorithms in analyses where random loadings are involved, since it takes into account the “plasticity-induced crack closure” phenomenon (EN13103, 2001). Moreover, NASGRO has been used in several papers addressing the propagation of fractures in railway axles (U. Zerbst M. V., 2005) (S.Beretta M. , Simulation of fatigue crack propagation in railway axles, 2005)(S. Beretta M. C., 2004).

The considered software adopts the Paris-based crack propagation law called “NASGRO equation”:

$$\frac{da}{dN} = C \left[\left(\frac{1-f}{1-R} \right) \Delta K \right]^n \frac{\left(1 - \frac{\Delta K_{th}}{\Delta K} \right)^p}{\left(1 - \frac{\Delta K}{(1-R)\Delta K_{crit}} \right)^q} \quad 2.1$$

where “C”, “n”, “p” and “q” are empirical constants, “R” is the stress ratio, “ ΔK_{th} ” is the threshold SIF range and “ ΔK_{crit} ” the critical SIF.

To analyze cracked bodies under combined loading, the stress intensity factor is expressed as:

$$\Delta K_{nom} = \left[\sum_{i=1}^6 \alpha_i \left(\frac{a}{D} \right)^i + \beta \right] (1-R)(S + \varepsilon) \sqrt{\pi a} \quad 2.2$$

Where α_i and β are empirical constants, S is the applied bending stress, a is the crack size and ε is a random coefficient (introduced later in the paragraph). The bending stress is considered plane since NASGRO is not able to consider rotating bending conditions. This assumption has not a significant influence on estimated life predictions as demonstrated in (S.Beretta M. , Rotating vs. plane bending for crack growth in railway axles, 2005)(S.Beretta M. M., 2006).

The closure function is defined as:

$$f = A_0 + A_1 R \quad 2.3$$

Where

$$A_0 = 0.825 - 0.34\vartheta + 0.05\vartheta^2 \left[\cos \left(\frac{\pi}{2} S_0 \right) \right]^{\frac{1}{\vartheta}} \quad 2.4$$

$$A_1 = (0.415 - 0.071\vartheta) S_0$$

ϑ is a plane stress/strain constraint and S_0 is the ratio of the maximum applied stress to the flow stress.

Since NASGRO does not consider the geometry of the typical transitions of axles, equation 2.5 is modified in terms of the maximum SIF present at the notch root and calculated as

$$\Delta K = K_t \Delta K_{nom} \quad 2.6$$

K_t represents the experimental stress concentration (S. Beretta M. C., 2004).

As demonstrated by (S. Beretta M. C., 2006), the crack growth randomness can be described considering the stress intensity factor threshold as a random variable. Particularly, it is demonstrated that ΔK_{th} can be considered as belonging indifferently to a lognormal distribution or normal distribution. In this work is considered as a normal variable with mean ΔK_{th} and standard deviation $\sigma_{\Delta K_{th}}$. The empirical calibration of all the other parameters is carried out by means of dedicated fracture mechanic experiments. Their values are listed in Appendix. Another relevant source of uncertainty is the randomness of the applied load (U. Zerbst M. V., 2005)(M. Carboni, 2007). Therefore service loads have been considered derived from experimental results on a high speed train. Next, the service stress spectrum has been approximated with a simple block loading consisting of twelve blocks (Table 1). To take into account the within block variability a random term ε is added in the Eq.2.9. It is assumed to be uniformly distributed with mean equal to 0 and with a span of 2ε .

The so defined block loadings were then applied to growth calculations with a time sequence in accordance to Gassner suggestions (Gassner, 1956). Starting from the discrete spectrum in Table 1, the random history loads sequence is built by permutations of the whole set of the blocks. Each load sequence is 3.222.887 km long, composed of 20 consecutive complete permutations. Some simulated crack growth path, considering all the uncertainties described (load history, ΔK_{th} and ε) are shown in Figure 1.

Cycles	Load [MPa]
1	145
8	135
75	125
825	115
15,000	105
110,025	95
357,675	85
678,900	75
1,621,725	65
3,046,500	55
8,165,775	45

39,718,275	35
------------	----

Table 1 The 12 service time blocks

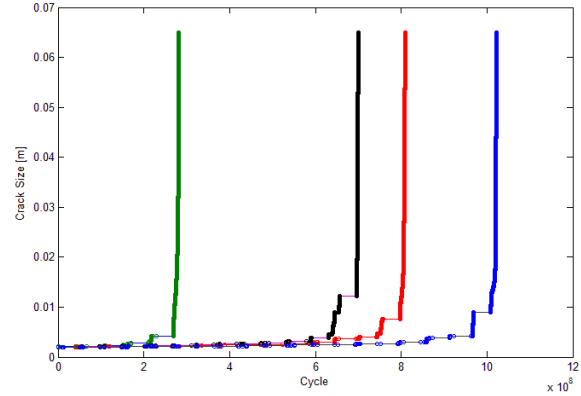


Figure 1 Examples of simulated crack growth paths

Eventually, once determined an initial crack size and a limiting crack depth value at failure, through the Monte Carlo technique is possible to estimate the TTF pdf. Each simulation run is characterized by a random ΔK_{th} and a random load history. Considering an initial crack size of 2 mm and a limiting crack size of 60 mm, the TTF pdf is shown in Figure 2.

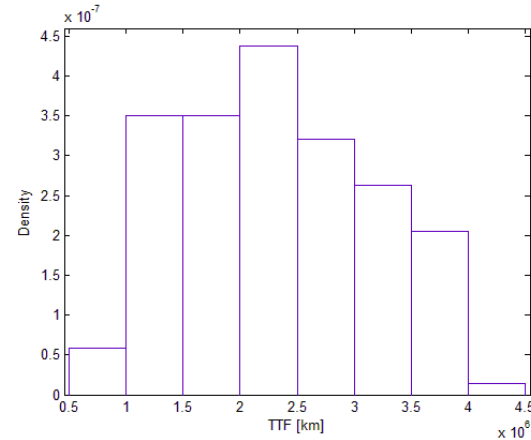


Figure 2 TTF probability distribution

The TTF pdf for the purposes of this work is considered as a lognormal distribution as can be observed in Figure 3. It can be noticed how a lognormal distribution fits well the TTF data for almost the whole TTF variability range, only the right hand tail significantly diverge for the TTF. This is demonstrated also by Beretta et al. (S. Beretta M. C., 2006) and Schijve (Schijve, 2001).

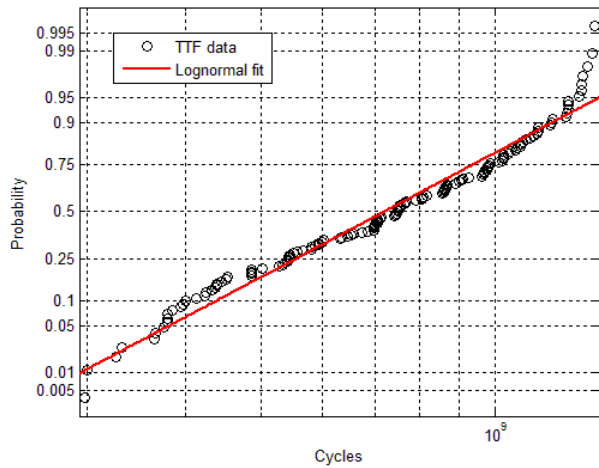


Figure 3 Lognormal fit plot for TTF pdf

2.2 Design of the preventive maintenance approach

The preventive maintenance approach is designed according to the damage tolerant approach well described by (U. Zerst M. V., 2005) (U. Zerst K. M., 2005). The steps that have to be followed to design a design an axle preventive maintenance plan are:

1. establishment of the initial crack shape and size for further analysis
2. within a damage tolerance concept the initial crack size, a_0 , is not identical to the size of a real flaw, e.g., from the manufacturing process but is a fictitious size, which usually refers to the detection limit of the NDI technique. The basic idea is that the largest crack that could escape detection is presupposed as existent.
3. simulation of sub-critical crack extension, This kind of crack growth is designated as sub-critical since it will not lead to immediate failure until a critical length of the crack is reached. For railway applications the common mechanism is fatigue.
4. determination of critical crack size for component failure. The sub-critical crack extension is terminated by the failure of the component. This may occur as brittle fracture or as unstable ductile fracture. Critical states may, however, also be defined by other events such as stable ductile crack initiation or the breakthrough of a surface crack through the wall or setting a maximum allowable crack size threshold.
5. determination of residual lifetime of the component, The residual lifetime is that time or number of loading cycles which a crack needs for extending from the initial crack size, a_0 , (step 1) up to the allowable crack size, a_{max} , established in step (3).
6. specification of requirements for non-destructive testing.

The constitution of an inspection plan is the aim of a damage tolerance analysis. From the requirement that a potential defect must be detected before it reaches its critical size it follows immediately that the time interval between two inspections has to be smaller than the residual lifetime. Sometimes inspection intervals are chosen to be smaller than half this time span. The idea is to have a second chance for detecting the crack prior to failure if it is missed in the first inspection. It is, however, also obvious that frequently even two or more inspections cannot guarantee the crack being detected since this would require a 100% probability of detection.

The procedure described by (U. Zerst M. V., 2005) aims to define the NDT specifications following the ‘last chance’ approach introduced in (M. Carboni, 2007). In this case, the PoD is not a variable to be optimized but is given. Therefore the maximum inspection interval was defined instead of the requirements for non destructive testing. The steps from 1 to 4 has already been done in the previous paragraph.

2.2.1 The PoD curve

The PoD can be derived from the calibration function of the particular NDE equipment used that relates the crack dimension (length, depth or area) to the output. In this case, the NDE method considered is the ultrasonic inspection. Since output from an NDE measurement process is a continuous response, the calibration curve is modeled as a linear function in which the measurement (dB of the signal) is given by a linear combination of two parameters and the crack area (\hat{a} [mm^2]) plus a normal zero mean error with constant variance (Eq.2.7).

$$Y(\hat{a}) = \beta_0 + \beta_1 \log_{10} \hat{a} + \epsilon(0, \sigma_r) \quad 2.7$$

The parameters $\beta_0, \beta_1, \sigma_r$ are estimated through the LSE or through the MLE methods. It is assumed that 1000 dB and -1000dB are respectively the saturation and observable limits.

The data provided from which the parameters are estimated have been obtained from real inspections of railway axles. The parameters’ values are reported in Table 2.

Parameter	Value*
β_0	Xxo
β_1	Yyo
σ_r	Zzo

Table 2: Calibration Curve Parameters

In order to use the calibration curve in the following analysis, the crack size has to be expressed in term on depth instead of surface area. The crack geometry is assumed to

* Values are omitted for confidentiality reasons

be semicircular (M. Carboni, 2007). Therefore, the resulting calibration curve function becomes:

$$Y = \beta_0 + \beta_1 \log_{10} \left(\frac{\pi a^2}{2} \right) + \epsilon (0, \sigma_r) \quad 2.8$$

In order to derive the PoD function, a threshold is fixed that represents the measure's bound that if it's overcome, the presence of a crack is diagnosed. This limit is set at 50.6 dB that corresponds to a crack depth of 5.492 mm.

The reference limit and the final calibration curve with the constant $3\sigma_r$ confidence limits is shown in Figure 4.

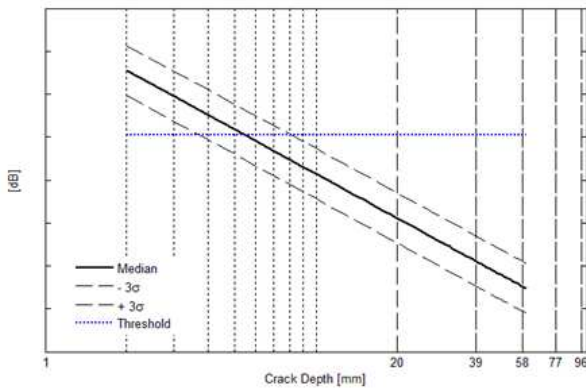


Figure 4: Final Calibration Curve

At this point the PoD curve can be derived as it represents the probability that a crack of size a can be detected, given that the threshold is set at a_{th} . According to this statement and making the hypothesis of a normal distributed error, the PoD of a crack depth a is:

$$\begin{aligned} PoD(a) &= P[Y(a) > Y(a_{th})] = \\ &= 1 - \Phi \left(\frac{Y(a_{th}) - \left(\beta_0 + \beta_1 \log_{10} \left(\frac{\pi a^2}{2} \right) \right)}{\sigma_r} \right) \quad 2.9 \\ &= 1 - \Phi \left(\frac{50.6 - \left(\beta_0 + \beta_1 \log_{10} \left(\frac{\pi a^2}{2} \right) \right)}{\sigma_r} \right) \end{aligned}$$

where Φ is the standard normal cdf. In Figure 5 is shown the resulting PoD curve.

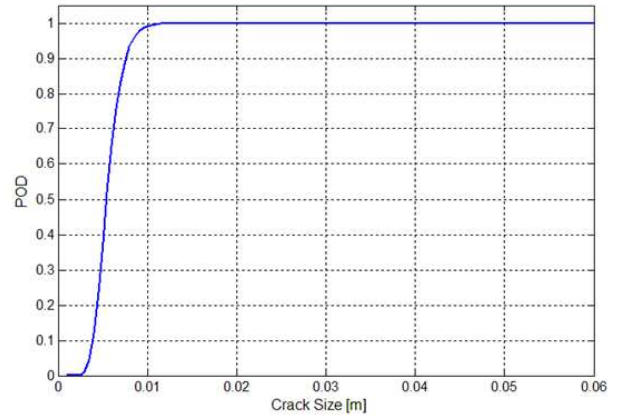


Figure 5: PoD

The PoD as discussed above in paragraph 2.2 is used to determine the maximum inspection interval in order to detect with a probability R the maximum allowed crack size a_{max} . In the following paragraph, according to the problem defined in paragraph 2.2, the maximum inspection interval is determined.

2.2.2 Identification of the maximum inspection interval

The maximum safe inspection interval is determined through examining the effect of the interval of inspection on the overall probability of detection in the case of a fast growing crack. The inspection interval is therefore the maximum interval of inspection that allows the detection of the maximum allowable crack size with a defined reliability. The worst case is when the time (or distance) before the failure occurs (TTF) is minimum. This happens when, once the maximum defect present in the system is set, the crack growth rate is the highest. The inspection interval is therefore dependent on the largest defect present in the system, that is the defect that will eventually cause failure.

The maximum defect size is set at 2 mm as suggested by the literature reviewed (M. Carboni, 2007) (U. Zerbst M. V., 2005) and as set in the crack growth simulations. At this point the fastest growth crack has to be chosen as the reference upon which the maximum allowable inspection interval has to be defined.

Starting from the TTF distribution shown in Figure 2, the fastest growth crack has been chosen. It is the crack growth path with the minimum TTF in 300 simulations and that falls in the first bin of the TTF distribution. In Figure 6 is shown the path selected and its relative position with respect to the TTF distribution (blue line). As can be seen it falls in the left tail of TTF pdf.

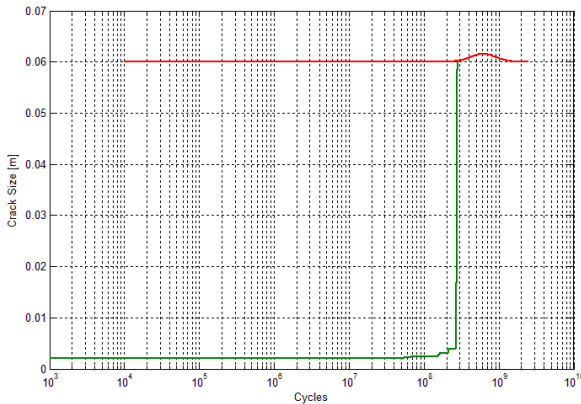


Figure 6 Fastest growth crack

Once the worst case is chosen and the reference PoD has been defined, the maximum inspection interval can be found.

Given an inspection interval, 'I', the cumulative PoD PC_{DET} of a defect, potentially observable in a given number of inspections, i , is calculated from the PoD curve of the adopted NDT technique. Figure 7 shows how the cumulative probability of detection is calculated, that in formulae results.

$$PC_{DET} = 1 - \prod_{i=1}^n PonD_i \tag{2.10}$$

$$PonD_i = 1 - Pod_i$$

Here, PC_{DET} is the theoretical cumulative PoD and $PonD$ ('probability of non-detection') represents the probability of failing to detect in a given inspection.

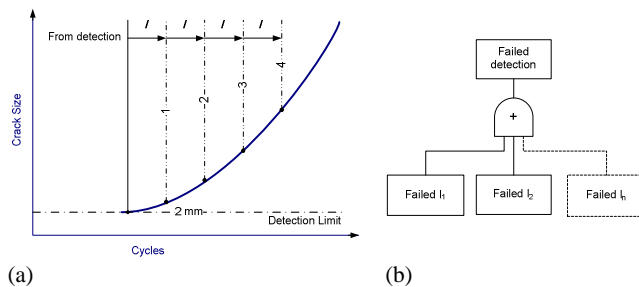


Figure 7 Calculation of the cumulative probability of detection (a) and the fault tree of the inspection (b) (adopted from (M. Carboni, 2007))

The PoD_i depends on the actual crack size a that corresponds to the cycle i according to the Eq. 2.9. The more the inspections the more the PC_{DET} will be.

Since a 100% PC_{DET} is impossible to reach theoretically, a PC_{DET} threshold was set at 0.99.

In order to determine the inspection interval the final PC_{DET} is evaluated at different intervals of inspection. Particularly, the final PC_{DET} was evaluated starting from 1 to 60 inspections that result in the same number of intervals.

The final PC_{DET} is the P_{DET} that results from the last inspection. Figure 8 shows the results of the assessment, it shows the PC_{DET} as a function of the inspection interval. The figure confirm what stated previously: as the number of inspection increases and the inspection interval decreases PC_{DET} increases. The optimal inspection interval is the largest that guarantee a $PC_{DET} = 0.99$.

From Table 3, can be seen that the inspection interval at 0.99 falls between 34,988 km and 32,297 km. By linear interpolation we can find that the interval at 99% PC_{DET} is **33,663 km**.

N° inspections	Inspection Interval [km]	PC_{DET}
1	419,856	0.000000
3	209,928	0.000014
5	139,952	0.003680
7	104,964	0.003694
9	83,971	0.007346
11	69,976	0.007360
13	59,979	0.010992
15	52,482	0.011006
17	46,651	0.026369
19	41,986	0.026967
21	38,169	0.397817
23	34,988	0.834808
25	32,297	0.999981

Table 3 PC_{DET} with different inspection interval

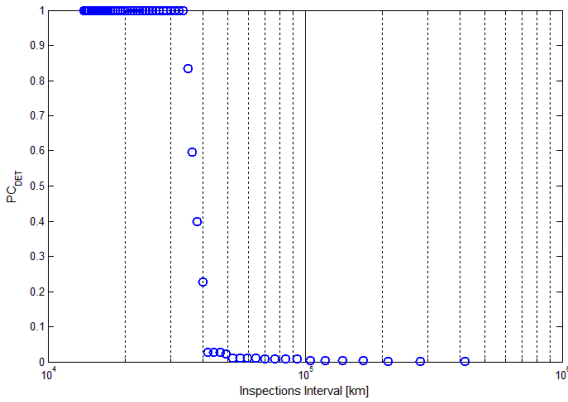


Figure 8 PC_{DET} as function of the inspection intervals

The literature reviewed (U. Zerbst M. V., 2005)(M. Carboni, 2007) (S. Beretta M. C., 2006) suggests to determine the inspection intervals referring to the average crack growth path, i.e whose TTF is equal to the mean TTF. In this case, once selected the right crack propagation lifetime, the maximum inspection interval is computed as well. The result is that the optimal inspection interval should be performed each 153,197.8 km. It is worth noting that in case of the fast crack growth crack, with an inspection interval equal to **153,197.8** km the PC_{DET} is equal to 0.2986%.

2.3 Prognostic Modeling of the Crack Size Growth

In this section two methods able to predict the RUL of cracked railway axles are introduced and compared in terms of their prediction performances.

The first model uses a statistical approach based on the Bayesian probabilistic theory and the second one uses the physical model introduced in the paragraph 2.1, the same used to generate the crack growth paths. Since the model accurately describe the real crack growth in railway axles(S. Beretta M. C., 2006), it can be used both to substitute experimental tests and to generate the database needed to support a statistical approach to evaluate the axles' TTF and RUL.

The aim of the section is to introduce and give evidence of the capability of a prognostic approach based on these algorithms to reduce the uncertainties associated to the prediction of the TTF of a continuously monitored cracked axle meanwhile it operates. This approach can be helpful to increase the inspection interval and, as a best result, inspects the axle only when the wheels have to be maintained without reducing the system's safety.

2.3.1 Setting the threshold

In order to design a prognostic algorithm capable of updating the axle's TTF the concept of failure has to be

clearly determined. In this case it is trivially derived since the axle is considered faulty when the maximum allowable crack size is reached. Obviously, the threshold has to be fixed considering the errors that affects the whole monitoring and prognostic system. Figure 10 shows a scheme of the different types of errors that has to be considered in setting the threshold. A safety margin has to be introduced against the errors that affect the estimation. The first error was introduced in the paragraph 2.2.1.

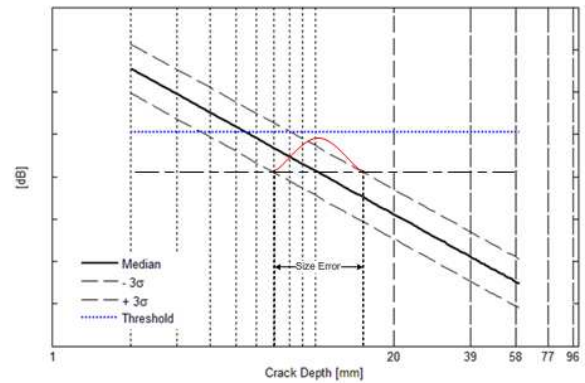


Figure 9 Illustration of the meaning of the size error

It is the error associated with the calibration curve of the ultrasonic inspection. This error introduces an uncertainty in the determination of the crack size given that the ultrasonic probe measures x dB.

Figure 9 illustrates what is meant for the size error. Given the calibration curve in Eq.2.8, the size error ϵ_s is defined as:

$$\epsilon_s = \frac{\epsilon}{\beta_1} \tag{2.11}$$

$$\epsilon_s = N\left(0, \frac{\sigma_r}{\beta_1}\right)$$

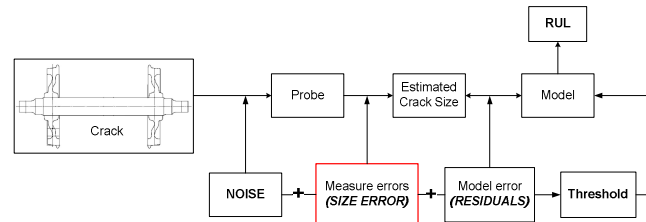


Figure 10 The errors affecting the monitoring and prognostic system

The other errors that are present are those associated with the model describing the crack growth, that are the residuals between the actual crack size and the that one predicted by the model and eventually the noise that affects the measurements process.

In this case the size error is only considered since no data are available about the other error sources. The error considered can be considered as the sum of those making the hypothesis that the used diagnostic system's performances are better.

Given a crack depth a_{max} as the maximum crack size allowed, the threshold that will be used as a reference for estimating the axle TTF is that one that guarantees at 99% of confidence that a_{max} won't be missed.

Starting from the calibration function in Eq.2.8 we have to find \tilde{a}_{th} that corresponds to $P(a_{max} \leq \tilde{a}_{th}) = 0.99$.

Starting from Eq.2.9, given the measure Y , the related crack size is:

$$a = \sqrt{\frac{2}{\pi} 10^{\frac{Y-\beta_0}{\beta_1}} 10^{\frac{\epsilon}{2\beta_1}}} \quad 2.12$$

Remembering that $\epsilon_s = \frac{\epsilon}{\beta_1}$, we have:

$$a = \sqrt{\frac{2}{\pi} 10^{\frac{Y-\beta_0}{\beta_1}} 10^{\epsilon_s}} \quad 2.13$$

Given that \hat{Y} corresponds to the measurement of the crack size a_{max} , we have:

$$a_{max} = \sqrt{\frac{2}{\pi} 10^{\frac{\hat{Y}-\beta_0}{\beta_1}}} \quad 2.14$$

The crack size that corresponds to the measurement \hat{Y} is:

$$c \quad a = \sqrt{\frac{2}{\pi} 10^{\frac{\hat{Y}-\beta_0}{\beta_1}} 10^{\frac{\epsilon_s}{2}}} \quad 2.15$$

$$a = a_{max} 10^{\frac{\epsilon_s}{2}}$$

From Eq.2.15 we have that given a real crack depth of a_{max} the crack size associated a (estimated from the measurement) is a random variable distributed as a lognormal with an associated mean of $\log_{10}(a_{max})$ and a standard deviation of $\frac{\sigma_r}{2\beta_1}$.

$$\log_{10} a = \log_{10} \left(a_{max} 10^{\frac{\epsilon_s}{2}} \right)$$

$$\log_{10} a = \log_{10}(a_{max}) + \log_{10} \frac{\epsilon_s}{2} \quad 2.16$$

$$\log_{10} \frac{\epsilon_s}{2} = N \left(0, \frac{\sigma_r}{2\beta_1} \right)$$

Now we can define the threshold \tilde{a}_{th} :

$$P(\tilde{a}_{th} - a_{max} \leq 0) \geq 0.99$$

$$\Phi \left(\frac{\log_{10} \tilde{a}_{th} - \log_{10} a_{max}}{\frac{\sigma_r}{2\beta_1}} \right) \geq 1 - 0.99 \quad 2.17$$

The result is $\tilde{a}_{th} = \mathbf{0.044}$.

If we let vary both σ_r and a_{max} and calculate the corresponding \tilde{a}_{th} we obtain a surface plotted in Figure 11. As we can see the relation is not linear and as the standard error increases, given a maximum crack size, the corresponding crack depth threshold decreases.

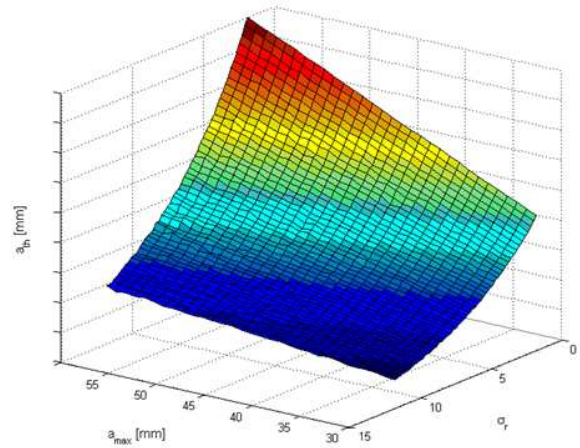


Figure 11 Crack size threshold as a function of σ_r and a_{max}

2.3.2 Bayesian updating algorithm

This section develops methods that combine two sources of information, the reliability characteristics of a axle's population and real-time sensor information from a functioning axle, to periodically update the distribution of the axles's residual life.

We first model the degradation signal for a population of axles with an appropriate model assuming error terms from an iid random error process. A Bayesian updating method is used to estimate the unknown stochastic parameters of the model for an individual component. Once we have determined the posterior distribution for these unknown parameters, we derive the residual-life distribution for the individual component.

In this case there is not simple functional form that fit well the simulated crack growth pattern. Nevertheless, an approximation of the paths can be performed by splitting the signal in two parts, that can be modeled as two exponential functions as shown in Figure 12.

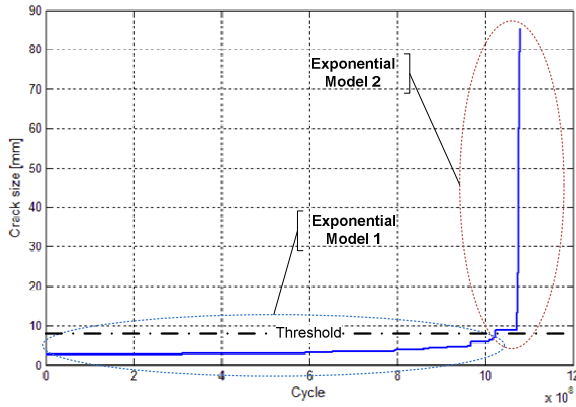


Figure 12 The two exponential models

The shift from the first model to the second is based on a crack depth threshold that is plotted in Figure 12 as a black dash dotted line. The TTF of the axle monitored is therefore defined as:

$$TTF = T_1 + T_2 \quad 2.18$$

Where T_1 is a random variable that express the predicted time to reach the threshold \tilde{S}_{th} and T_2 is a random variable as well that denote the time that takes the crack to grow from the threshold to \tilde{a}_{th} .

Let $S(t)$ denote the degradation signal as a continuous stochastic process, continuous with respect to cycle n . We observe the degradation signal at some discrete points in cycles, n_1, n_2, \dots , where $n_i \geq 0$. Therefore, we can model the degradation signal at cycles $n_i = n_1, n_2, \dots$, as follows:

$$\begin{cases} S(n_i) = \varphi_1 + \theta_1 \exp[\beta_1 n_i + \epsilon_1] & S \leq \tilde{S}_{th} \\ S(n_i) = \varphi_2 + \theta_2 \exp[\beta_2 n_i + \epsilon_2] & \tilde{S}_{th} \leq S \leq \tilde{a}_{th} \end{cases} \quad 2.19$$

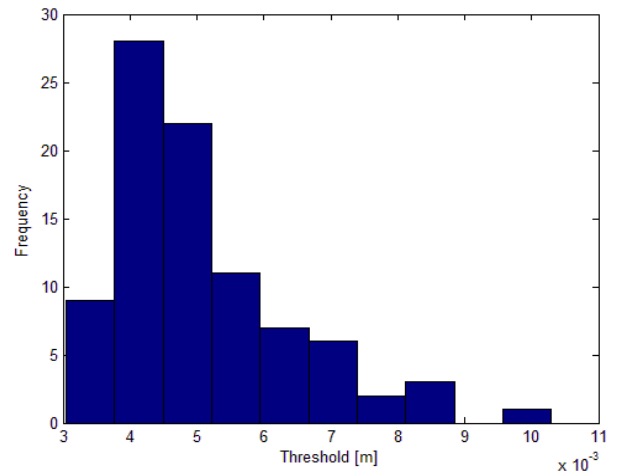
If we redefine $L_1(n_i) = S(n_i) - \varphi_1$ for $S \leq \tilde{S}_{th}$ and $L_2(n_i) = S(n_i) - \varphi_2$ for $\tilde{S}_{th} \leq S \leq \tilde{a}_{th}$ we obtain:

$$\begin{cases} L_1(n_i) = \theta_1 \exp[\beta_1 n_i + \epsilon_1(n_i)] & S \leq \tilde{S}_{th} \\ L_2(n_i) = \theta_2 \exp[\beta_2 n_i + \epsilon_2(n_i)] & \tilde{S}_{th} \leq S \leq \tilde{a}_{th} \end{cases} \quad 2.20$$

The choice of threshold \tilde{S}_{th} has to be based on an optimization rule. In this case, the threshold is that one that bound the maximum residual of the first fitted model to 0.0012. Obviously the rule can be changed, for example the threshold could be that one that minimize the overall fitting error. The value 0.0012 at which the first residual error is

bounded is chosen upon that willingness to prefer a better fit in the first part of the signal in order to achieve better predictions in the first stage of the degradation process. The reason is that good predictions (more precise) in the first part of the degradation path can restrict the uncertainties on the final RUL estimation from the beginning. As matter of facts, the main part of the uncertainty on the TTF comes from the uncertainty associated with the variable T_1 . In other words, the variance of the cycles taken by the crack to grow from the initial size to \tilde{S}_{th} is much greater that the number of cycles taken by the crack to grow from \tilde{S}_{th} to \tilde{a}_{th} .

After several simulations, the threshold that bound the maximum residual error of the first part of S is a random variable as shown in Figure 13.


 Figure 13 Threshold \tilde{S}_{th} distribution

Eventually the final threshold chosen is the mean value of distribution, that is $\tilde{S}_{th} = 5,1 \text{ mm}$.

Once determined the threshold, through an appropriate number of crack growth simulations, we can build our a priori information on the crack growth behavior. Our a priori information, a part from the a priori TTF distribution shown in Figure 2, is composed of the random parameters $\theta_1, \theta_2, \beta_1$ and β_2 probability distributions. Their values are obtained through the LSE technique though fitting the crack growth functions with the models in Eq.2.19. The final distribution PDFs are plotted in Figure 14.

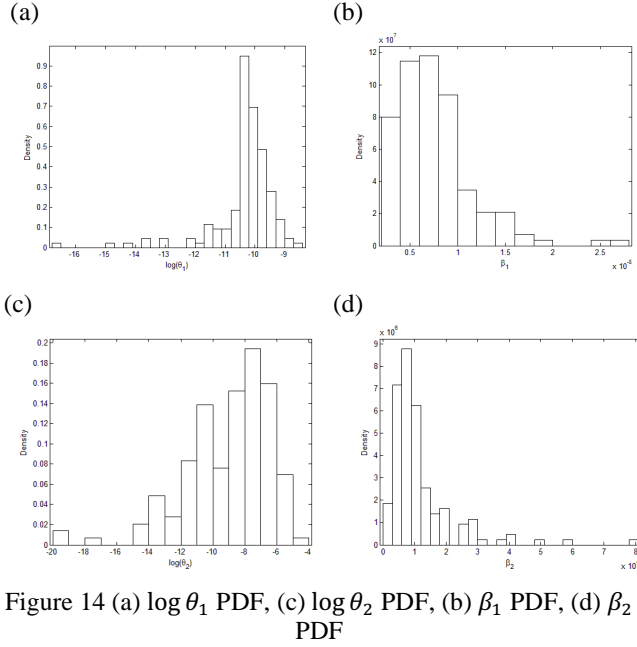


Figure 14 (a) $\log \theta_1$ PDF, (c) $\log \theta_2$ PDF, (b) β_1 PDF, (d) β_2 PDF

As can be noted from the figure above, $\theta_1, \theta_2, \beta_1$ and β_2 can be approximated by lognormal distributions[†] with parameters:

$$\begin{aligned} \theta_1 &= LN(\mu_{\theta_1}, \sigma_{\theta_1}) & \theta_2 &= LN(\mu_{\theta_2}, \sigma_{\theta_2}) \\ \beta_1 &= LN(\mu_{\beta_1}, \sigma_{\beta_1}) & \beta_2 &= LN(\mu_{\beta_2}, \sigma_{\beta_2}) \end{aligned}$$

The probability charts of those distributions can be found in the Appendix.

For these exponential models, it will be convenient to work with the logged signal S . We can then define the logged signal at cycle n_i as follows:

$$\begin{cases} LS_1(n_i) = \log \theta_1 + \beta_1 n_i + \epsilon_1(n_i) & c \leq \tilde{S}_{th} \\ LS_2(n_i) = \log \theta_2 + \beta_2 n_i + \epsilon_2(n_i) & \tilde{S}_{th} \leq S \leq \tilde{a}_{th} \end{cases} \quad 2.21$$

We will use the observations $LS_{i,1}, LS_{i,2}, \dots$, obtained at cycles n_1, n_2, \dots , as our data. Next, suppose we have observed $LS_{i,1}, \dots, LS_{i,k}$ at cycles n_1, \dots, n_k .

Since the error terms, $\epsilon_i(n_t)$, $i = 1, 2$ and $t = 1, \dots, k$, are assumed to be iid normal random variables, if we know $\theta_{1,2}$ and $\beta_{1,2}$, then the likelihood function of $LS_{i,1}, \dots, LS_{i,k}$, given $\theta_{1,2}$ and $\beta_{1,2}$, is:

$$f(LS_{1,1}, \dots, LS_{1,k} | \theta_1, \beta_1) = \left(\frac{1}{\sqrt{2\pi\sigma_{r1}^2}} \right) \exp \left(- \sum_{j=1}^k \left(\frac{LS_{1,j} - \log \theta_1 - \beta_1 n_j}{2\sigma_{r1}^2} \right)^2 \right) \quad 2.22$$

$$S \leq \tilde{S}_{th}$$

$$f(LS_{2,1}, \dots, LS_{2,k} | \theta_2, \beta_2) = \left(\frac{1}{\sqrt{2\pi\sigma_{r2}^2}} \right) \exp \left(- \sum_{j=1}^k \left(\frac{LS_{2,j} - \log \theta_2 - \beta_2 n_j}{2\sigma_{r2}^2} \right)^2 \right) \quad 2.23$$

$$\tilde{S}_{th} \leq S \leq \tilde{a}_{th}$$

Assumed that $\theta_1, \theta_2, \beta_1$ and β_2 are lognormal random variables with parameters defined above, their a posteriori joint distributions, according to the Bayes theorem are:

$$\begin{aligned} f(\theta_1, \beta_1 | LS_{1,1}, \dots, LS_{1,k}) &= \frac{f(LS_{1,1}, \dots, LS_{1,k} | \theta_1, \beta_1) \Pi(\theta_1) \Pi(\beta_1)}{\int_{-\infty}^{+\infty} f(LS_{1,1}, \dots, LS_{1,k} | \theta_1, \beta_1) \Pi(\theta_1) \Pi(\beta_1) d\theta_1 d\beta_1} \\ & \quad S \leq \tilde{S}_{th} \end{aligned} \quad 2.24$$

$$\begin{aligned} f(\theta_2, \beta_2 | LS_{2,1}, \dots, LS_{2,k}) &= \frac{f(LS_{2,1}, \dots, LS_{2,k} | \theta_2, \beta_2) \Pi(\theta_2) \Pi(\beta_2)}{\int_{-\infty}^{+\infty} f(LS_{2,1}, \dots, LS_{2,k} | \theta_2, \beta_2) \Pi(\theta_2) \Pi(\beta_2) d\theta_2 d\beta_2} \\ & \quad \tilde{S}_{th} \leq S \leq \tilde{a}_{th} \end{aligned}$$

Where $f(LS_{1,1}, \dots, LS_{1,k} | \theta_1, \beta_1)$ and $f(LS_{2,1}, \dots, LS_{2,k} | \theta_2, \beta_2)$ are defined in Eq.2.22 and Eq.2.23 respectively and:

$$\begin{aligned} \Pi(\theta_1) &= \left(\frac{1}{\sqrt{2\pi\theta_1^2\sigma_{\theta_1}^2}} \right) \exp \left(- \frac{1}{2} \left(\frac{\log \theta_1 - \mu_{\theta_1}}{\sigma_{\theta_1}} \right)^2 \right) \\ \Pi(\beta_1) &= \left(\frac{1}{\sqrt{2\pi\beta_1^2\sigma_{\beta_1}^2}} \right) \exp \left(- \frac{1}{2} \left(\frac{\log \beta_1 - \mu_{\beta_1}}{\sigma_{\beta_1}} \right)^2 \right) \\ \Pi(\theta_2) &= \left(\frac{1}{\sqrt{2\pi\theta_2^2\sigma_{\theta_2}^2}} \right) \exp \left(- \frac{1}{2} \left(\frac{\log \theta_2 - \mu_{\theta_2}}{\sigma_{\theta_2}} \right)^2 \right) \\ \Pi(\beta_2) &= \left(\frac{1}{\sqrt{2\pi\beta_2^2\sigma_{\beta_2}^2}} \right) \exp \left(- \frac{1}{2} \left(\frac{\log \beta_2 - \mu_{\beta_2}}{\sigma_{\beta_2}} \right)^2 \right) \end{aligned} \quad 2.25$$

The a posteriori mean of the parameters can be obtained from:

[†] In the Appendix can be found the probability charts of those distributions.

[‡] i is used to denote the belongings of LS to the first ($i = 1$) or second model ($i = 2$) in Eq 2.19.

$$\begin{aligned} \hat{\mu}_{\theta_1} &= \int_{-\infty}^{+\infty} \theta_1 \int_{-\infty}^{+\infty} f(\theta_1, \beta_1 | LS_{1,1}, \dots, LS_{1,k}) d\beta_1 d\theta_1 & TTF(n_k = 0) &= \hat{T}_1 + \hat{T}_2 & 2.28 \\ \hat{\mu}_{\beta_1} &= \int_{-\infty}^{+\infty} \beta_1 \int_{-\infty}^{+\infty} f(\theta_1, \beta_1 | LS_{1,1}, \dots, LS_{1,k}) d\beta_1 d\theta_1 & & & \\ \hat{\mu}_{\theta_2} &= \int_{-\infty}^{+\infty} \theta_2 \int_{-\infty}^{+\infty} f(\theta_2, \beta_2 | LS_{2,1}, \dots, LS_{2,k}) d\beta_2 d\theta_2 & & & \\ \hat{\mu}_{\beta_2} &= \int_{-\infty}^{+\infty} \beta_2 \int_{-\infty}^{+\infty} f(\theta_2, \beta_2 | LS_{2,1}, \dots, LS_{2,k}) d\beta_2 d\theta_2 & & & \end{aligned}$$

2.26

Where \hat{T}_1 and \hat{T}_2 are the a priori pdf distributions of T_1 and T_2 . They can be expressed as:

$$\begin{aligned} \hat{T}_1(n_i | n_k = 0) &= P(LS_1(n_i) \geq \tilde{S}_{th} | \hat{\omega}_1, \hat{\beta}_1) & 2.29 \end{aligned}$$

$$\begin{aligned} \hat{T}_2(n_i | n_k = 0) &= P(LS_2(n_i) \geq \tilde{a}_{th} | \hat{\omega}_2, \hat{\beta}_2) & 2.30 \end{aligned}$$

And their a posteriori variances from:

$$\begin{aligned} \hat{\sigma}_{\theta_1} &= \int_{-\infty}^{+\infty} (\theta_1 - \hat{\mu}_{\theta_1})^2 \int_{-\infty}^{+\infty} f(\theta_1, \beta_1 | LS_{1,1}, \dots, LS_{1,k}) d\beta_1 d\theta_1 \\ \hat{\sigma}_{\beta_1} &= \int_{-\infty}^{+\infty} (\beta_1 - \hat{\mu}_{\beta_1})^2 \int_{-\infty}^{+\infty} f(\theta_1, \beta_1 | LS_{1,1}, \dots, LS_{1,k}) d\beta_1 d\theta_1 & 2.27 \\ \hat{\sigma}_{\theta_2} &= \int_{-\infty}^{+\infty} (\theta_2 - \hat{\mu}_{\theta_2})^2 \int_{-\infty}^{+\infty} f(\theta_2, \beta_2 | LS_{2,1}, \dots, LS_{2,k}) d\beta_2 d\theta_2 \\ \hat{\sigma}_{\beta_2} &= \int_{-\infty}^{+\infty} (\beta_2 - \hat{\mu}_{\beta_2})^2 \int_{-\infty}^{+\infty} f(\theta_2, \beta_2 | LS_{2,1}, \dots, LS_{2,k}) d\beta_2 d\theta_2 \end{aligned}$$

Where $\hat{\omega}_1, \hat{\beta}_1, \hat{\omega}_2$ and $\hat{\beta}_2$ are the a priori pdf of $\omega_1, \omega_2, \beta_1$ and β_2 respectively.

Given that $\hat{\omega}_1, \hat{\omega}_2, \hat{\beta}_1$ and $\hat{\beta}_2$ are normal random variables, the degradation signal LS_1 and LS_2 computed at cycles n_i and n_j respectively, are normal variables as well (N. Gebraeel J. P., 2008)(N. Gebraeel M. L., 2005)(C.J. Lu, 1993) with mean variance given by:

$$\begin{aligned} \mu_{LS_1}(n_i) &= \mu_{\omega_1} + \mu_{\beta_1} n_i & 2.31 \\ \sigma_{LS_1}^2(n_i) &= \sigma_{\omega_1}^2 + \sigma_{\beta_1}^2 n_i^2 & 2.32 \\ &\quad + 2\rho_1 \sigma_{\omega_1} \sigma_{\beta_1} + \sigma_{r_1}^2 \end{aligned}$$

$$\begin{aligned} \mu_{LS_2}(n_j) &= \mu_{\omega_2} + \mu_{\beta_2} n_j & 2.32 \\ \sigma_{LS_2}^2(n_j) &= \sigma_{\omega_2}^2 + \sigma_{\beta_2}^2 n_j^2 & 2.33 \\ &\quad + 2\rho_2 \sigma_{\omega_2} \sigma_{\beta_2} + \sigma_{r_2}^2 \end{aligned}$$

Since the solution to the problem stated has not been found in the statistical literature and recognizing the computation problem associated with solving the equations numerically, we have to make other assumptions on the parameters' pdf functional forms. In order to reduce problem complexity the assumption of β_1 and β_2 as normal distributed parameters is reasonable. This assumption let us to exploit the problem solution proposed by Lindley (D. V. Lindley, 1972) and Gebraeel (N. Gebraeel J. P., 2008). Therefore, $\log \theta_1, \log \theta_2, \beta_1$ and β_2 are assumed to be normal random variables with parameters:

$$\begin{aligned} \log \theta_1 = \omega_1 &= N(\mu_{\omega_1}, \sigma_{\omega_1}) & \log \theta_2 = \omega_2 &= N(\mu_{\omega_2}, \sigma_{\omega_2}) \\ \beta_1 &= N(\mu_{\beta_1}, \sigma_{\beta_1}) & \beta_2 &= N(\mu_{\beta_2}, \sigma_{\beta_2}) \end{aligned}$$

Before proceeding to the formal definition of the problem statement, an assessment of the errors computed after relaxing the hypothesis of lognormal distributed β_1 and β_2 can be done through a comparison of the a priori TTF calculated by the model with β_1 and β_2 as normal random variables with the true TTF computed through the crack growth simulations.

The a priori TTF probability distribution, given the model described by the Eq.2.20, can be computed as the probability that the degradation signal (crack size) LS is smaller than the crack maximum size allowed for each cycle $n_i > 0$, given the a priori model parameters pdfs. The statement, remembering the Eq.2.18, can be formally written as,

Remembering the Eq.2.29 and 2.30, we can write for \hat{T}_1 :

$$\begin{aligned} \hat{T}_1(n_i | n_k = 0) &= 1 - P(LS_1(n_i) \leq \tilde{S}_{th} | \hat{\omega}_1, \hat{\beta}_1) = \\ &= 1 - P\left(Z < \frac{\tilde{S}_{th} - \mu_{LS_1}(n_i)}{\sqrt{\sigma_{LS_1}^2(n_i)}}\right) \\ &= \Phi\left(\frac{\tilde{S}_{th} - \mu_{LS_1}(n_i)}{\sqrt{\sigma_{LS_1}^2(n_i)}}\right) & 2.33 \end{aligned}$$

And for \hat{T}_2 :

$$\begin{aligned}
 \hat{T}_2(n_j | n_k = 0) &= \\
 1 - P(LS_2(n_j) \leq \tilde{a}_{th} | \hat{\omega}_1, \hat{\beta}_1) &= \\
 = 1 - P\left(Z < \frac{\tilde{a}_{th} - \mu_{LS_1}(n_i)}{\sqrt{\sigma^2_{LS_1}(n_i)}}\right) & \\
 = \Phi\left(\frac{\tilde{a}_{th} - \mu_{LS_2}(n_j)}{\sqrt{\sigma^2_{LS_2}(n_j)}}\right) & \quad 2.34
 \end{aligned}$$

Where Φ stands for the standard normal cdf. The domain of \hat{T}_1 and \hat{T}_2 , is ≤ 0 , thus can take on negative values, which is practically impossible from an implementation standpoint. Consequently, we use the truncated cdf for \hat{T}_1 and \hat{T}_2 with the constraint $\hat{T}_i \geq 0$, $i=1,2$ which is given as:

$$\begin{aligned}
 \hat{T}_1 &= \frac{\hat{T}_1 - \hat{T}_1(n_i = 0)}{\hat{T}_1(n_i = 0)} \\
 \hat{T}_2 &= \frac{\hat{T}_2 - \hat{T}_2(n_j = 0)}{\hat{T}_2(n_j = 0)} \quad 2.35
 \end{aligned}$$

As observed by (N. Gebraeel M. L., 2005), \hat{T}_1 and \hat{T}_2 are three parameter truncated Bernstein distributed random variables for which the first and second moment closed form don't exist (A.K Sheikh, 1983). As suggested by (N. Gebraeel M. L., 2005) the median is taken as the central moment. This can be justified, from one side by the non-existence of a closed form for the mean, and for the other hand, considering that the T_i pdfs are skewed and therefore the use of the median is more appropriate and conservative.

To compute the sum of the two random variables the Monte Carlo technique is followed, given the \hat{T}_1 and \hat{T}_2 numerical pdfs shown in Figure 15. The $\hat{\omega}_1, \hat{\beta}_1, \hat{\omega}_2$ and $\hat{\beta}_2$ a priori pdfs parameters are reported in Table 4.

	$\hat{\omega}_1$	$\hat{\beta}_1$	$\hat{\omega}_2$	$\hat{\beta}_2$	ϵ_1	ϵ_2
μ	-10.35	6.95e-009	-8.85	1.07e-007	0	0
σ^2	0.69	6.92e-035	47.65	3.55e-029	1.76e-008	1.5e-005
ρ	-0.1421		-0.2039			

Table 4 $\hat{\omega}_1, \hat{\beta}_1, \hat{\omega}_2$ and $\hat{\beta}_2$ a priori pdfs parameters

The pdfs are simply obtained differentiating the two cdfs with respect to n .

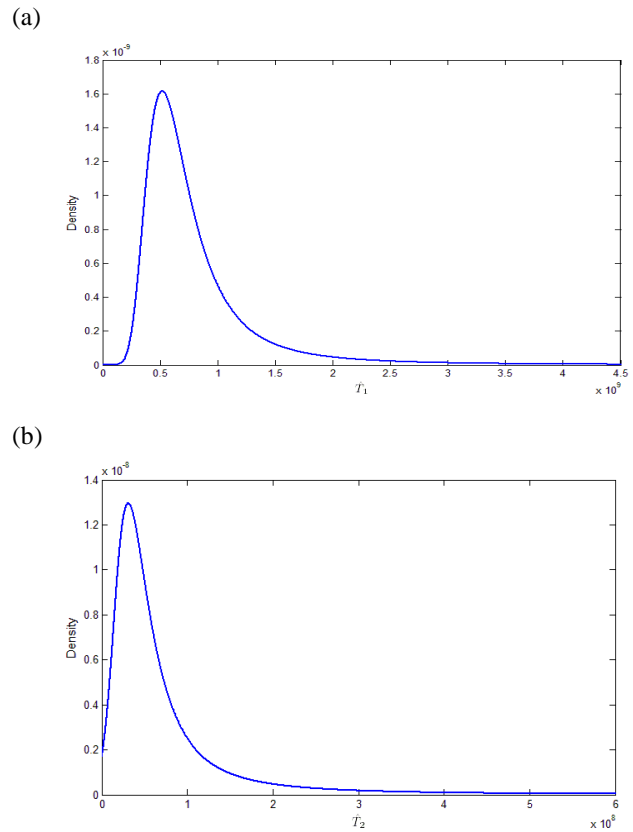


Figure 15 (a) \hat{T}_1 pdf (b) \hat{T}_2 pdf

Eventually the modeled a priori TTF is shown in Figure 16 compared to the simulated a priori TTF on a lognormal probability plot. The green circles belong to the simulated a priori TTF, while the black ones belong to the modeled a priori TTF.

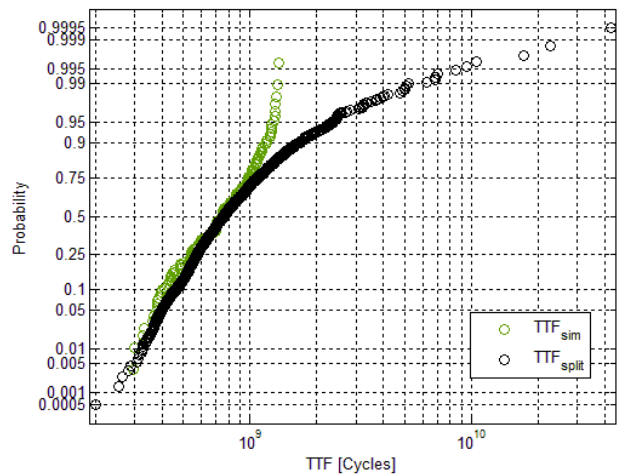


Figure 16 Simulated a priori TTF and a priori modeled TTF comparison – probability plot

A further comparison is between the two TTF pdfs is shown in Figure 17 in which both the cdfs are plotted. From the two figures can be observed that the left hand distributions' tail are similar, while for large values of TTF the two distributions differs. The modeled TTF has the right hand tail longer than the simulated one. However, for our purposes the left hand tail is much more important than the right one. For this reason the $\hat{\beta}_1$ and $\hat{\beta}_2$ normality assumption can be acceptable.

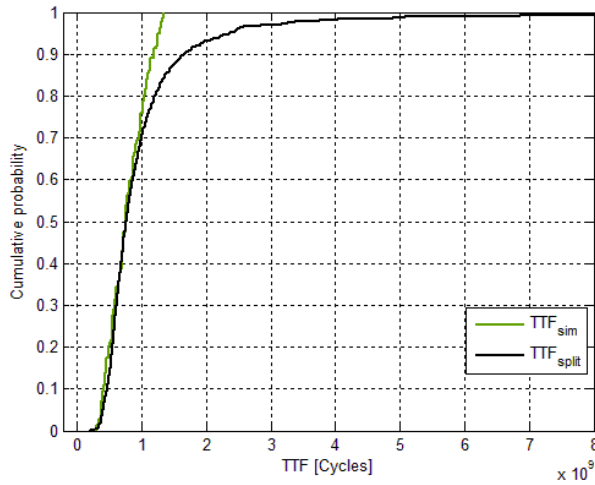


Figure 17 Simulated a priori TTF and a priori modeled TTF comparison – cdf

It is worth noting that if the two model's parameters are somehow correlated, It would be possible to update the second model's parameter instead of using the a priori information to compute the TTF till the threshold \tilde{S}_{th} is reached. This situation would be valuable to exploit because better predictions could be performed since the beginning of the crack growth. Unfortunately this is not the case since the two pairs of coefficients are not significantly correlated as can be observed from Figure 18.

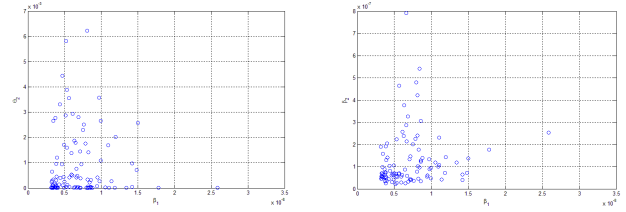
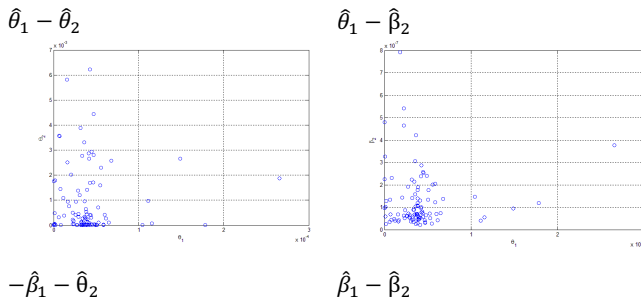


Figure 18 Correlations between the couple of model parameters

Now, once we have computed the a priori parameters' pdfs, we can write the equations that update these pdfs' parameters once obtained the signals $LS_{1,1}, \dots, LS_{1,k}$ or $LS_{2,1}, \dots, LS_{2,k}$ from the monitoring system, depending in which S interval the signals are. Below is just reported the final formulas from which the updated pdfs parameters are obtained.

The models can be rewritten as:

$$\begin{cases} LS_1 = X_1[A]_1 \\ LS_2 = X_2[A]_2 \end{cases} \quad \begin{matrix} S \leq \tilde{S}_{th} \\ \tilde{S}_{th} \leq S \leq \tilde{a}_{th} \end{matrix} \quad 2.36$$

Where:

$$\begin{matrix} [A]_1 & X_1 & [A]_2 & X_2 \\ = \begin{bmatrix} \omega_1 \\ \beta_1 \end{bmatrix} & = \begin{bmatrix} 1 & n_1 \\ \vdots & \vdots \\ 1 & n_n \end{bmatrix} & = \begin{bmatrix} \omega_2 \\ \beta_2 \end{bmatrix} & = \begin{bmatrix} 1 & n_{1,2} \\ \vdots & \vdots \\ 1 & n_{n,2} \end{bmatrix} \end{matrix}$$

At a cycle n_t , given the measures $LS_{i,1}, LS_{i,2}, \dots, LS_{i,t}$, $i = 1,2$ the updated $\omega_1, \beta_1, \omega_2, \beta_2$ pdfs parameters are:

$$\begin{aligned} \tilde{\mu}_1^T = & \left(\left[(X_1^T X_1)^{-1} X_1^T LS_1 \right]^T \frac{X_1^T X_1}{\sigma_{r1}^2} \right. \\ & \left. + \hat{\mu}_1^T \hat{\Sigma}_1^{-1} \right) \left(\frac{X_1^T X_1}{\sigma_{r1}^2} \right. \\ & \left. + \hat{\Sigma}_1^{-1} \right)^{-1} \end{aligned} \quad 2.37$$

$$\tilde{\Sigma}_1 = \left(\frac{X_1^T X_1}{\sigma_{r1}^2} + \hat{\Sigma}_1^{-1} \right)^{-1} \quad 2.38$$

$$\begin{aligned} \tilde{\mu}_2^T = & \left(\left[(X_2^T X_2)^{-1} X_2^T LS_2 \right]^T \frac{X_2^T X_2}{\sigma_{r2}^2} \right. \\ & \left. + \hat{\mu}_2^T \hat{\Sigma}_2^{-1} \right) \left(\frac{X_2^T X_2}{\sigma_{r2}^2} \right. \\ & \left. + \hat{\Sigma}_2^{-1} \right)^{-1} \end{aligned} \quad 2.39$$

$$\tilde{\Sigma}_2 = \left(\frac{X_2^T X_2}{\sigma_{r2}^2} + \hat{\Sigma}_2^{-1} \right)^{-1} \quad 2.40$$

Where:

$$\begin{aligned} \hat{\mu}_1 &= [\mu_{\omega 1} \quad \mu_{\beta 1}] & \hat{\mu}_2 &= [\mu_{\omega 2} \quad \mu_{\beta 2}] \\ \hat{\Sigma}_1 &= \begin{bmatrix} \sigma_{\omega 1} & \sigma_{\omega 1, \beta 1} \\ \sigma_{\omega 1, \beta 1} & \sigma_{\beta 1} \end{bmatrix} & \hat{\Sigma}_2 &= \begin{bmatrix} \sigma_{\omega 2} & \sigma_{\omega 2, \beta 2} \\ \sigma_{\omega 2, \beta 2} & \sigma_{\beta 2} \end{bmatrix} \end{aligned}$$

are the vectors of the a priori pdfs means and the covariance matrixes while:

$$\begin{aligned} \tilde{\mu}_1 &= [\tilde{\mu}_{\omega 1} \quad \tilde{\mu}_{\beta 1}] & \tilde{\mu}_2 &= [\tilde{\mu}_{\omega 2} \quad \tilde{\mu}_{\beta 2}] \\ \tilde{\Sigma}_1 &= \begin{bmatrix} \tilde{\sigma}_{\omega 1} & \tilde{\sigma}_{\omega 1, \beta 1} \\ \tilde{\sigma}_{\omega 1, \beta 1} & \tilde{\sigma}_{\beta 1} \end{bmatrix} & \tilde{\Sigma}_2 &= \begin{bmatrix} \tilde{\sigma}_{\omega 2} & \tilde{\sigma}_{\omega 2, \beta 2} \\ \tilde{\sigma}_{\omega 2, \beta 2} & \tilde{\sigma}_{\beta 2} \end{bmatrix} \end{aligned}$$

are the vectors of the a posteriori pdfs means and the covariance matrixes.

Now, given the a posteriori pdfs' parameters the T_1 or T_2 distribution can be computed.

Remembering Eq.2.31 and 2.32 the updated mean and the variance of the degradation signal at a cycle n_i or n_j will be:

$$\tilde{\mu}_{LS_1}(n_i) = \tilde{\mu}_{\omega 1} + \tilde{\mu}_{\beta 1} n_i \quad 2.41$$

$$\tilde{\sigma}_{LS_1}^2(n_i) = \tilde{\sigma}_{\omega 1}^2 + \tilde{\sigma}_{\beta 1}^2 n_i^2 + 2\tilde{\rho}_1 \sigma_{\omega 1} \tilde{\sigma}_{\beta 1} + \sigma_{r1}^2$$

$$\tilde{\mu}_{LS_2}(n_j) = \tilde{\mu}_{\omega 2} + \tilde{\mu}_{\beta 2} n_j \quad 2.42$$

$$\tilde{\sigma}_{LS_2}^2(n_j) = \tilde{\sigma}_{\omega 2}^2 + \tilde{\sigma}_{\beta 2}^2 n_j^2 + 2\tilde{\rho}_2 \tilde{\sigma}_{\omega 2} \tilde{\sigma}_{\beta 2} + \sigma_{r2}^2$$

And therefore from Eq.2.33 and 2.34 the updated T_1 or T_2 pdf will be:

$$\begin{aligned} \tilde{T}_1(n_i | LS_{1,1} \quad LS_{1,2}, \dots, LS_{1,t}) \\ = \Phi \left(\frac{\tilde{S}_{th} - \mu_{LS_1}(n_i)}{\sqrt{\sigma_{LS_1}^2(n_i)}} \right) \\ \xrightarrow{\tilde{T}_1 \geq 0} \frac{\tilde{T}_1 - \tilde{T}_1(0)}{\tilde{T}_1(0)} \end{aligned} \quad 2.43$$

And for \tilde{T}_2 :

$$\begin{aligned} \tilde{T}_2(n_j | LS_{2,1} \quad LS_{2,2}, \dots, LS_{2,t}) \\ = \Phi \left(\frac{\tilde{a}_{th} - \mu_{LS_2}(n_j)}{\sqrt{\sigma_{LS_2}^2(n_j)}} \right) \\ \xrightarrow{\tilde{T}_2 \geq 0} \frac{\tilde{T}_2 - \tilde{T}_2(0)}{\tilde{T}_2(0)} \end{aligned} \quad 2.44$$

An Example:

Given a crack growth path shown in Figure 19, at each time step we can update the a priori TTF given in Figure 2, exploiting the information gained from monitoring the crack growth.

Using Eq.2.37, 2.38 for the first part of the degradation pattern (T_1 in Figure 19) and the Eq.2.39 and 2.40 for the second part, we can compute the a posteriori $\hat{\omega}_1, \hat{\beta}_1, \hat{\omega}_2$ and $\hat{\beta}_2$ pdfs' parameters, that are the means and the standard deviations.

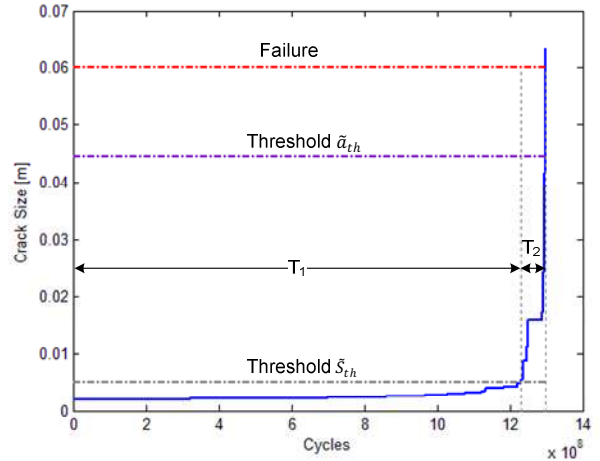
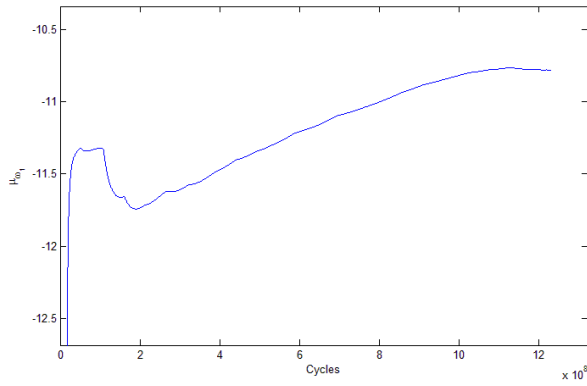


Figure 19 Crack growth path

From the initial cycle to that one that corresponds to a crack size of 5.1 mm the updated TTF is given by Eq.2.7 where T_2 is given by Eq.2.35, that is the a priori modeled T_2 .

a)



b)

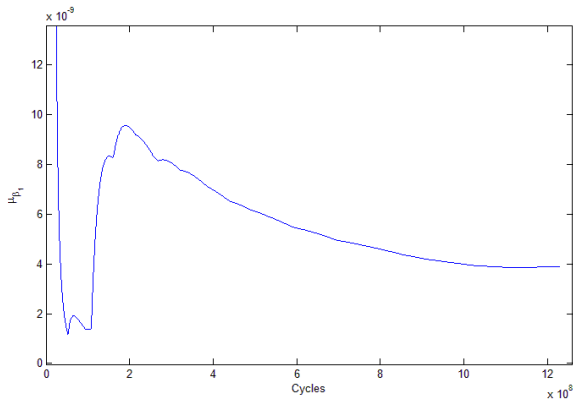


Figure 20 a) updated $\tilde{\mu}_{\omega_1}$ and b) updated $\tilde{\mu}_{\beta_1}$
 b) shows the updated $\tilde{\mu}_{\omega_1}$ as a function of cycles, while the plot b) shows the updated $\tilde{\mu}_{\beta_1}$.

At each time step, given the updated $\tilde{\mu}_{\omega_1}$ and $\tilde{\mu}_{\beta_1}$ we can compute the actual TTF where \tilde{T}_1 is given by the Eq.2.43. For each time step the TTF median and the 1st percentile is stored. These two values are plotted in Figure 21. As can be observed, cycle after cycle the predictions converge to the true TTF even before the second degradation phase. In this case, both the 1st percentile and the mean fall within the 5% error interval. The interval in which the TTF median and its 1st percentile lines are interrupted means that the predicted TTF falls beyond the timescale.

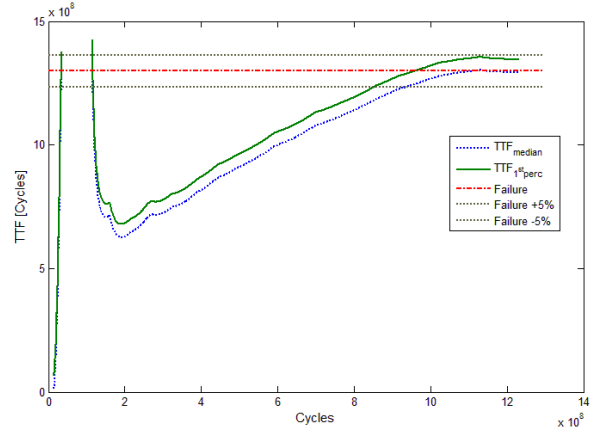


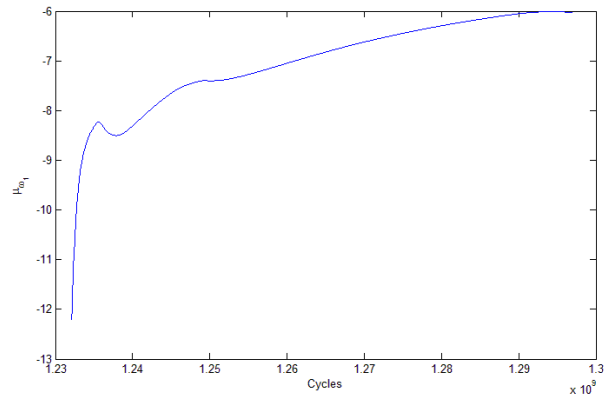
Figure 21 Predicted TTF - 1st phase

Once the threshold \tilde{S}_{th} is passed, the TTF is equal to the cycle T_1 , that is no more a random variable (it is deterministic), plus the predicted \tilde{T}_2 .

\tilde{T}_2 is given by Eq.2.44, once computed the updated μ_{β_2} , μ_{ω_2} and the related variances given by Eq.2.39 and 2.40.

Figure 22 shows the updated μ_{β_2} and μ_{ω_2} respectively.

a)



b)

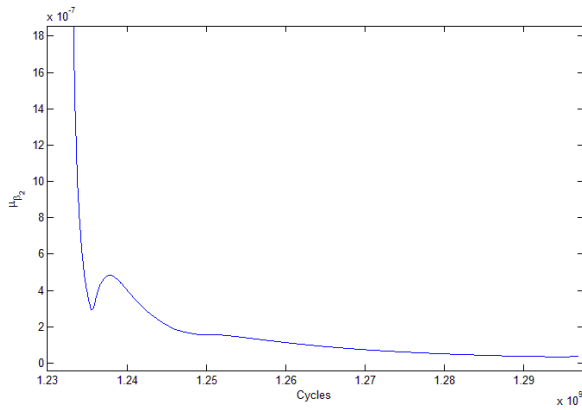


Figure 22 a) updated $\tilde{\mu}_{\omega_2}$ and b) updated $\tilde{\mu}_{\beta_2}$

As previously done for the first degradation phase, the *TTF* pdf can be computed using Eq.2.39, 2.40, 2.42 and eventually 2.44. The updated *TTF* median and its 1st percentile are shown in Figure 23.

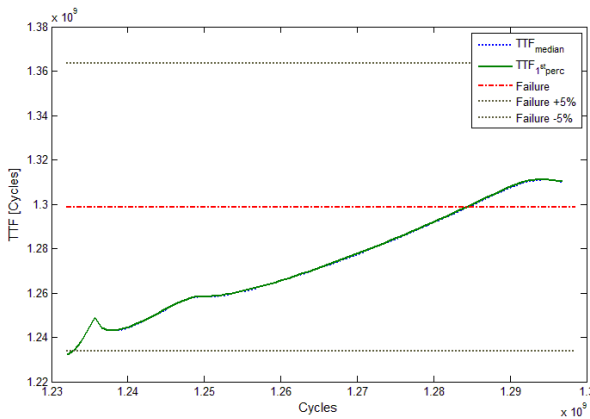


Figure 23 Predicted *TTF* – 2nd phase

Can be observed how the predictions converge to actual failure time. This time the prediction variances are smaller than those of the first phase. This is due to the fact that the 1st phase predictions include the uncertainties related to the a priori T_2 pdf.

2.3.3 Prognostic through the physical model

The same problem faced by the Bayesian prognostic model can be pursued through a recursive application of the crack growth model presented in paragraph 2.1. The physical phenomenon analyzed in this context has been faced by numerous researches, therefore numerous models have been proposed capable of describing and highlighting the main variables and their relations that influence the crack growth. The NASGRO model used in this context is recognized to be the most reliable to describe crack growth in railway

axles (S. Beretta M. C., 2006) (U. Zerbst M. V., 2005) (S. Beretta M. C., 2004), therefore can be used to predict accurately the *TTF*.

The main idea at the basis of this approach is that, once measured and estimated the actual crack size and the loads history, we can estimate the *TTF* through simulating the possible growth paths by using a Monte Carlo technique.

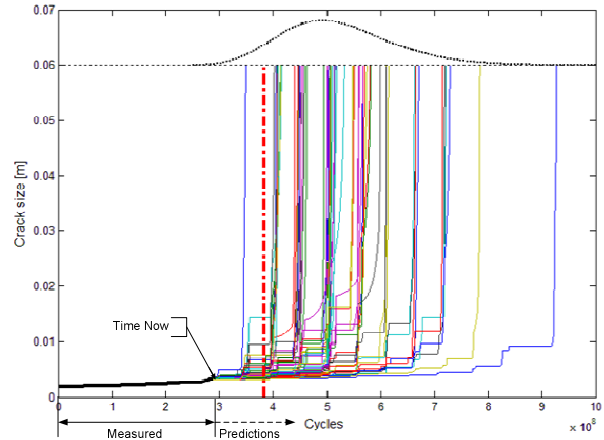


Figure 24 *TTF* prediction through the NASGRO crack growth model

This approach is shown in Figure 24. Let suppose that through the monitoring infrastructure we have measured the crack size at the time now, we can simulate the crack propagation considering as random variables the load applied and the SIF threshold and the initial crack size equal to the measured one. The functions plotted and originating from the time now, are some simulated crack growth paths. Starting from the crack growth paths set, it is possible to estimate the *TTF* pdf. In Figure 24 the black dotted line represents the predicted *TTF* pdf, while the red line represents the actual failure time.

The estimated *TTF* at each time step can be approximated by lognormal distribution, as shown in Figure 25.

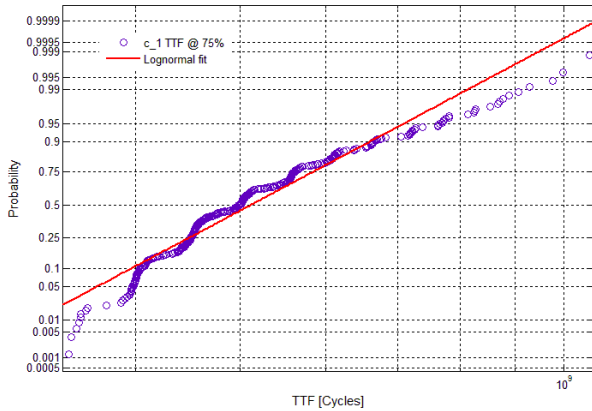


Figure 25 The approximated TTF probability plot

As in the Bayesian approach, at each time step, the TTF 1st percentile, the median and the TTF at 98% level of confidence is stored. However, for computational reasons, the TTF updating times are set at the 5%, to the 99% of the actual TTF with a 5% gap. Figure 27 shows the TTF estimations at different time steps. Can be observed how the predictions converge to the actual failure. At the last updating time step all the TTF distributions's lower and upper bounds fall into the 5% error interval.

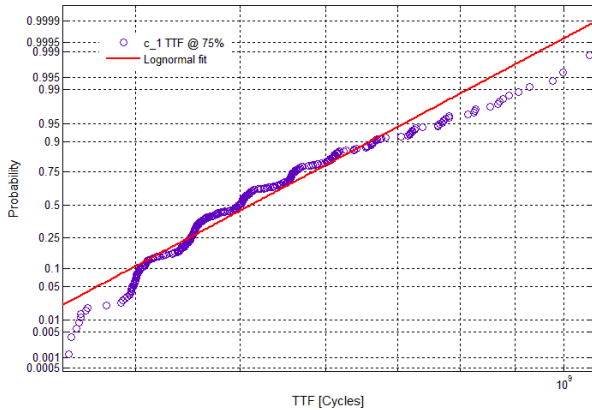


Figure 26 The approximated TTF probability plot

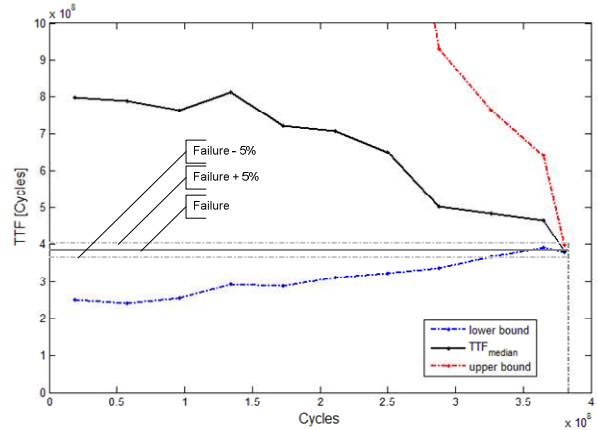


Figure 27 TTF predictions

Figure 28 shows how the confidence interval diminish as we approach to the actual failure. The green dotted line represents the difference between the TTF median and the TTF at the 0.01 confidence level, while the red dashed dotted line represents the TTF pdf upper bound, at the 0.99 confidence level.

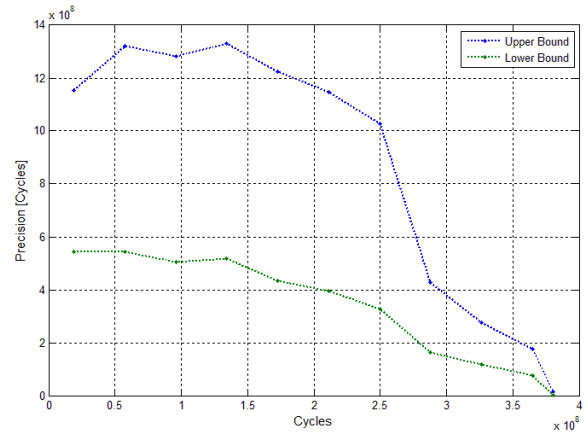


Figure 28 Estimated TTF at the 0.01 and 0.98 confidence level

2.3.4 The size error and the updating frequency effect on TTF predictions

In the case of the physical model, the size error and the updating frequency effect on the estimations can be approximately evaluated through simple geometrical considerations. The assessment of these effects on the predictions performances is an important issue since they characterize the monitoring and diagnostic equipment goodness. Higher size errors characterize low performance diagnostic, while lower updating frequency entails lower monitoring equipment cost.

In this case the effect of the updating frequency on the prediction performances is not relevant since the *TTF* estimation relies on just the last crack size measurement and not, as in the Bayesian case, on the complete set of measurements. The *TTF* updating frequency effect can be considerable when maintenance scheduling decisions is considered. By this point of view, high frequency updating is preferable since the decisions can be based on an updated *TTF*.

In this case we can apply a predictive maintenance policy similar to that one proposed by Kaiser et.al. in (N.Z Gebraeel, 2009). The stopping rule, i.e the cycle n_k at which the axle should be substituted, is defined as in Eq.2.45.

$$n_k \rightarrow TTF_{lb}(n_k) - n_k - \delta \leq 0 \quad 2.45$$

Where n_k is the first cycle at which the rule is verified, $TTF_{lb}(n_k)$ is the *TTF* prediction computed at a 0.01 confidence level at the cycle n_k , δ is the updating interval. From this simple rule is self-evident that the greater δ the lower n_k .

This simple rule can be easily understood by analyzing the graph shown in Figure 29. The blue line represents the estimated *TTF* at the 0.01 confidence level while the black dotted line represents the equality $n = TTF_{lb}$. The dashed line represents the equality $n = TTF_{lb} + \delta$. Therefore, for Eq.2.45, the cycle n_k is the first intersection point of the TTF_{lb} (blue line) with the black dashed line. Particularly, referring to what stated in the previous chapters, the quantity $TTF_{lb}(n_k) - n_k$ is the RUL computed at the 0.01 confidence level (RUL_ in Figure 29). The main idea associated with this rule is that the axle can be safely run till it reaches the last TTF_{lb} estimation.

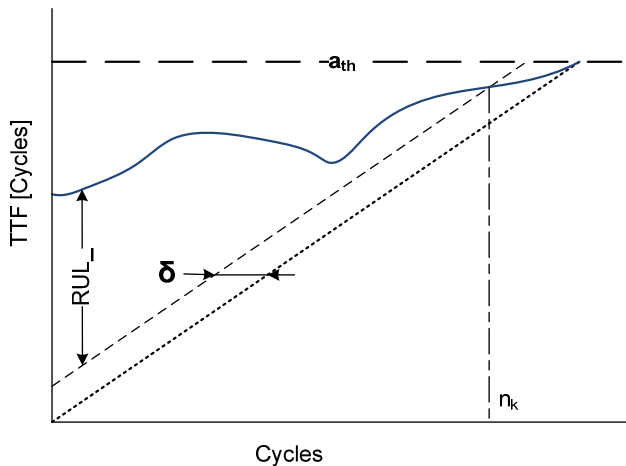


Figure 29 The effect of updating frequency on *TTF* predictions

The size error effect on the *TTF* predictions can be approximately computed making the hypothesis that the crack growth path can be approximated with an exponential function. Generally, as described in 2.3.1, the more the size error the lesser the threshold. The analysis framework is shown in Figure 30. Let us suppose that for a given size error, the failure threshold is set at the value a_{th} and that we are at the cycle n_i and we measure the crack size $exp(LS_i)$. Through the method explained in paragraph 2.3.3, we can compute the *TTF* pdf (blue line) and therefore we know the TTF_{median} and the TTF_{1st_p} at the 0.01 confidence level.

Next, suppose that the new size error is greater to the previous one, consequently, from Eq.2.17 keeping a_{max} constant, we obtain the failure threshold a_{th2} lower than a_{th} . This threshold shift causes a change in the *TTF* pdf parameters and therefore to the reference points TTF_{median} and TTF_{1st_p} .

The new reference points TTF'_{median} and TTF'_{1st_p} computed at cycle n_i , thanks to the hypothesis made, can be computed as follows:

$$TTF'_{median} = TTF_{median} - \frac{\log a_{th} - \log a_{th2}}{\beta} \quad 2.46$$

$$TTF'_{1st_p} = TTF_{1st_p} - \frac{\log a_{th} - \log a_{th2}}{\alpha} \quad 2.47$$

Where:

$$\alpha = \frac{\log a_{th} - LS_i}{TTF_{1st_p} - n_i} \quad 2.48$$

$$\beta = \frac{\log a_{th} - LS_i}{TTF_{1st_p} - n_i} \quad 2.49$$

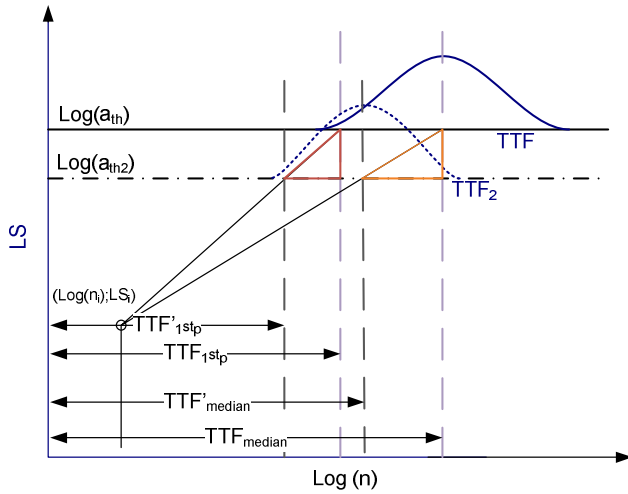


Figure 30 The error size effect on TTF predictions

The lower confidence interval $CI = (TTF_{median} - TTF_{1stp})$, decreases when the size error increases, i.e the prediction is more accurate. This can be easily demonstrated, subtracting term by term Eq. 2.46 with Eq. 2.47 we obtain:

$$CI' = CI - \Delta(\log a_{th}) \left(\frac{1}{\beta} - \frac{1}{\alpha} \right) \quad 2.50$$

Since $\beta < \alpha$ and $\Delta(\log a_{th}) > 0$ for increasing size errors $CI' < CI$.

It is worth noting that, from Eq. 2.47, the ratio $\frac{TTF_{1stp}}{TTF'_{1stp}}$ is not linear with respect to the ratio $\frac{a_{th}}{a_{th2}}$ and from Eq.2.17 the ratio $\frac{a_{th}}{a_{th2}}$ is not a linear function of the size error ratio.

The updating frequency and size error combined effect on the cycle n_k normalized with respect the actual failure (i.e % of the life exploited) on particular crack growth curve is shown in Figure 31. As we can see the relationship between the size error and the ratio $\frac{n_k}{n_{failure}}$. As the size error increases, for a given updating frequency, the life exploited decreases, while the relationship between the updating frequency and the life exploited for a given size error is linear: the more frequent the TTF updating the greater the life exploited.

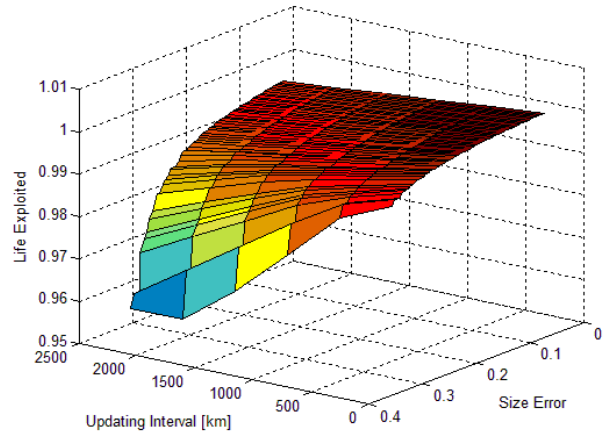


Figure 31 The updating frequency and size error combined effect

3. RESULTS

Our goal, as stated in paragraph 2, is to assess the predictive performances of both the prognostic models and eventually highlight the differences between the predictive and preventive maintenance policy.

The probabilistic aspect of the issue has clearly arisen during the dissertation, therefore a reliable and a definitive answer to the questions proposed has to be given after numerous simulations that guarantee a reliable representation of the probabilistic aspects involved. However, some preliminary considerations can be outlined analyzing a limited number of instances.

The method used to select the instances analyzed is based on the stratified sampling technique. Particularly, the TTF pdf represented in Figure 2 has been divided in 10 equal spaced intervals, that corresponds to the bins shown in the same figure. For each bin a crack growth path was selected obtaining a set of 10 possible degradation curves as shown in Figure 32.

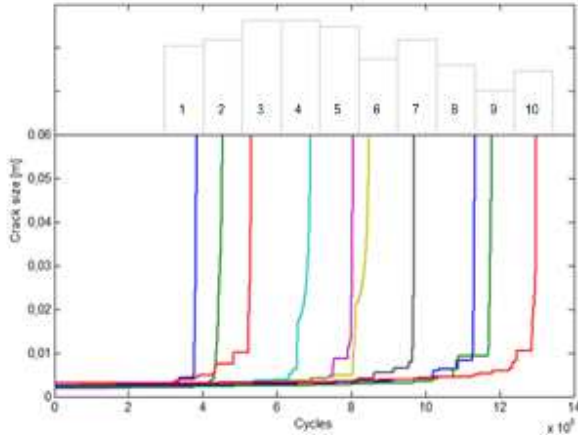


Figure 32: The 10 crack growth paths

For the whole set of track selected, the Bayesian prognostic algorithm and the physical model was applied. Moreover, the maximum number of inspections N_{insp} and the expected number of inspections \overline{N}_{insp} was computed.

In order to evaluate the prognostic algorithms described, two metrics were used, one of which suggested by (A.Saxena, 2008).

This metric, called Timeliness φ , exponentially weighs RUL prediction errors through an asymmetric weighting function. Penalizes the late predictions more than early prediction. The formula is:

$$\Phi(n) = \begin{cases} \exp\left(\frac{|z(n)|}{a}\right) & z \geq 0 \\ \exp\left(\frac{|z(n)|}{b}\right) & z \leq 0 \end{cases} \quad 3.1$$

$$\varphi = \frac{1}{N} \sum_{n=1}^N \Phi(n) \quad 3.2$$

Where $z(n) = TTF_{actual} - TTF_{median}(n)$ is the prediction error computed at cycle n , while a and b are two constants where $a > b$. In this case $a = 100$ and $b = 10$.

Ideally the perfect score is $\varphi=1$. To be comparable, the updating frequency has to be the same between the two algorithms, therefore the TTF predictions in the physical model case have been linearly interpolated.

The other metric chosen is simply the predictions percentage error computed at fixed time steps $n_k = 0.25FT, 0.5FT, 0.75FT, 0.98FT$, where FT is the cycle at which the failure occurs.

In the appendix the comparison of the predictions at different time steps and the PC_{DET} for each of 10 sampled

paths can be found. Moreover, the size error and the updating frequency effect on the exploited life are plotted for each instance.

As can be noticed from these figures, both the algorithms' predictions converge to the actual failure time. The information about the actual degradation path increase as time elapses, resulting in an improved knowledge about the actual TTF. Better knowledge of the crack growth behavior allow more accurate predictions. The advantage of continuous monitoring with respect to the a priori information is clearly evident observing Figure 33. It shows the TTF pdf obtained from the prognostic algorithms described and the a priori TTF pdf (black line). It is clearly noticeable how prognostics can improve the knowledge on the actual failure path followed by an individual axle.

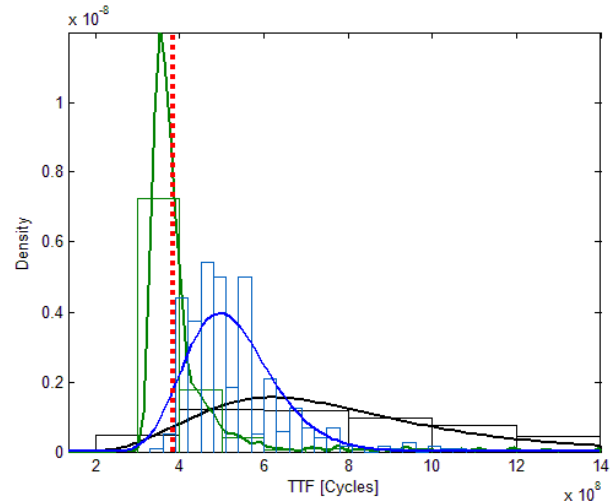


Figure 33 Comparison of the a priori TTF pdf and the updated TTF pdf obtained from the prognostics algorithms described (green-Bayesian, blue physical based model, black - a priori)

However, substantial differences among the two prognostic approach exists. Particularly, what differs is the distribution of the prediction errors along the degradation timeline and the prediction confidence interval. The last statement is evident observing the figures in the appendix in which the predictions paths are compared. In all the instances selected the physical model confidence interval is larger than that one computed by the Bayesian approach.

However, the most important differences among the two approaches have to be evaluated in term of the prediction errors. The following graphs display the prediction errors for both the algorithms and for the whole crack growth track set at fixed residual life percentile (i.e 0.25, 0.5, 0.75, 0.98). The same information are displayed in a tabular form in Table 5. The percentage prediction error is simply calculated as:

$$err\% = \frac{FT - TTF_{median}}{FT} 100 \quad 3.3$$

From the graphs can be concluded that:

1. Physical model prediction errors decrease approaching the FT
2. Bayesian algorithm prediction errors decreases till the 75° percentile of the residual lifetime, while at 98% the errors are greater that in the 75 percentile
3. Physical model predictions are lower for FT near the average (bins 3,4,5)
4. Bayesian predictions seems to outperform the physical model predictions for till the 75th percentile, while for the 98th the physical model predictions are more accurate.

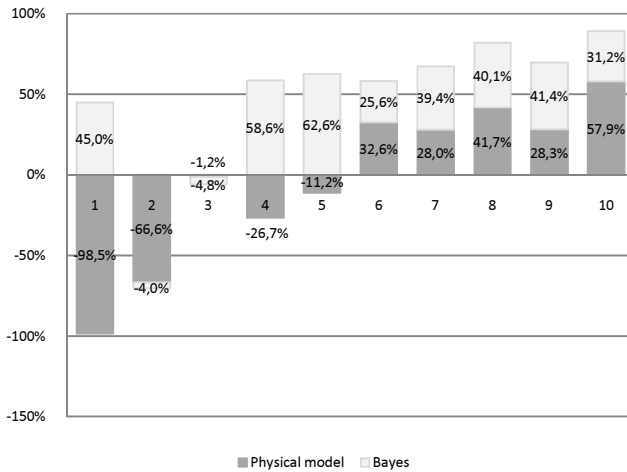


Figure 34 Percentage prediction error @ 25% FT

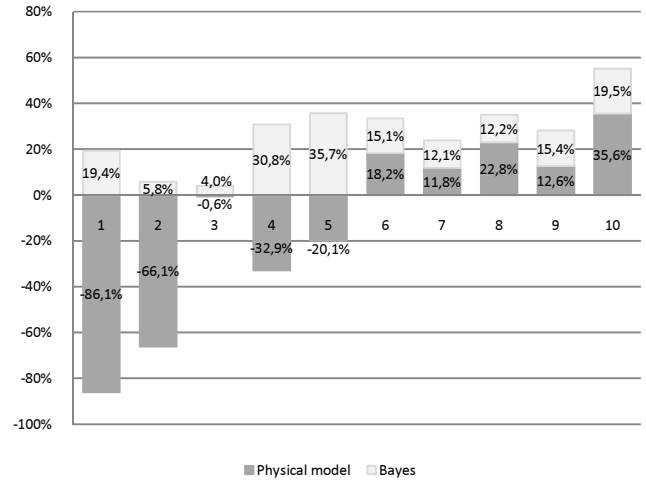


Figure 35 Percentage prediction error @ 50% FT

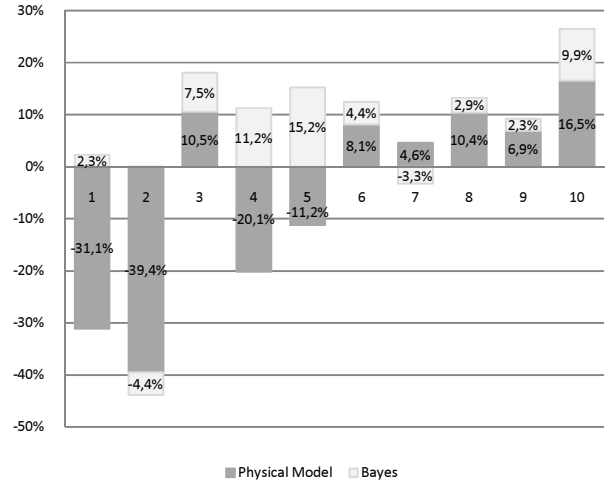


Figure 36 Percentage prediction error @ 75% FT

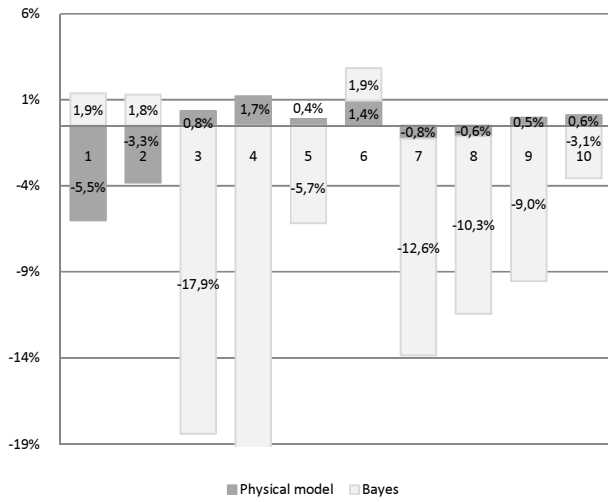


Figure 37 Percentage prediction error @ 98% FT

% Life	Model	c_1	c_2	c_3	c_4	c_5
25	Physical model	-98,50%	-66,59%	-1,20%	26,69%	11,19%
	Bayes	44,95%	-3,96%	-4,85%	58,58%	62,57%
50	Physical model	-86,08%	-66,11%	-0,55%	32,95%	20,08%
	Bayes	19,36%	5,77%	4,00%	30,77%	35,70%
75	Physical model	-31,09%	-39,44%	10,52%	20,11%	11,21%
	Bayes	2,31%	-4,42%	7,53%	11,24%	15,22%
98	Physical model	-5,52%	-3,31%	0,84%	1,72%	0,39%
	Bayes	1,87%	1,81%	-	-	-5,68%
25	Physical model	32,56%	27,98%	41,74%	28,29%	57,91%
	Bayes	25,64%	39,37%	40,14%	41,39%	31,19%
50	Physical model	18,24%	11,76%	22,84%	12,63%	35,59%
	Bayes	15,11%	12,08%	12,15%	15,37%	19,54%
75	Physical model	8,09%	4,64%	10,39%	6,87%	16,52%
	Bayes	4,38%	-3,27%	2,87%	2,34%	9,91%
98	Physical model	1,44%	-0,76%	-0,61%	0,48%	0,61%
	Bayes	1,90%	-12,61%	-	-9,04%	-3,06%

Table 5 Percentage prediction errors

General considerations can be drafted from the conclusive graph in Figure 38 that displays the mean squared percentage error among the whole set for each residual life percentile. The statements of the list above are confirmed.

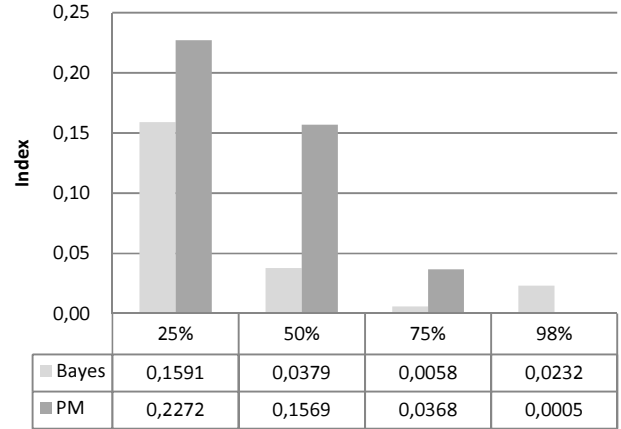


Figure 38 MS of the percentage prediction errors for each residual life percentile

Using the other metric chosen, expressed by Eq.3.2 the results displayed in Table 6 are obtained. The main difference between the metric defined before, is that this metric considers the whole set of predictions and not only those that corresponds to particular moments. The results found are very similar among the two approaches. The physical model index is slightly smaller than the Bayesian one.

	Physical model	Bayes	NDI - max	NDI - mean
c_1	1.07595	1.02061	34	33.24
c_2	1.05471	1.00486	40	39.49
c_3	1.00225	1.02235	47	42.41
c_4	1.02188	1.01337	61	58.98
c_5	1.01014	1.01547	71	68.61
c_6	1.00199	1.00774	75	73.04
c_7	1.00143	1.00769	86	82.14
c_8	1.00251	1.00787	100	96.19
c_9	1.00163	1.00240	105	100.35
c_10	1.00355	1.00484	115	110.70
MS	1.01791	1.01074		

Table 6 Results – φ , N_{insp} and \bar{N}_{insp}

The last two columns of Table 6 reports respectively the maximum non destructive inspections number and the expected NDI number. The last result is obtained multiplying the NDI cumulative number with the corresponding PC_{DET} .

Obviously, the expected NDI number increases as the FT increases. The NDI number that should be performed to guarantee a 99% chance to detect a crack before it reaches the length of 6cm is relevant. As a consequence, the availability of the asset is highly affected from this maintenance policy. The loose of availability and the numerous maintenance activities imply a considerable maintenance costs build up.

In Figure 39 the effect of an increase of the size error is displayed[§], considering the updating frequency of 90 km. Can be noticed that generally, as previously stated, the greater the size error, the lower the life exploited. However, the life exploited reduction is not relevant. An increase of 3 times of the size error causes a life exploited reduction of about 5% on average. For the figures in appendix can be noticed that the effect of the updating frequency is lower with respect to the error size effect.

The scarce effect of this important variables to the exploited life is due to the fact that an increase of the size error cause a reduction of the threshold a_{th} that however corresponds to a negligible life loss reduction thanks to the high crack growth rate that characterize the last part of the degradation phase. Greater effects shall be noticed when the size error is large enough to force the threshold a_{th} to be set at crack sizes at which the growth rate is lower (i.e at the end of the first degradation phase).

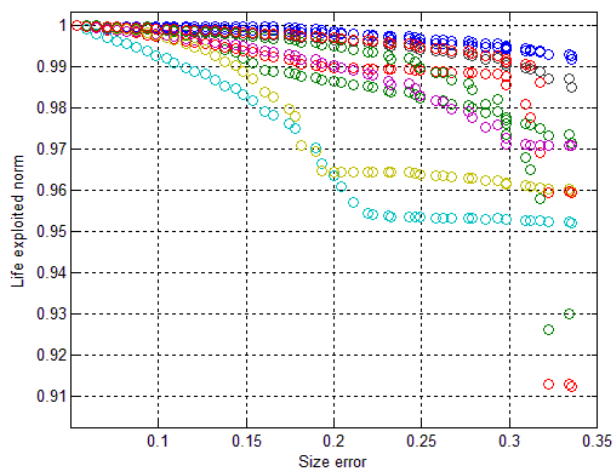


Figure 39 The size error effect on life exploited given $\delta = 90 km^{**}$

[§] Computed considering the physical model predictions only

** Life exploited is normalized with respect to the life exploited that corresponds to the first size error considered

4. CONCLUSIONS

The objective of this research was to propose an approach to a condition based maintenance policy assessment in order to preliminary evaluate its benefits and to understand the main variables that influence the overall approach performance. Particularly, an explanatory study was carried out to evaluate the possibility to introduce prognostic concepts into railway axle maintenance management.

Through a reliable probabilistic crack growth model a comparison between a prognostic maintenance approach based on Bayesian probabilistic theory, a prognostic maintenance approach based on the same crack growth physical model and the classical preventive maintenance policy based on regular NDT was carried out. The probabilistic crack growth model considers the SIF as a random normal variable and a random load history derived from measured load spectra. The diagnostic-monitoring infrastructure precision was described by a size error, directly derived from the calibration curve of an ultrasonic NDT. Assuming the hypothesis introduced in paragraph 2.3.4, the results suggests that further research should be conducted validating the approach proposed on a real case study. As matter of facts both the prognostic algorithms described guarantee an average absolute predictions errors lower than 50 % at 25% of the actual axle life. The later predictions guarantees lower prediction errors, approaching the 7% on average. Earlier predictions errors are generally lower for the Bayesian prognostic algorithm than those computed through the physical model. Whereas, for later predictions the physical model seem to provide more accurate RUL estimations. However, the gap between predictions error computed by the two models are, on average, comparable. The effect of the updating frequency and the size error on predictions errors in case of prognostic physical model algorithm scenario and therefore, on the overall approach performance (life exploited with a determined reliability threshold) is assessed as well. The results show that the higher the size error and the lower updating frequency the lower life exploited. However the effect of updating frequency and size error in terms of life exploited is limited till the maximum crack size threshold, derived from the error size of the diagnostic infrastructure, becomes lower than about 5 mm, i.e the crack size at which the crack growth rate significantly increases.

Generally speaking, a PHM approach needs a deep system/component knowledge. This need implies high investment costs to perform experimental tests (high fixed costs). System/component knowledge in high safety requirement environments, such as in the aviation industry, has to be known before commissioning for obvious safety reasons. Low Accuracy PHM May Be Worse Than No PHM. Costs and the benefits resulting from a prognostic approach could be distributed differently across the actors involved, therefore an “integrator” that manages all the

process is suggested or partnership between the main actors involved committed to share the investment costs. Moreover it is worth noting that a trade off exists between system usage pattern and the resulting benefits, higher usage allows a better return on investment but lowers t_{ADV} , i.e. the main prognostic benefits driver.

After all these considerations, it is possible to sum up the results in the matrix displayed in Figure 40. Profitability of a PHM approach can be thought as a function of two variables:

- Component criticality
- Easiness to acquire data of component's failure modes

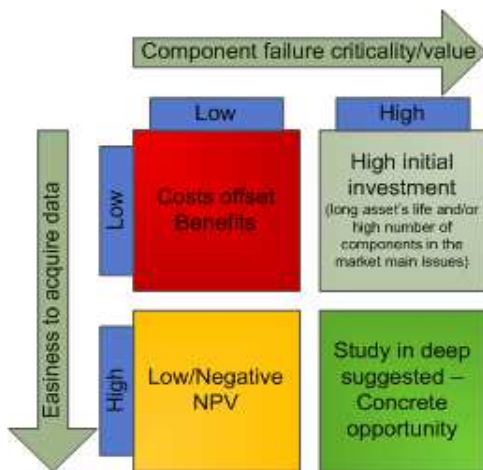


Figure 40: PHM applicability

Difficulties to describe and acquire data on the component's failure behavior imply high R&D costs while the components criticality and value can boost the benefits allowed by a PHM approach. The case in which a PHM approach is suggested is the case in which it is easy to acquire and data and knowledge on the component failure behavior and in which the component monitored and maintained is critical for the whole system availability and/or it has a very high value. In the other two situations further investigation aimed to better estimate the costs and the benefits involved is suggested.

ACKNOWLEDGEMENT

The author wishes to thank professors S. Beretta of the Polytechnic Institute of Milan (Italy) and G. Jacazio of the Polytechnic Institute of Turin (Italy) for their encouragement and support in the preparation of the paper

M. Vismara, Milan, 09/09/1985. Bachelor degree in Transport Engineering at Polytechnic of Milan. Master

degree in Mechanical Engineering at Polytechnic of Milan and Turin, ASP diploma (High Polytechnic School).

Maintenance Engineer at Hupac SA in Chiasso, Switzerland

REFERENCES

A.K. Sheikh, M. A. (1983). Renewal Analysis Using Bernstein Distribution. *Reliability Engineering*, 5, 1-19.

A. Saxena, J. C. (2008). Metrics for Evaluating Performance of Prognostic Techniques. *International Conference on prognostics and health management*, (pp. 1-17). Denver, CO.

Anonymus. (2006). Fracture Mechanics and Fatigue Crack Growth 4.2. *NASA Technical report*.

C.J. Lu, W. M. (1993). Using Degradation Measures to Estimate a Time-to-Failure Distribution. *American Society for Quality*, 35 (2), 161-174.

C.P. Lonsdale, D. S. (2004). North American axle failure experience. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 218 (4), 293-298.

D. V. Lindley, A. F. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B, Statistical*, 34 (1), 1-41.

D.A. Virkler, B. H. (1979). The statistical nature of fatigue crack propagation. *ASME, Transactions, Journal of Engineering Materials and Technology*, 101, 148-153.

EN13103. (2001). *Railway applications – wheelsets and bogies – non powered axles – design method*.

Gassner, E. (1956). Performance fatigue testing with respect to aircraft design. In E. Gassner, *Fatigue in Aircraft Structures*. New York: Academic Press.

Hoddinot, D. (2004). Railway axle failure investigations and fatigue crack growth monitoring of an axle. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 218, 283-292.

J.L. Bogdanoff, F. K. (1985). *Probabilistic models of cumulative damage*. New York: John Wiley & Sons.

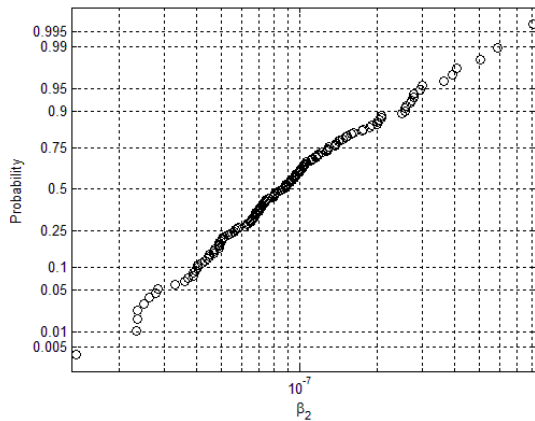
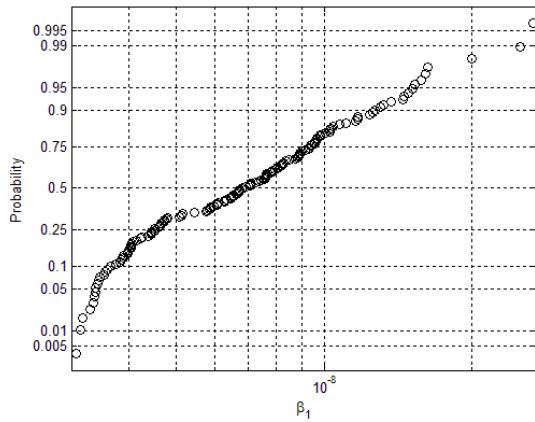
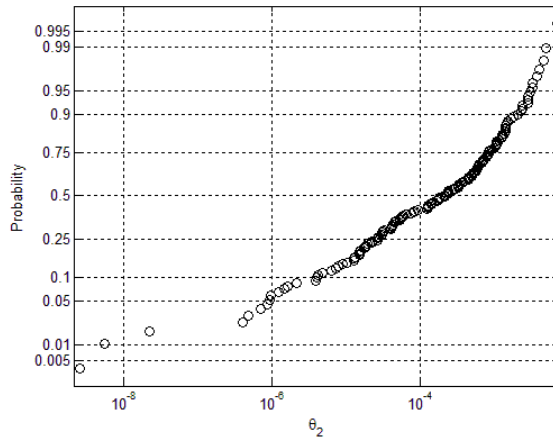
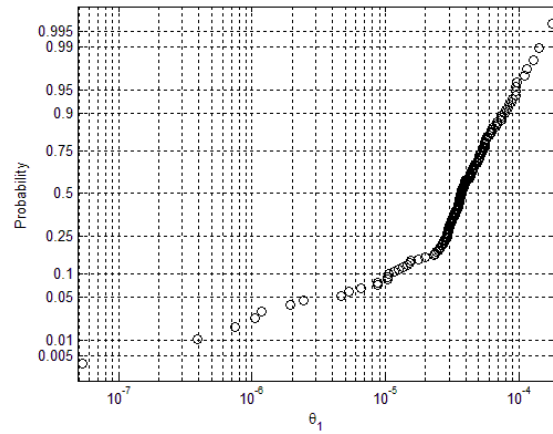
- K.Ortiza, A. (1988). Stochastic modeling of fatigue crack growth. *Engineering Fracture Mechanics* , 29 (3), 317-334.
- M. Carboni, S. B. (2007). Effect of probability of detection upon the definition of inspection intervals for railway axles. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit* , 221 (3), 409-417.
- N. Gebraeel, J. P. (2008). Prognostic Degradation Models for Computing and Updating Residual Life Distributions in a Time-Varying Environment. *IEEE Transaction on Reliability* , 57 (4), 539-549.
- N. Gebraeel, M. L. (2005). Life distributions from component degradation signals: A Bayesian approach. *IIE Trans.* , 37 (6), 543–557.
- N.Z Gebraeel, K. K. (2009). Predictive Maintenance Management Using Sensor-Based Degradation Models. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans* , 39 (4), 840-849.
- R.A. Smith, S. (2004). A brief historical overview of the fatigue of railway axles. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit* , 218 (4), 267-277.
- S. Beretta, M. C. (2004). Application of fatigue crack growth algorithms to railway axles and comparison of two steel grades. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit* , 218 (4).
- S. Beretta, M. C. (2006). Experiments and stochastic model for propagation lifetime of railway axles. *Engineering Fracture Mechanics* , 73, 2627–2641.
- S.Beretta, M. M. (2006). SIF solutions for cracks at notches under rotating bending. *Proceedings of the 16th European Conference on Fracture (ECF16)*. Alexandroupoulos.
- S.Beretta, M. (2005). Rotating vs. plane bending for crack growth in railway axles. *ESIS-TC24 Meeting*. Geesthacht.
- S.Beretta, M. (2005). Simulation of fatigue crack propagation in railway axles. *J ASTM Int* , 2 (5), 1-14.
- Schijve, J. (2001). *Fatigue of structures and materials*. Dordrecht: Kluwer Academic Publishers.
- U. Zerbst, K. M. (2005). Fracture mechanics in railway applications—an overview. *Engineering Fracture Mechanics* , 72, 163–194.
- U. Zerbst, M. V. (2005). The development of a damage tolerance concept for railway components and its demonstration for a railway axle. *Engineering Fracture Mechanics* , 72, 209–239.

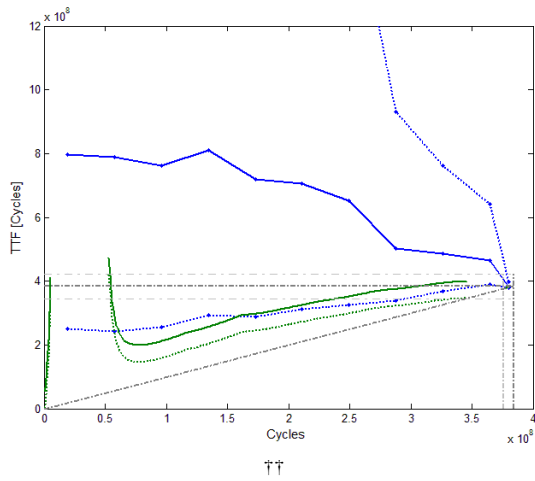
APPENDIX

In this paragraph graphs related to the first simulated crack growth path. They represent respectively:

- The predictions (lower bound, median and upper bound) on the TTF for
 - the prognostic physical model (blu lines)
 - the bayesian model (green lines)
- The probability of detection at each inspection
- The effect of the updating interval in km and the size error on the % of life exploited (physical model only)

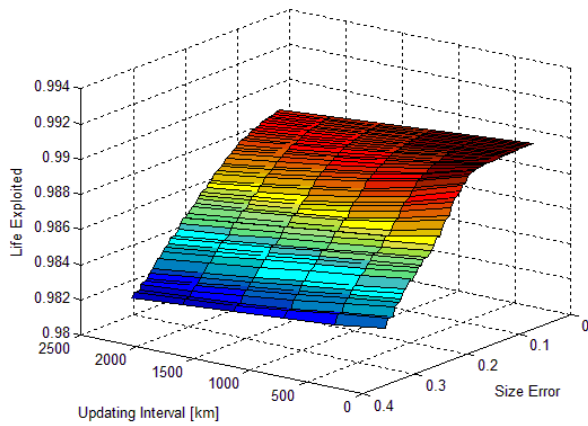
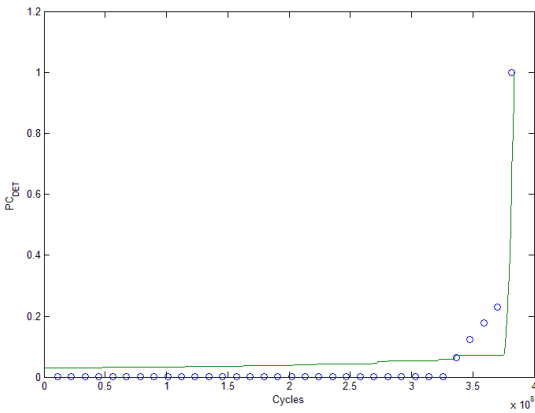
The first four probability plots represent the coefficients of the two exponential models used in the bayesian prognostic model.





DATA

$\Delta K_{th} = N(11.32, 0.857) \text{ MPa}\sqrt{\text{m}}$ $n = 1.9966$
 $\Delta K_{th0} = 5.96 \text{ MPa}\sqrt{\text{m}}$ $C_{th} = -0.02$
 $R = -1$ $\alpha_1 = -194.024$
 $\Delta K_{crit} = 24 \text{ MPa}\sqrt{\text{m}}$ $\alpha_2 = 322.544$
 $p = 1.3$ $\alpha_3 = -177.24$
 $q = 0.001$ $\alpha_4 = 41.957$
 $\alpha_5 = -1.916$ $D = 160 \text{ mm}$
 $\alpha_6 = -0.3927$ $K_t = 1.2$
 $\beta = 0.656$
 $\varepsilon = 10 \text{ MPa}$
 $\vartheta = 2.5$
 $S_0 = 0.2$



$\dagger\dagger$ Blue line: Physical model TTF estimation with confidence bounds (dotted)
 Green Line: Bayesian model TTF estimations with lower confidence bound (dotted)

Lithium-ion Battery State of Health Estimation Using Ah-V Characterization

Daniel Le¹ and Xidong Tang²

¹*National Science Foundation & American Society for Engineering Education
with Corporate Partnership from General Motors, Warren, MI 48090, USA
dbl2t@virginia.edu*

²*General Motors, Warren, MI 48090, USA
xidong.tang@gm.com*

ABSTRACT

The battery state of health (SOH) is a measure of the battery's ability to store and deliver electrical energy. Typical SOH methods characterize either the battery power or energy. In this paper, new SOH estimation methods are investigated based on the battery energy represented by the Ampere-hour throughput (Ah). The methods utilize characteristics of the Ah to estimate the battery capacity or the useable energy for state of health estimation. Three new methods are presented and compared. The simulation results indicate the effectiveness of the methods for state of health estimation.

1. INTRODUCTION

Battery diagnostic and prognostic methods are important to maintain proper battery operation. Battery damage occurs due to a number of reasons, such as over-charging and over-depleting the battery. Also, battery operation is dynamic and its performance varies significantly with age. An important aspect of battery diagnostics is the battery state of health (SOH) which is a qualitative measure of the battery's ability to store energy and deliver power. Battery diagnostics track the degradation of battery's performance to estimate battery SOH. There are two common methods to calculate the battery SOH. One method uses the battery impedance, or equivalently the battery power, to determine the battery SOH. The SOH using the impedance, R , can be calculated using Eq. (1).

$$SOH = \left(\frac{R_i}{R_0} \right) * 100 \quad [\%] \quad (1)$$

where R_i is the i^{th} impedance measurement in time and R_0 is the initial value. In the other method, the battery capacity, C , is used to determine the battery SOH as given in Eq. (2).

$$SOH = \left(\frac{C_i}{C_0} \right) * 100 \quad [\%] \quad (2)$$

where C_i is the i^{th} capacitance measurement in time and C_0 is the initial value. There are many studies that have researched the degradation of the battery as it ages (Zhang, 2011). As the battery ages, the battery's performance degradation is related to changes in the battery chemistry. First, the growth of a solid electrolyte interface (SEI) layer reduces the electrical efficiency of the battery. This contributes to an increase of the high-frequency resistance of the battery, reducing the maximum power output of the battery (Troltzsch, 2006). Considerable loss of battery power will result in ineffective vehicle operation or vehicle failure, i.e. vehicle inoperation. Second, the battery capacity degrades as the battery ages (Liaw, 2005). Capacity degradation results from several factors, such as loss of bonding sites in the active material and loss of active Lithium-ions. Considerable loss of battery capacity will result in ineffective battery operation and reduced vehicle range.

There have been several attempts to estimate the battery SOH using the battery impedance or the battery capacity. Haifeng et al (2009) defined SOH as a function of the battery's high-frequency resistance. Using a Kalman Filter, the authors estimated the battery resistance to estimate the battery SOH. Also, Kim (2010) developed a technique to estimate the battery capacity for SOH estimation. The author implements a dual-sliding mode observer to estimate battery capacity fade.

Although there has been much progress in the area of SOH estimation, it is still uncertain and still requires research to develop new and more accurate methods. The research presented in this paper investigates new methods which are based on the battery energy storage capability to estimate the battery SOH. The Ampere-hour throughput (Ah) is the current throughput by the battery and represents the energy that is delivered or stored by the battery. The battery terminal voltage and open-circuit voltage varies with the battery state of charge. The Ampere-hour throughput can be

related as a function of the battery terminal or open-circuit voltage, i.e. Ah-V. The methods presented in this paper capitalize on unique characteristics of the Ah-V function as the battery ages to estimate the battery SOH.

2. PROBLEM FORMULATION

As stated above, there are two main methods used to estimate the battery state of health (SOH). One method is based on the battery impedance and the other based on the battery capacity. For this paper, the battery capacity is used as the baseline method for SOH calculation. The battery capacity is especially important to electric vehicles (EV) and plug-in hybrid electric vehicles (PHEV) due to the range constraint of the battery. In this section, the problem of SOH estimation will be discussed and the basis for a practical method for online SOH estimation.

The battery capacity degrades over the life of the battery and varies with temperature. As the battery ages, irreversible losses reduce the amount of energy that can be stored and delivered. Also, over-charging and over-depleting the battery also cause the battery capacity to be reduced further. Determining the battery's state of health (SOH) provides a qualitative measure of its ability to function properly. The battery SOH can be calculated based on capacity measurements from the capacity test shown in Figure 1. The capacity test cycles the battery through constant current charge and discharge profiles. The battery is initially discharged to achieve 0% SOC, i.e. the terminal voltage is 2.5 V. The battery is charged to 100% SOC, i.e. the terminal voltage is 4.15V. These values are defined by the battery manufacturer. The

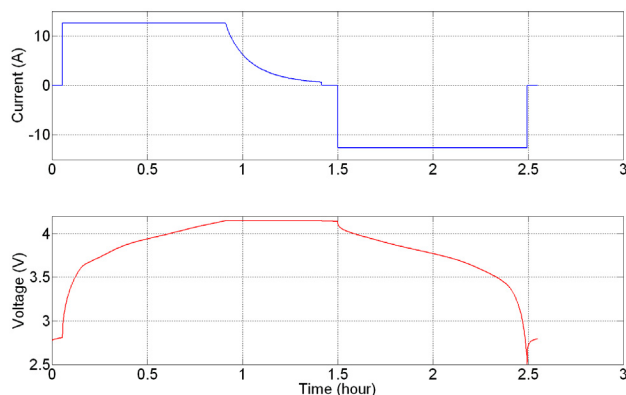


Figure 1. Measured terminal battery current and voltage during capacity test

As the battery ages, battery capacity slowly degrades as shown in Figure 2. The figure shows the measured capacity for three battery data sets. The battery for each data set was subject to capacity tests, performance tests, and accelerated ageing. The capacity test, shown in Figure 1, for the battery in each data set was conducted at 25 °C and was repeated to obtain an average capacity value. The performance tests included Charge Rate, Hybrid Pulse

Power Characterization (HPPC), Charge Depleting (EV mode), and Charge Sustaining battery current profiles. During the performance tests, the temperature of the battery for each data set, Data Set 1, Data Set 2, and Data Set 3, was maintained at 20, 30, and 40 °C, respectively. Each battery then underwent accelerated aging at 35 °C. The battery voltage rails, i.e. lower and upper operating voltage limits, were set at 2.5 and 4.15 V, which spans the nonlinear battery operating range. The terminal current and voltage were measured during all tests.

The battery SOH is calculated using the measured battery capacity, as given in Eq. (2) where C_i is the measured capacity of the i^{th} ageing iteration. The battery SOH over time is shown in Figure 3. The SOH is an indication of the battery health over the age of the battery. As the battery continues to age, the capacity will degrade further. At some point, the SOH will indicate that the battery is unhealthy, meaning the battery is unable to store and deliver energy for proper vehicle operation. In an ideal scenario, the battery capacity would be readily available to provide an accurate estimation of the battery SOH. However, in practical vehicle operation, this is not the case. In this context, new methods must be developed which can accurately estimate the battery SOH using online algorithms with the available battery data during vehicle operation.

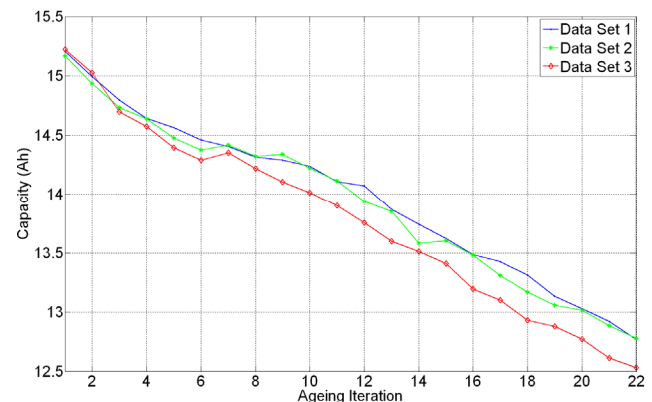


Figure 2. Battery capacity over time.

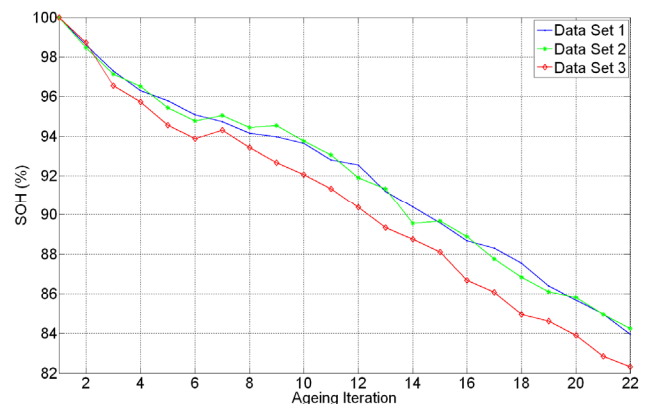


Figure 3. Battery state of health over time using battery capacity value

The battery capacity is the Ampere-hour throughput between the voltage rails of 2.5 and 4.15 V and is measured using the capacity test. However, during vehicle battery operation, the voltage rails are restricted to a smaller voltage range to maintain linear operating behavior and to protect the battery from damage due to over-charging and over-depletion. The relationship between the voltage rails and battery operation is illustrated in Figure 4. The Ampere-hour throughput between the restricted voltage rails is the useable energy during vehicle operation. A distinction is made between the terms “battery capacity” and “useable energy”. The “battery capacity” is the total Ampere-hour throughput between the voltage rails of 2.5-4.15 V. The “useable energy” is the Ampere-hour throughput between the restricted voltage rails of 3.4-4 V.

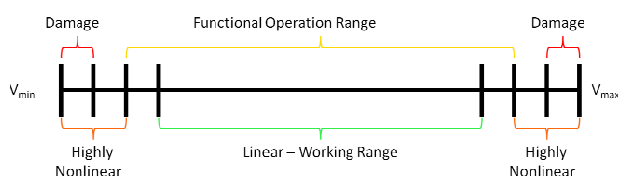


Figure 4. Illustration of the relationship between battery voltage rails and operating behavior

During vehicle battery operation, the voltage rails are restricted and the battery capacity cannot be measured. This inhibits the ability to calculate the battery SOH using the battery capacity. However, online methods can be developed which can provide accurate estimation of the battery SOH using characteristics of the relationship between the Ampere-hour throughput and voltage, i.e. Ah-V, which varies with the age of the battery.

In this study, constant current charge and discharge profiles are used to generate the Ah-V function, using the terminal voltage and open-circuit voltage. Although online constant current profiles will be limited, this study will illustrate that the Ah-V profiles can reflect battery ageing. In particular, the Ah-V profile using open-circuit voltage is not subject to battery loads and may be used to provide Ah-V profiles which can be used to estimate the battery SOH.

Onboard sensors measure the battery terminal voltage and current. The Ampere-hour throughput is the integrated current over time and represents the energy delivered or stored by the battery. The Ah can then be related as a function of the battery terminal voltage or the open-circuit voltage. Although the open-circuit voltage cannot be measured online, several studies have shown that the open-circuit voltage can be accurately estimated using filtering techniques.

In Figure 5, Ah is shown as a function of voltage as the battery ages, i.e. the test iteration number (Itr). For each iteration, the battery was cycled through a capacity test, which comprised of a constant current charge and discharge cycle. The terminal voltage was measured and the open-circuit voltage was estimated offline. The figure shows that

the Ah-V function, based on the measured terminal voltage or the estimated open-circuit voltage, gradually changes as the battery is aged. Similar results are seen for Data Set 2 and Data Set 3 but are not shown here. The upper most profiles are generated from constant current discharge. The lower most profiles are generated from constant current charge cycles. The middle profiles are Ah-V profiles generated using measured open-circuit voltage values. Several methods, presented in the next section, are developed which characterize the variations in the Ah-V function to estimate the battery SOH.

The Ah-V profile will be investigated to develop practical methods for SOH estimation. The Ah-V profiles are readily available through onboard sensors and algorithms. Although the current will fluctuate during vehicle operation, constant current charging operation may be available during vehicle battery recharging in EV and PHEV. Also, it is possible to implement an onboard filter to generate Ah-V discharge profiles. The Ah-V using the open-circuit voltage, however, is readily available given onboard estimation algorithms.

The following section will present several methods which utilize the characteristics of the Ah-V profile to estimate the battery SOH. The results will be compared to the battery SOH calculated using the measured battery capacity.

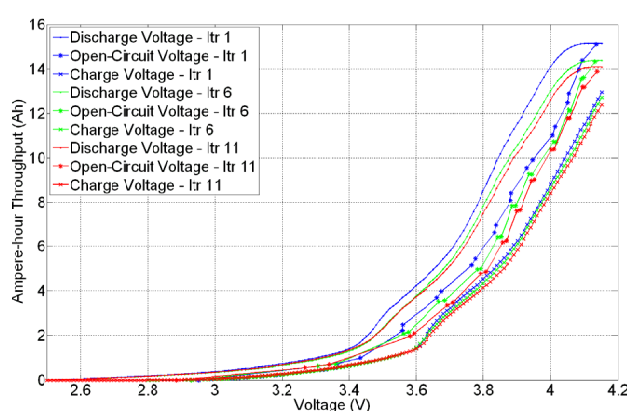


Figure 5. Ah as a function of terminal and open-circuit voltage during constant current charge and discharge cycles for Data Set 1.

3. METHODS

As the battery ages, variations in the battery Ah-V profile can be seen. Several new methods are developed which characterize the variations in the Ah-V profile to estimate the battery state of health (SOH). The methods estimate the battery capacity or useable energy to estimate the battery SOH. Also, the methods presented in this paper can be applied to online vehicle operation. New Ah-V data can be continually input and update the SOH estimation.

3.1 Non-linear Model

As seen in Figure 5, the Ampere-hour throughput is a nonlinear function of the battery terminal and open-circuit voltage between the voltage rails of 2.5 to 4.15 V. The Ah-V function can be modeled using a logistics growth curve, i.e. Richard’s curve (Richards, 1959), as given in Eq. (3).

$$F(x) = A + \frac{C}{\{1 + Sexp[-\beta(x - x_o)]\}^{1/S}} \quad (3)$$

where A = Lower Limit Value, C = Upper Limit Value, S = Symmetry, β = Growth Rate, x_o = Inflection Point.

This equation can be used to model the Ah-V function. The lower limit value, A, is set to zero, assuming that the Ah value is 0 when the battery is completely discharged, i.e. when the measured terminal voltage is 2.5V. The upper limit, C, represents the battery’s maximum energy storage potential, i.e. its capacity. Online battery data can be used to fit the model to determine parameter values. New Ah-V data can be used to update the model parameters.

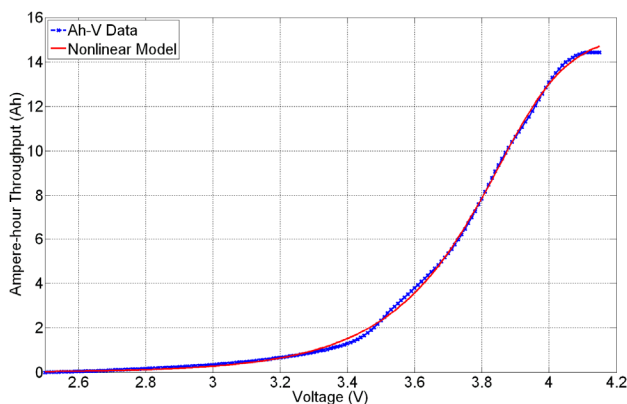


Figure 6. Nonlinear model fit using Ah-V data

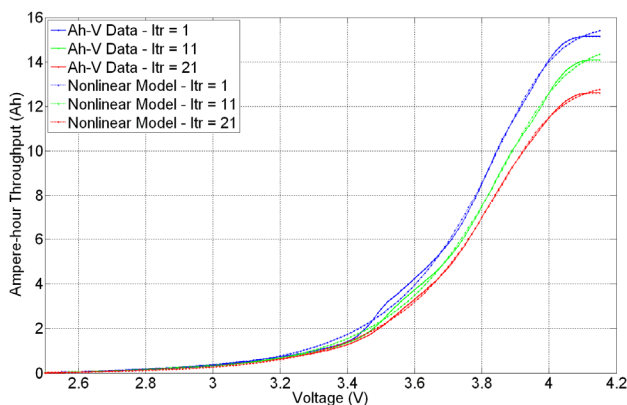


Figure 7. Ah-V data and fitted nonlinear model at different ageing iterations for Data Set 1

The Ah-V profile generate using constant current discharge data has a shape very similar to the logistics curve. Therefore, the Ah-V profile using the terminal voltage during constant current discharge is used to fit the logistics curve using least squares method, shown in . As seen in this

figure, the model fits the data relatively well. In addition, the maximum Ampere-hour throughput from the nonlinear model matches the battery capacity well.

For each ageing iteration, the constant current discharge data is used to generate the Ah-V profile. The Ah-V data is then used to fit the model parameters. shows the Ah-V data and fitted nonlinear model at different ageing iterations. As expected, the fitted nonlinear model is matches the relative shape. In addition, the maximum Ampere-hour throughput is approximately equal to measured battery capacity.

Using the nonlinear model, the estimated battery capacity is defined as the Ampere-hour throughput between the voltage rails of 2.5 and 4.15 V. The estimated battery capacity for each data set over the ageing iteration is shown in .

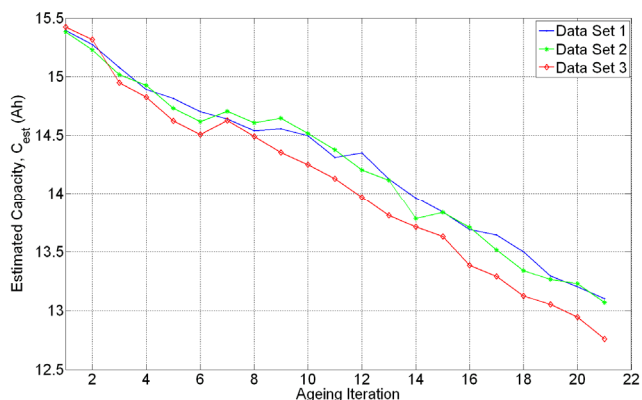


Figure 8. Estimated battery capacity over time

Using the estimated capacity, the estimated battery SOH is calculated using Eq. (4).

$$SOH_{est} = \left(\frac{(C_{est})_i}{(C_{est})_1} \right) * 100 \quad (4)$$

where the subscripts 1 and i indicate the test iteration number. The estimated SOH, SOH_{est} , is compared to the battery SOH in for each data set. In the figure, the label “Battery Data” refers to the SOH calculated using the measured battery capacity. The label “Nonlinear Model” refers to the SOH estimated using the estimated capacity from the nonlinear model. The estimated battery capacity using the fitted nonlinear model provides relatively accurate estimates for the battery SOH.

This method observes the battery behavior over the nonlinear operating region between the voltage rails of 2.5 and 4 V and uses the Ah-V battery data to fit the model parameters. Once the parameters are determined, the estimated capacity can be calculated. However, this method requires the battery to function between the voltage rails of 2.5 and 4.15 V to capture the nonlinear behavior. Ah-V Slope vs Battery Age

As shown above, the battery voltage rails define the working range of battery. In the capacity tests, the voltage rails were defined using manufacture specifications to measure the battery capacity. The Ah-V profile using the

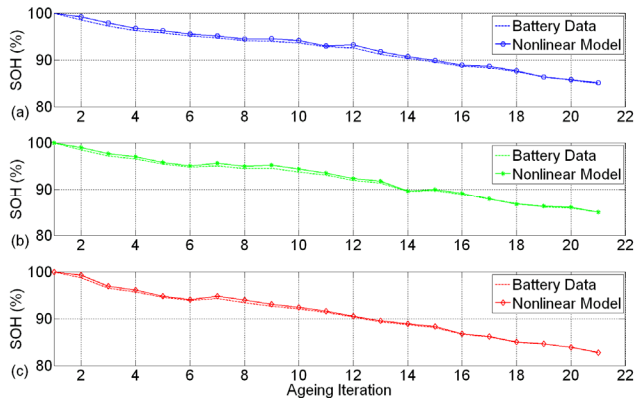


Figure 9. Compare battery SOH to estimated SOH for (a) Data Set 1, (b) Data Set 2, and (c) Data Set 3

voltage rails of 2.5V and 4.15V has a nonlinear profile as seen in Figure 5. However, in electric vehicles, battery voltage rails are restricted to maintain linear operating behavior, i.e. 3.4V and 4V. The Ah-V profiles between the restricted voltage rails of 3.4 to 4 V over the age of the battery are shown in . The figure presents three sets of Ah-V profiles at three ageing iterations. The upper three are the Ah-V generated using the terminal voltage during constant current discharge data. The middle three Ah-V profiles use the open-circuit voltage. The lower three are the Ah-V profiles generated using the terminal voltage during constant current charge data.

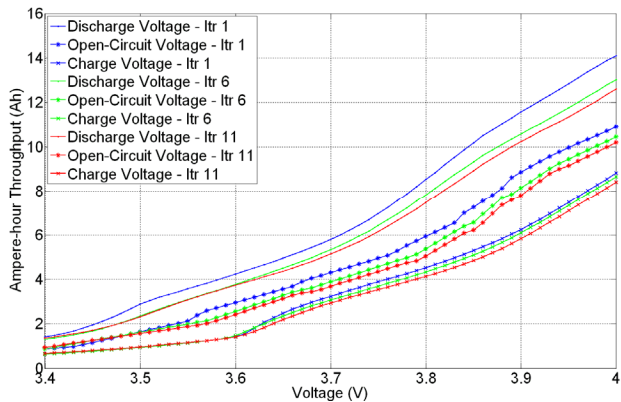


Figure 10. Ah as a function of terminal and open-circuit voltage for constant current charge and discharge over linear battery operation region for Data Set 1

The Ah-V profiles for the terminal and open-circuit voltages are relatively linear and vary with the age of the battery. Specifically, it can be seen that the slope of the Ah-V profiles vary with the battery age. The Ah-V data for the discharge, charge, and open-circuit Ah-V profiles were fitted to a linear model to estimate the Ah-V slope. shows an example of the linear fit of the Ah-V data. The slope, i.e. dAh/dV , of the Ah-V profiles were calculated for each ageing iteration and is shown in . The results show that the slope of the linear fit, for the discharge, charge, and open-circuit voltage Ah-V profiles, is approximately linear over

the age of the battery. The linear relationship between the slope of the linear fit to the battery age could be used to estimate the battery SOH from Eq. (4).

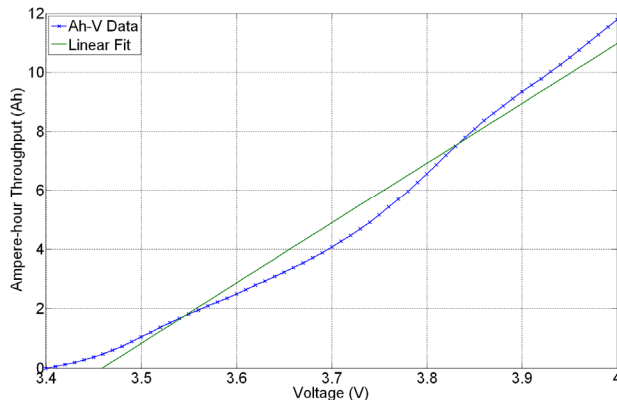


Figure 11. Example of linear fit to Ah-V Data

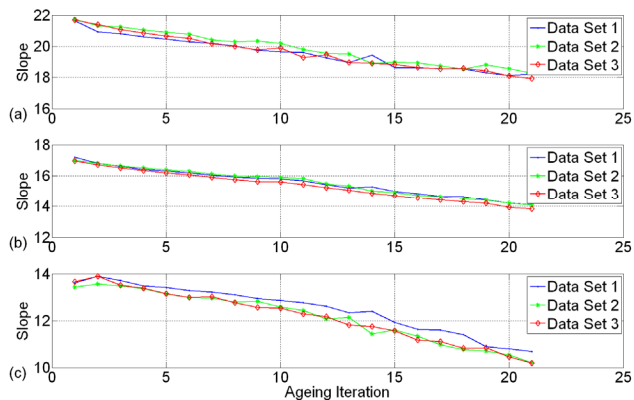


Figure 12. Slope of (a) discharge, (b) open-circuit, and (c) charge Ah-V profile using linear fit over ageing iteration

The slope of the linear fit was related to the battery’s measured capacity, shown in Figure 13. The results show that the battery capacity is a linear function of the Ah-V slope. A linear model can be generated to relate the capacity to the slope of the Ah-V function. In this way, online battery data can be used to generate the Ah-V profile and a linear fit can be used to calculate its slope. The slope can then be used to calculate the estimated battery capacity and then estimate the battery SOH.

This method does have some drawbacks. The capacity-slope relationship does vary with temperature and with current rate. However, variations are minimized if the open-circuit voltage Ah-V profile is used. This method for capacity estimation is also sensitive to small errors in the slope. Noise and uncertainty in the Ah-V profile will affect the linear fit and will produce inaccurate slope estimation, which will then affect the SOH estimation.

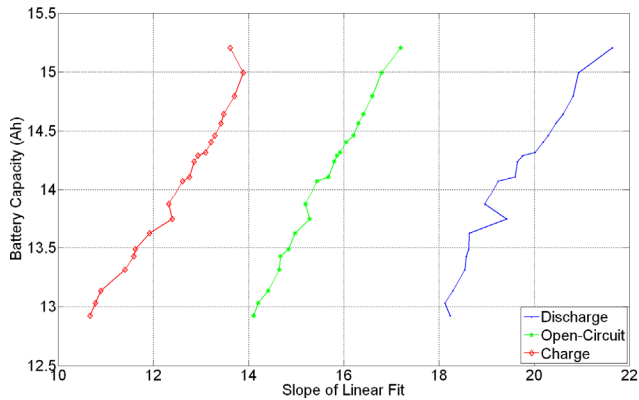


Figure 13. Battery capacity as a function of Ah-V slope over linear battery operating region for Data Set 1

3.2 Estimated Useable Energy Using Quadratic Fit

In the previous section, a linear fit was used to model the Ah-V profile which can then be used to estimate the battery SOH. The major limitation of the previous method is the sensitivity to error, which is largely due to the inaccuracy of the linear fit. The Ah-V data between the voltage rails of 3.4 to 4 V is approximately linear, however, small nonlinearities in the Ah-V function over this region introduce some inaccuracies.

In this method, a quadratic fit is used to model the Ah-V profile. A quadratic model provides a more accurate relationship and will be more tolerant to small errors. Also, the quadratic model can be easily updated to reflect new Ah-V data. This quadratic model can be used to estimate the battery useable energy for SOH estimation. The useable energy is the Ampere-hour throughput between the restricted voltage rails.

The following steps was used to estimate the battery SOH for each ageing iteration.

Step 1: The battery capacity is measured from the capacity tests. The battery capacity of the first ageing iteration is defined as the reference capacity value.

Step 2: The Ah-V profile using the open-circuit voltage is generated between the restricted voltage rails of 3.4 to 4 V.

Step 3: A quadratic fit is generated using the Ah-V data. The quadratic fit is constrained to 0 Ah, i.e. zero useable energy, at the lower voltage rail of 3.4 V. Figure 14 shows an example of the Ah-V profile using the open-circuit voltage and the quadratic fit.

Step 4: The estimated useable energy from the quadratic model is used to estimate the battery SOH using Eq. (5).

$$SOH_{est} = \left(\frac{(Useable\ Energy)_i}{(Useable\ Energy)_1} \right) * 100 \quad (5)$$

where the subscripts 1 and i indicate the test iteration number. A quadratic fit is used to model the Ah-V profile using the battery open-circuit voltage. The Ah-V profile can

also be generated using the terminal voltages. However, the Ah-V using the terminal voltages will vary with operating conditions such as temperature and current rate. For clarity, only the Ah-V using the open-circuit voltage is shown. The quadratic model is constrained to 0 Ah, i.e. zero useable energy, at the lower voltage rail of 3.4 V. The Ah-V data, using the open-circuit voltage, and the quadratic fit are shown in Figure 14. The quadratic fit is more accurate than a linear fit.

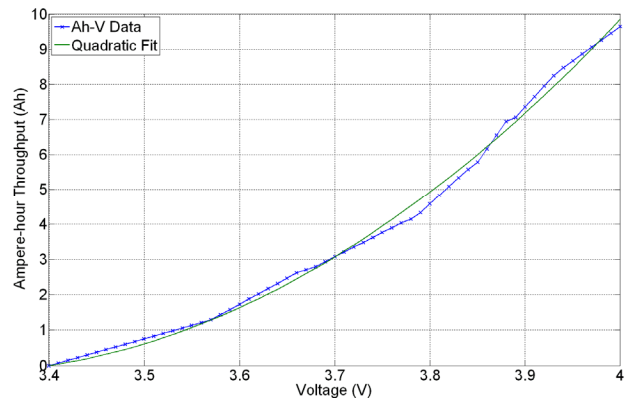


Figure 14. Example of quadratic fit of the Ah-V using the open-circuit voltage over linear operating region

The estimated useable energy is calculated using the quadratic model and is shown in over the battery ageing iteration.

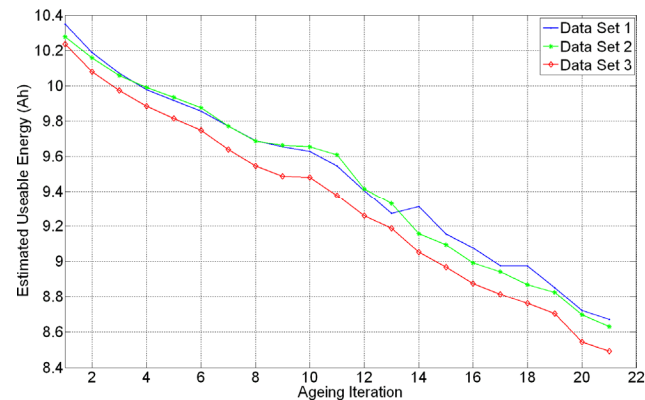


Figure 15. Estimated useable energy calculated from quadratic model of Ah-V

The battery SOH is calculated based on the useable energy determined from the data and quadratic fit of the Ah-V function between the restricted voltage rails of 3.4 to 4 V. The estimated SOH using the useable energy is compared the SOH calculated from battery capacity values are shown in Figure 16 for each battery data set. The SOH using the quadratic fit matches the SOH calculated using the numerical results well. Also, the figure includes the calculated SOH using the measured battery capacity. The figure shows that the SOH calculated based on the quadratic model also match well.

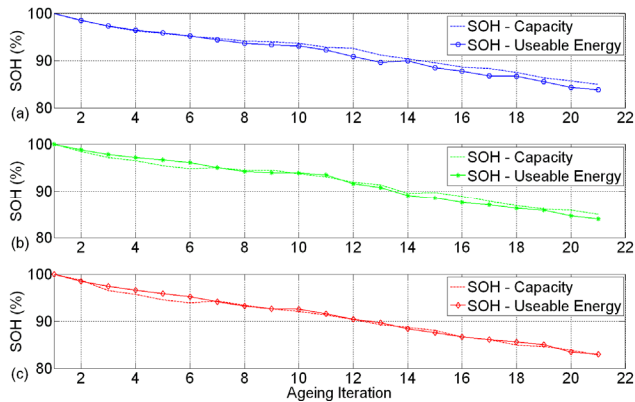


Figure 16. Battery SOH using battery capacity estimated useable energy for (a) Data Set 1, (b) Data Set 2, and (c) Data Set 3

The results from this method show that the battery SOH can be accurately estimated using a quadratic model of the Ah-V. The quadratic model is used to estimate the useable energy over the restricted voltage rails of 3.4 to 4 V which is then used estimate the battery SOH. The estimated SOH matches the battery SOH well. This method can be applied to battery vehicle operation.

4. CONCLUSIONS

Several new methods for capacity estimation were developed and investigated. Each method has a potential to provide capacity estimation for SOH evaluation. The first method models the linear and nonlinear regions of the Ah-V curve using Richard's equation. This method requires a high degree of training effort. The slope of the Ah-V curve was correlated to the battery capacity. This is a relatively simplistic method that provides a linear relationship between the slope and the battery capacity. This method is sensitive to small errors and requires complete charge and discharge cycles to maintain accuracy. The last method uses a quadratic fit to model the Ah-V function. Using the open-circuit voltage, a reliable estimation of the battery useable energy can be used to estimate the battery SOH. This results of this method match well to the SOH calculated using battery capacity values.

ACKNOWLEDGEMENT

This research work was supported by the National Science Foundation under Grant # EEC-0946373 to the American Society for Engineering Education. The authors would like to acknowledge Dr. Xiaofeng Mao for data collection and analysis.

REFERENCES

Haifeng, D., Xuezhe, W., Zechang, S. (2009). A new SOH prediction concept for the power lithium-ion battery used on HEVs. *Vehicle Power and Propulsion*

Conference (pp. 1649-1653), Sept. 7-11, Dearborn, MI. doi: 10.1109/VPPC.2009.5289654

- Kim, I. (2010). A Technique for Estimating the State of Health of Lithium Batteries Through a Dual-Sliding-Mode Observer. *IEEE Transactions on Power Electronics*. vol. 25 (4), pp. 1013-1022.
- Kim, H., Heo, S., Kang, G. (2010). Available Power and Energy Prediction Using a Simplified Circuit Model of HEV Li-ion Battery. *SAE 2010 World Congress & Exhibition*, April 13-15, Detroit, MI.
- Richards, F. (1959). A flexible growth function for empirical use. *Journal of Experimental Botany*. vol. 10, pp. 290-300.
- Liaw, B., Jungst R., Nagasubramanian, G., Case, H., Doughty, D. (2005). Modeling capacity fade in lithium-ion cells. *Journal of Power Sources*, vol. 140 (1), pp. 157-161.
- Rong, P. and Pedram, M. (2006). An analytical model for predicting the remaining battery capacity of lithium-ion batteries. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 14 (5), pp 441- 451.
- Spotnitz, R. (2003). Simulation of capacity fade in lithium-ion batteries. *Journal of Power Sources*, vol. 113, pp. 72-80.
- Zhang, Y. Wang, C., Tang, X. (2011). Cycling degradation of an automotive LiFeP04 lithium-ion battery. *Journal of Power Sources*. vol. 196 (3), pp. 1513-1520.
- Troltzsch U., Kanoun O., Trankler H. (2006). Characterizing aging effects of lithium ion batteries by impedance spectroscopy. *Electrochimica Acta*. vol. 5 (8,9), pp. 1664-1672.

Daniel B. Le received his B.S. from Arizona State University in Aerospace Engineering in 2002 and his M.S. and Ph.D. in Mechanical and Aerospace Engineering from the University of Virginia in 2005 and 2009, respectively. He is currently a Research Scientist at General Motors through a research fellowship through NSF and ASEE. His research is focused on developing diagnostic and prognostic methods for energy storage systems.

Xidong Tang received his B.S. degree and M.S. degree in Automatic Control from Shanghai Jiao Tong University, China in 1997 and 2000 respectively. In 2005 he received his Ph.D. degree in Electrical Engineering from the University of Virginia. Since 2005 he has worked as a senior researcher in the Electrical and Controls Integration Lab at GM R&D in Warren, MI. His research covers diagnosis, prognosis, fault tolerance, control theory and applications, estimation theory, signal processing, and pattern recognition. He has co-authored one book, published over 30 papers, and filed over 20 patent applications.

Model-Based Prognostics Under Non-stationary Operating Conditions

Matej Gašperin¹, Pavle Boškoski¹, Dani Juričić¹

¹ *Jožef Stefan Institute, Ljubljana, Slovenia*

matej.gasperin@ijs.si, pavle.boskoski@ijs.si, dani.juricic@ijs.si

ABSTRACT

The paper presents a novel approach for prognostics of faults in mechanical drives under non-stationary operating conditions. The feature time series is modeled as an output of a dynamical state-space model, where operating conditions are treated as known model inputs. An algorithm for on-line model estimation is adopted to find the optimal model at the current state of failure. This model is then used to determine the presence of the fault and predict the future behavior and remaining useful life of the system. The approach is validated using the experimental data on a single stage gearbox.

1. INTRODUCTION

An important emerging feature of new generation of condition monitoring systems enables prediction of future evolution of the fault and thus enables the plant personnel to accommodate maintenance actions well in advance. Even more, it can predict the remaining useful life of the component under changing operating condition, thus providing information to operators on how the different operating regimes will lengthen or shorten the components useful life. This is a relatively new research area and has yet to receive its prominence compared to other condition monitoring problems (Heng, Zhang, Tan, & Mathew, 2009).

The focus in this paper will be on mechanical drives. They are the most ubiquitous item of equipment in manufacturing and process industries as well as transportation. During the operational life-cycle, these items are subjected to wear, fatigue, cracks and other destructive processes. These processes can not be directly observed or measured without interrupting the operation of the machine. The extent of the damage has to be

This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

inferred from the available signals, which are usually vibrations, acoustic emissions, oil contaminants, etc.

In this work, we follow an established approach for model-based prognostics, which is to model the fault progression using a dynamical model. This approach has been applied to specific cases where the exact model of the fault was derived. The model, combined with an appropriate state estimation algorithm (e.g. Particle Filter) can be used to estimate the current state and predict its future evolution (M. Orchard, Kacprzyński, Goebel, Saha, & Vachtsevanos, 2008; M. E. Orchard & Vachtsevanos, 2009; Zhang et al., 2009; DeCastro, Liang, Kenneth, Goebel, & Vachtsevanos, 2009). However, most of the authors assume constant operating conditions of the machine. Recently, (Edwards, Orchard, Tiang, Goebel, & Vachtsevanos, 2010) analyzed the impact of variable operating conditions on the remaining useful life in terms of uncertainty.

The aim of this work is to propose a new approach toward model-based prognostics in which the operating conditions are considered as a measured input into the model. Because the exact relations between the model inputs, fault dimension and measured signals are hard to derive, we propose an algorithm for on-line estimation of these relations. The model obtained in this manner can therefore be used to determine the current state and trend of the fault, predict its future evolution in different operating regimes and estimate its remaining useful life (RUL).

The paper is organized as follows. Section 2 presents the conceptual idea behind the proposed approach for a general setup. Section 3 introduces the algorithm for model estimation that can be used to apply the proposed approach. Section 4 presents the experimental setup that was used to collect the data for algorithm validation. Section 5 shows the results in terms of estimating the current state and trend of the fault and predict its future evolution. Finally, Section 6 summarizes the most important results and outlines the directions for further

research.

2. THE IDEA OF THE PROPOSED APPROACH

Let us assume that there exists at least one feature that provides the information about the current extent of the fault in a mechanical system and its value is available through noisy measurements. Furthermore, different operating conditions affect the extent and the rate of change of the underlying fault as well as the current feature value. Finally, when the fault occurs, its progression can be described by a stochastic dynamical process (Gašperin, Juričić, Bošković, & Vižintin, 2011).

Following the above assumptions, the evolution of fault dimension in time can be described by the following model (M. E. Orchard & Vachtsevanos, 2009):

$$\mathbf{x}_{t+1} = \mathbf{f}(\mathbf{x}_t, \mathbf{u}_t, \boldsymbol{\theta}) + \mathbf{w}_t \quad (1a)$$

$$\mathbf{y}_t = \mathbf{g}(\mathbf{x}_t, \mathbf{u}_t, \boldsymbol{\theta}) + \mathbf{v}_t \quad (1b)$$

where \mathbf{x}_t is the system state, \mathbf{y}_t is the observed feature value, \mathbf{u}_t is the vector of model inputs, $\boldsymbol{\theta}$ is the vector of model parameters, finally \mathbf{w}_t and \mathbf{v}_t are random variables describing system and measurement noise, respectively. The first equation in the model represents the fault evolution dynamics and the second one describes the feature extraction. Assuming that the values of the model parameters $\boldsymbol{\theta}$ are known, this model can be used to predict the future evolution of the fault for any given sequence of the operating conditions (fixed or variable) \mathbf{u}_t .

Nonlinear models (1) are a very powerful description of the process dynamics and can describe a broad range of dynamic behavior. Usually the estimation methods include only a specific family of models, e.g. as shown by (DeCastro et al., 2009) or rely on approximation methods (M. Orchard et al., 2008). If linearized, the expression (1) takes the form (Gašperin et al., 2011)

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \mathbf{w}_t \quad (2a)$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \mathbf{D}\mathbf{u}_t + \mathbf{v}_t \quad (2b)$$

In the model (2), \mathbf{w}_t and \mathbf{v}_t are random variables that follow a normal distribution:

$$\begin{bmatrix} \mathbf{w}_t \\ \mathbf{v}_t \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{Q} & \mathbf{S} \\ \mathbf{S}^T & \mathbf{R} \end{bmatrix} \right) \quad (3)$$

If the functions governing the dynamical behavior of the fault in (1) are known, the linear approximation can be computed analytically. However, this has only been done for a limited number of special cases and for a general setup, the model parameters have to be assumed unknown. To alleviate this problem we propose a data-driven approach for modeling and prognostics, where the parameters of the linear model (2) are estimated on-line based on the past data of the feature value.

The benefit of using a linear model is that the parameter estimation algorithm can be implemented with minimal computational load and the analysis of the model (in terms of stability)

is less demanding than in the nonlinear case. The downside is that linear model can only adequately describe the system in a limited subspace of fault dimension and operating conditions. However, this is partially alleviated by on-line parameter estimation that provides an updated model as soon as the conditions change.

2.1 Prognostics under variable operating conditions

It is well known (Heng et al., 2009) that the changes in operating conditions (e.g., load, temperature) can greatly affect the fault in mechanical systems. A schematic representation of different scenarios is given in Figure 1, where it can be seen that under more favorable load, the life of the machine can be significantly extended.

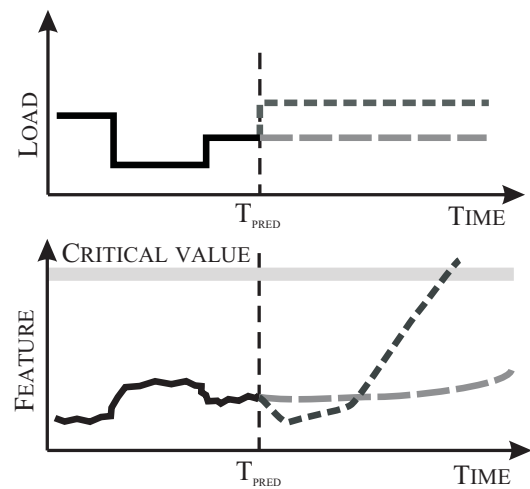


Figure 1. Fault progression under different load scenarios

The exact relations between them and the fault dimension can be obtained by advanced and complex modeling approaches, which are usually not applicable to real-world condition monitoring problems. The main advantage of implementing the approach presented here is that it offers a systematic solution to finding the relation between the machine operating conditions, feature value and fault dimension. The added functionality of our solution can be summarized as follows:

- **Detection of fault progression:** The approach can separate the fault evolution dynamics from the dynamics enforced by the variable operating conditions. This means that we can detect the rate at which the fault is progressing.
- **Estimation of the remaining useful life:** If the future load profile of the machine is known, it can be used as an input to the model and predict the future evolution of the fault.

3. MODEL ESTIMATION

In this chapter we will address the problem of estimating unknown model parameters of the linear state-space models (2). Estimating the state-space models is challenging because the internal system states are not directly observed and therefore all the information about them has to be inferred from the measured data. The state sequence can be estimated from the data, but the procedure requires the knowledge of the model parameters. As this is usually not the case, an approach that allows both the estimation of system states and unknown model parameters is required.

3.1 Maximum likelihood estimator

Suppose \mathbf{x} is a random variable with probability density function $p(\mathbf{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of unknown parameters. Let $\mathbf{X}_T = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$ be the set of observed values. The probability density function of \mathbf{X}_T is

$$p(\mathbf{X}_T|\boldsymbol{\theta}) = p(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T|\boldsymbol{\theta}) \quad (4)$$

The pdf $p(\mathbf{X}_T|\boldsymbol{\theta})$ is deterministic function of $\boldsymbol{\theta}$ and is referred to as the *likelihood function*. A reasonable estimator for $\boldsymbol{\theta}$ could then be to select the values in such a way that the observed realization \mathbf{X}_T becomes as likely as possible. Maximum Likelihood (ML) estimator for unknown parameters is defined by

$$\hat{\boldsymbol{\theta}}_{ML}(\mathbf{X}_T) = \arg \max_{\boldsymbol{\theta}} p(\mathbf{X}_T|\boldsymbol{\theta}) \quad (5)$$

where the maximization is performed with respect to $\boldsymbol{\theta}$ and for a fixed \mathbf{X}_T .

Rather than (5) it is often convenient to operate with the log-likelihood function.

$$L(\boldsymbol{\theta}) = \log p(\mathbf{X}_T|\boldsymbol{\theta}) \quad (6)$$

Since logarithmic function is monotonically increasing, maximizing the likelihood function is the same as maximizing its logarithm,

$$\hat{\boldsymbol{\theta}}_{ML}(\mathbf{X}_T) = \arg \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \quad (7)$$

3.2 Likelihood function for dynamical models

Consider a dynamic state-space model, where $\mathbf{Y}_T = \{y_1, y_2, \dots, y_T\}$ are the measured system outputs, $\mathbf{X}_T = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$ is the unobserved sequence of system states and $\boldsymbol{\theta}$ is vector of model parameters. A straightforward way to define the maximum likelihood parameter estimator for this case is

$$\hat{\boldsymbol{\theta}}_{ML}(\mathbf{Y}_T) = \arg \max_{\boldsymbol{\theta}} p(\mathbf{Y}_T|\boldsymbol{\theta}) \quad (8)$$

where the data likelihood function can be expressed using chain rule

$$p(\mathbf{Y}_T|\boldsymbol{\theta}) = p(y_1|\boldsymbol{\theta}) \prod_{t=2}^T p(y_t|\mathbf{Y}_{t-1}, \boldsymbol{\theta}) \quad (9)$$

However, it is convenient to consider the log-likelihood function

$$L(\boldsymbol{\theta}) = \log p(\mathbf{Y}_T|\boldsymbol{\theta}) = \sum_{t=2}^T \log p(y_t|\mathbf{Y}_{t-1}, \boldsymbol{\theta}) + \log p(y_1|\boldsymbol{\theta}) \quad (10)$$

And the maximum likelihood estimator is thus

$$\hat{\boldsymbol{\theta}}_{ML}(\mathbf{Y}_T) = \arg \max_{\boldsymbol{\theta}} p(\mathbf{Y}_T|\boldsymbol{\theta}) = \arg \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \quad (11)$$

A closer look at the expression $p(y_t|\mathbf{Y}_{t-1}, \boldsymbol{\theta})$ in (10) reveals that it depends on system states. Indeed

$$p(y_t|\mathbf{Y}_{t-1}, \boldsymbol{\theta}) = \int p(y_t|\mathbf{x}_t, \boldsymbol{\theta}) p(\mathbf{x}_t|\mathbf{Y}_{t-1}, \boldsymbol{\theta}) d\mathbf{x}_t \quad (12)$$

The formulation of the above integral is problematic and in general case no closed form solutions exist.

3.3 The Expectation-Maximization algorithm

The expectation-maximization algorithm can solve the ML estimation problem in the case of incomplete or missing data. Therefore, if the states \mathbf{X}_T are considered as missing data, this algorithm can be successfully deployed to solve the system identification problem. Consider an extension to (8).

$$\hat{\boldsymbol{\theta}}_{ML}(\mathbf{X}_T, \mathbf{Y}_T) = \arg \max_{\boldsymbol{\theta}} \log p(\mathbf{X}_T, \mathbf{Y}_T|\boldsymbol{\theta}) \quad (13)$$

The EM algorithm then solves the problem of simultaneously estimating system states and model parameters by alternating between two steps. First, it approximates the likelihood function with its expected value over the missing data (E-step), and secondly maximizes the likelihood function w.r.t. $\boldsymbol{\theta}$ (M-step). A short overview of the algorithm will be presented, while a more detailed explanation can be found in (Haykin, 2001; Gibson & Ninness, 2005).

1. Start with initial parameter estimate $\boldsymbol{\theta}_0$.
2. **Expectation (E) step:** Compute the expected value of the complete data log-likelihood function.

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_k) = E_{p(\mathbf{X}_T|\mathbf{Y}_T, \boldsymbol{\theta}_k)} \{\log p(\mathbf{X}_T, \mathbf{Y}_T|\boldsymbol{\theta})\} \quad (14)$$

3. **Maximization (M) step:** Compute the optimal parameter vector value by maximizing the function $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_k)$.

$$\boldsymbol{\theta}_{k+1} = \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_k) \quad (15)$$

4. If convergence criteria are not satisfied, set $k = k + 1$ and return to step 2.

According to the EM algorithm, the first task is to compute the expected value of the complete data log-likelihood function

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_k) = E_{p(\mathbf{X}_T|\mathbf{Y}_T, \boldsymbol{\theta}_k)} \{\log p(\mathbf{X}_T, \mathbf{Y}_T|\boldsymbol{\theta})\} \quad (16)$$

where the joint likelihood of the measured output and system states can be written as

$$\begin{aligned} p(\mathbf{Y}_T, \mathbf{X}_T | \boldsymbol{\theta}) &= p(y_1, \dots, y_T | x_1, \dots, x_T, \boldsymbol{\theta}) p(x_1, \dots, x_T | \boldsymbol{\theta}) \\ &= p(x_1 | \boldsymbol{\theta}) \prod_{t=1}^{T-1} p(x_{t+1} | x_t, \boldsymbol{\theta}) \prod_{t=1}^T p(y_t | x_t, \boldsymbol{\theta}) \end{aligned} \quad (17)$$

Taking into account Gaussian distributions and ignoring the constants, the complete data likelihood function can be written as

$$\begin{aligned} -2 \log p(\mathbf{X}_T, \mathbf{Y}_T | \boldsymbol{\theta}) &= \log |\mathbf{P}_1| + (\mathbf{x}_1 - \boldsymbol{\mu}_1)^T \mathbf{P}_1^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) \\ &+ \sum_{t=1}^T (\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t - \mathbf{B}\mathbf{u}_t)^T \mathbf{Q}^{-1} (\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t - \mathbf{B}\mathbf{u}_t) \\ &+ \sum_{t=1}^T (y_t - \mathbf{C}\mathbf{x}_t - \mathbf{D}\mathbf{u}_t)^T \mathbf{R}^{-1} (y_t - \mathbf{C}\mathbf{x}_t - \mathbf{D}\mathbf{u}_t) \\ &+ T \log |\mathbf{Q}| + T \log |\mathbf{R}| \end{aligned} \quad (18)$$

The expected value of the above expression can be maximized by the following choices (Gibson & Ninness, 2005):

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \boldsymbol{\Psi} \boldsymbol{\Sigma}^{-1} \quad (19)$$

$$\begin{bmatrix} \mathbf{Q} & \mathbf{S} \\ \mathbf{S}^T & \mathbf{R} \end{bmatrix} = \boldsymbol{\Phi} - \boldsymbol{\Psi} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Psi}^T \quad (20)$$

where

$$\boldsymbol{\Phi} = \frac{1}{T} \sum_{t=1}^N E_{p(\mathbf{X}_T | \mathbf{Y}_T, \boldsymbol{\theta}_k)} \left\{ \begin{bmatrix} \mathbf{x}_{t+1} \\ \mathbf{y}_t \end{bmatrix} \begin{bmatrix} \mathbf{x}_{t+1}^T & \mathbf{y}_t^T \end{bmatrix} \right\} \quad (21)$$

$$\boldsymbol{\Psi} = \frac{1}{T} \sum_{t=1}^N E_{p(\mathbf{X}_T | \mathbf{Y}_T, \boldsymbol{\theta}_k)} \left\{ \begin{bmatrix} \mathbf{x}_{t+1} \\ \mathbf{y}_t \end{bmatrix} \begin{bmatrix} \mathbf{x}_t^T & \mathbf{u}_t^T \end{bmatrix} \right\} \quad (22)$$

$$\boldsymbol{\Sigma} = \frac{1}{T} \sum_{t=1}^N E_{p(\mathbf{X}_T | \mathbf{Y}_T, \boldsymbol{\theta}_k)} \left\{ \begin{bmatrix} \mathbf{x}_t \\ \mathbf{u}_t \end{bmatrix} \begin{bmatrix} \mathbf{x}_t^T & \mathbf{u}_t^T \end{bmatrix} \right\} \quad (23)$$

and the required expected values of the system states can be computed using a standard Kalman smoother (Haykin, 2001).

The estimated values of model parameters at a time instance T , along with the estimated state sequence and the model structure defined by (2) constitute the model of the fault dynamics at this particular time instance and is labeled \mathcal{M}_T .

3.4 Algorithm Summary

The presented algorithm, adopted for machine health estimation and prognostics can be summarized as follows:

1. Select time window N and set $T = N + 1$.
2. Run the EM algorithm for model estimation using past data $y_{T-N}, y_{T-N-1}, \dots, y_T$ and $u_{T-N}, u_{T-N-1}, \dots, u_T$.

3. Use the estimated model \mathcal{M}_T and state \mathbf{x}_T to analyze the fault and predict future behavior of the system.
4. When the new feature value is collected, set $T = T + 1$ and return to step 2.

4. CASE STUDY

For the purpose of the development and verification of the model-based prognostics tools, the experimental test bed has been used (Figure 2). It consists of a motor-generator pair with a single stage gearbox. The motor is a standard DC motor powered through DC drive. A generator is being used as a break and the generated power is being fed back in the system, thus achieving the breaking force.

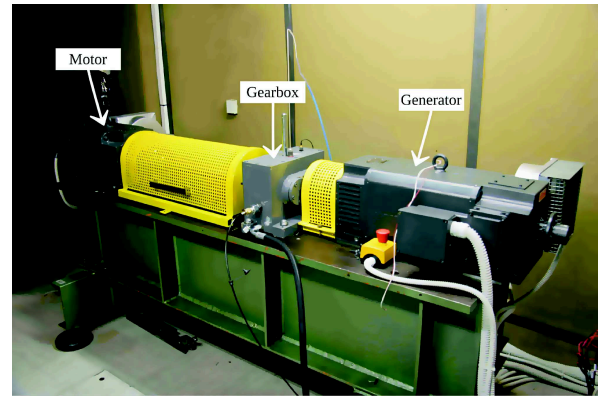


Figure 2. The test bed

The most informative and easily accessible signals that offer information on gear health are vibration signals (Combet & Gelman, 2009). In our setup, the vibration signals are acquired from a sensor placed on the output shaft bearing.

4.1 Experimental run

The set of gears was subjected to a time-varying load profile. The speed was kept constant throughout the experiment. Vibration signals were acquired every 5 minutes and each acquisition took 5 seconds.

The complete experiment lasted approximately 180 hours. At the end extensive pitting damage was clearly visible on both gear and pinion, as shown in Figure 3.

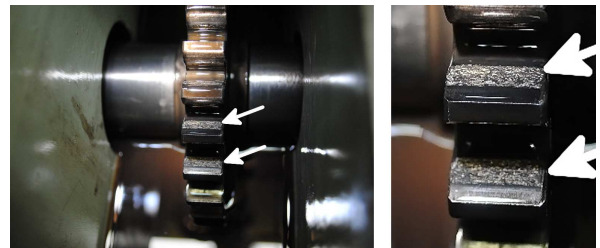


Figure 3. Gear condition after 180 hours of operation

4.2 Gear fault dynamics

The main source of vibrations in meshing gears originates from the changes in the bending stiffness of the gear teeth as well as variations in the torsional stiffness of the shafts and supporting bearings (Howard, Jia, & Wang, 2001). As gear teeth rotate through the meshing cycle the overall bending stiffness changes according to the number of teeth present in the meshing contact. Under constant operating conditions, these variations are expressed as a spectral component positioned at the gear mesh frequency.

A localized fault alters the original tooth stiffness profile. This alteration occurs every time the damaged tooth enters a meshing contact. This localized fault affects the produced vibrations by the appearance of an additional modulation component around the original gear mesh frequency (Randall, 1982). As the fault progresses and spreads on all teeth the changes in the gear mesh frequency component become more apparent.

As our goal is to perform the earliest possible estimation of the remaining useful life of the observed gears, we have based our algorithm on the information contained in the signal's energy portion extracted from the sidebands around the principle gear mesh component. This value was computed for each vibration acquisition session and the corresponding time series represents the feature values.

In terms of modeling the gear fault dynamics the feature value is the model output while the known inputs into the model are torque and temperature. The model inputs and outputs are shown in Figure 4.

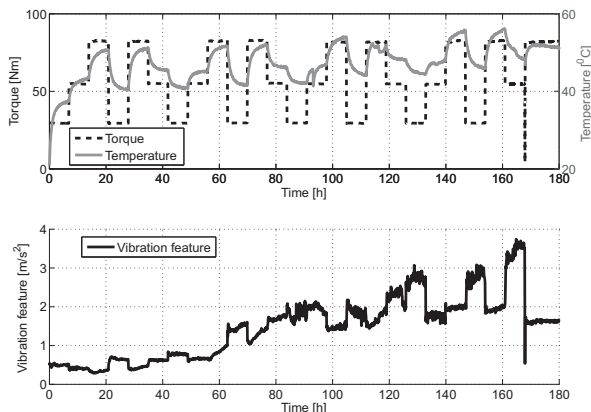


Figure 4. Top: torque and temperature (inputs), Bottom: vibration feature (output)

5. RESULTS

The developed algorithm for model estimation was implemented with the sample size of $N = 200$, which corresponds to approximately 16 hours. The unknown model parameters are:

$$\theta = [A, B, C, D, Q, R] \quad (24)$$

where Q and R are covariance matrices of Gaussian random variables w_t and v_t , respectively. The model structure is defined by selecting the number of hidden states, measured inputs and outputs. In our case, the state dimension is $m = 2$, the number of inputs is $n = 2$ (torque and temperature) and the model has $d = 1$ measured output (vibration feature). The unknown model parameters are thus matrices with the following dimensions:

$$\begin{aligned} A &\in \mathbb{R}^{m \times m}, B \in \mathbb{R}^{m \times n}, C \in \mathbb{R}^{d \times m}, \\ D &\in \mathbb{R}^{d \times n}, Q \in \mathbb{R}^{m \times m}, R \in \mathbb{R}^{d \times d} \end{aligned} \quad (25)$$

Prior to running the algorithm, these parameters have to be initialized to some values. In this problem formulation, the selection of the initial values is not crucial as the likelihood function for linear system is unimodal and there is no threat of divergence. The values of all the matrix entries were thus set to a neutral value of 0.1.

5.1 Detecting the trend of the fault

After a model \mathcal{M}_T is obtained at a certain time point T , it can be analyzed to determine the current trend of the fault, even under variable operating conditions in the period of data acquisition. This is made possible because the state-space model can distinguish between the feature dynamics that is due to the variable operating conditions (model input matrix B) and the dynamics due to the fault progression (system state matrix A). Therefore, by analyzing the eigenvalues of the system matrix A , one can determine whether the fault progression has a stable dynamics (i.e. it will remain of a constant size) or unstable dynamics (i.e. the fault dimension will increase in time).

A more illustrative way to present this is by visualizing the future evolution of the feature value at constant operating conditions. In Figure 5, this is done for two different times T_{pred} , one with stable and one with unstable dynamics.

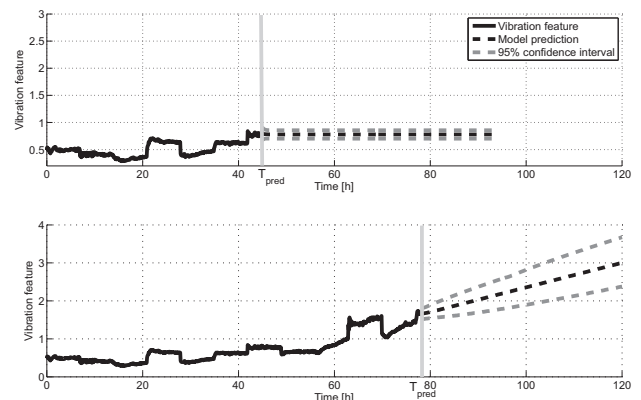


Figure 5. Detection of the fault progression at non-stationary operating conditions

It can be seen that in the first case (Figure 5 (top)), the predicted feature value is constant, which means that the fault

will not progress. The T_{pred} here was $44h$ and no fault was indeed present at that time. In the second case (Figure 5 (bottom)), the model was estimated at $T_{pred} = 78h$, where the fault started to increase and the model thus predicted the gradual increase of the feature value even at a constant load.

5.2 Model-based prognostics under non-stationary conditions

The model \mathcal{M}_T includes all the information about the current fault state as well as the relation between the operating conditions and the fault. Therefore it can be used to predict the evolution of the fault under variable operating conditions. For example, if the future time profile of the load is known, the model can predict the feature time series for that specific load profile. The example of such a prediction is shown in Figure 6

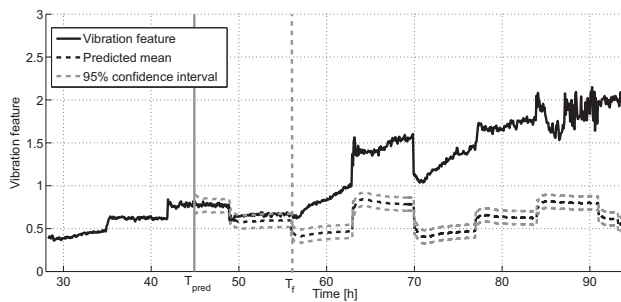


Figure 6. Long term prediction under variable load at $T_{pred} = 44h$

It can be seen that the model predicted a stable fault dynamics and the changes in the feature value only occur due to changes in the load. In the actual experiment, the initial fault occurred around the time $T_f = 55h$, which is impossible to predict with the model that is based only on the data up to time $T_{pred} < T_f$.

Effect like this may occur because the underlying model is linear and serves only as a local approximation. However, it is crucial to note that if such a fault occurs, it is reflected in the feature values data and the algorithm will quickly incorporate the new data into the model and produce the updated parameter values.

After the model is adapted to the new data, the prediction is updated and a result of a later prediction is shown in Figure 7.

It can be seen, that the actual feature value almost always lies within the 95% confidence interval of the prediction.

6. CONCLUSIONS

The paper presents a new approach for model-based prognostics of mechanical drives under non-stationary operating conditions. The novelty of the proposed algorithm lies in the use of dynamical model to describe the relations between operating conditions, fault dimension and vibration feature value.

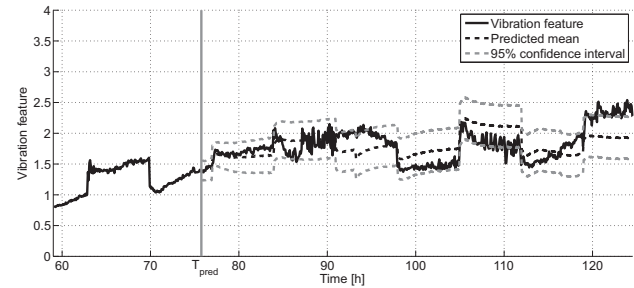


Figure 7. Long term prediction under variable load at $T_{pred} = 75h$

The model assumes linear relations between these quantities which can be interpreted as a local approximation of the otherwise complex nonlinear relations. The benefit of this approximation is that the model parameters can easily be estimated on-line. This means that the model is constantly updated as new data arrive.

The approach was validated on a laboratory test bed using a single-stage gearbox and vibration sensors. The problem was to detect and predict the faults in gear and the model analysis and prognostics on the experimental data validated our hypotheses.

Future work will include validation of the approach for estimation of the remaining useful life of the gear and examine how the RUL depends on the load profile. However, to properly conduct this study, further experiments are required.

REFERENCES

- Combet, F., & Gelman, L. (2009). Optimal filtering of gear next term signals for early damage detection based on the spectral kurtosis. *Mechanical Systems and Signal Processing*, 23, 652-668.
- DeCastro, J. A., Liang, T., Kenneth, L. A., Goebel, K., & Vachtsevanos, G. (2009). Exact nonlinear Filtering and Prediction in Process Model-Based Prognostics. In *Proceedings of the 1st Annual conference of the PHM Society, San Diego, USA, September 27 - October 1, 2009*.
- Edwards, D., Orchard, M. E., Tiang, L., Goebel, K., & Vachtsevanos, G. (2010). Impact of Input Uncertainty on Failure Prognostic Algorithms: Extending the Remaining Useful Life of Nonlinear Systems. In *Annual Conference of the Prognostics and Health Management Society, 2010*.
- Gašperin, M., Juričić, D., Boškoski, P., & Vižintin, J. (2011). Model-based prognostics of gear health using stochastic dynamical models. *Mechanical Systems and Signal Processing*, 25(2), 537-548.
- Gibson, S., & Ninness, B. (2005). Robust Maximum-Likelihood Estimation of Multivariable Dynamic Systems. *Automatica*, 41, 1667-1682.

- Haykin, S. (Ed.). (2001). *Kalman Filtering and Neural Networks*. John Wiley & Sons, New York, USA.
- Heng, A., Zhang, S., Tan, A. C., & Mathew, J. (2009). Rotating machinery prognostics: State of the art, challenges and opportunities. *Mechanical Systems and Signal Processing*, 23, 724-739.
- Howard, I., Jia, S., & Wang, J. (2001). The dynamic modelling of a spur gear in mesh including friction and crack. *Mechanical Systems and Signal Processing*, 15, 831-853.
- Orchard, M., Kacprzynski, G., Goebel, K., Saha, B., & Vachtsevanos, G. (2008). Advances in uncertainty representation and management for particle filtering applied to prognostics. In *International Conference on Prognostics and Health Management, 6-9 Oct. 2008, Denver, CO*.
- Orchard, M. E., & Vachtsevanos, G. J. (2009). A particle-filtering approach for on-line fault diagnosis and failure prognosis. *Transactions of the Institute of Measurement and Control*, 31, 221-246.
- Randall, R. (1982). A New Method of Modeling Gear Faults. *Journal of Mechanical Design*, 104, 259-267.
- Zhang, B., Khawaja, T., Patrick, R., Vachtsevanos, G., Orchard, M., & Saxena, A. (2009). A Novel Blind Deconvolution De-Noising Scheme in Failure Prognosis. In K. P. Valavanis (Ed.), *Applications of Intelligent Control to Engineering Systems* (Vol. 39, p. 37-63). Springer Netherlands.

Modeling wave propagation in Sandwich Composite Plates for Structural Health Monitoring

V. N. Smelyanskiy¹, V. Hafiychuk¹, D. G. Luchinsky¹, R. Tyson², J. Miller³, C. Banks³

¹ NASA Ames Research Center, Mail Stop 269-2, Moffett Field, CA 94035, USA

vadim.n.smelyanskiy@nasa.gov

² University of Alabama, 1804, Sparkman Dr., Huntsville, AL, 35816, USA

Richard.Tyson@nasa.gov

³ Marshal Space Flight Center/EM20, Huntsville, AL, 35812, USA

jim.miller@nasa.gov

ABSTRACT

Wave propagation is investigated in sandwich composite panels using analytical approach for layered materials, Mindlin plate theory and finite element modeling in the context of developing an on-board structural health monitoring system. It is shown that theoretical results are in agreement with the results of numerical simulations and with experimental results.

1. INTRODUCTION

Composite sandwich panels (CSP), consisting of fiber-reinforced facesheets separated by low-density cores, offer lightweight and flexible production capabilities and high performance: high strength, damage tolerance and thermal resistance (Zenkert, 1995), (Zenkert, 1997). During the past few decades, the CSPs have been steadily replacing the traditional materials in many industries including e.g. automotive, marine, and aerospace. Their stiffness-to-weight ratios and damage tolerance are especially attractive in aerospace industry leading to higher payloads (Bednarczyk, Arnold, Collier, & Yarrington, 2007). However, the multi-layered construction and laminate layout of the facesheets allow for debonding, delamination, and other internal flaws that are hardly visible and may severely damage the structural strength of the CSPs. In this context, it becomes important to develop reliable on-line structural health monitoring (SHM) systems of the composite panels. The aerospace industry has one of the highest payoffs for SHM since damage can lead to catastrophic failures.

There are several techniques currently under investigation (See, for example, (Raghavan & Cesnik, 2007)) for diagnostics including e.g. embedded fiber optic sensors for strain measurement, active ultrasonics, passive acoustic emission monitoring, and electromechanical impedance measure-

ments. The Lamb wave based diagnostics of CSPs is one of the most promising SHM techniques due to the similarity between the Lamb wavelength and the CSP thickness, the ability to travel far distances, high sensitivity, active sensing and low cost of piezoelectric wafer actuators/sensors (Raghavan & Cesnik, 2007). The development of the reliable SHM technique based on guided wave propagation in CSPs is complicated due to heterogeneity of the sandwich structures. This study is needed for better understanding and more reliable model predictions in the context of development of the in-flight SHM for the next generation of the heavy-lift vehicle.

2. MODELING WAVE PROPAGATION

A three-dimensional formulation relying on a global matrix technique provides a general framework for analysis of wave propagation in an anisotropic multi-layered medium (Zakharov, 2008). We consider symmetrical sandwich structures and the equation of motion in each layer reads

$$\partial_{\beta}\sigma_{mp}^j + \rho_j\omega^2 u_m = 0, \quad m, p, j = 1, 2, 3. \quad (1)$$

where for j -th layer ρ_j is the density. For an isotropic material the stresses σ_{mp}^j and strains ε_{mp}^j satisfy Hook's law and Kelvin-Voigt model of linear viscoelasticity and constitutive relations have the form

$$\sigma_{mp}^j = (\lambda'_j + \lambda''_j \partial_t) \delta_{mp} \varepsilon_{kk}^j + 2(\mu'_j + \mu''_j \partial_t) \varepsilon_{mp}^j.$$

For small displacements $u_j^{(n)}$ components of the deformation matrix $\varepsilon_{mp}^{(n)}$ are given by the following relation

$$\varepsilon_{mp}^j = \frac{1}{2} (\partial_p u_m^j + \partial_m u_p^j),$$

For the complex-valued representation of Lamé constants, wave speeds

$$\lambda_j = \lambda'_j - i\omega\lambda''_j, \mu_j = \mu'_j - i\omega\mu''_j. \quad (2)$$

V. N. Smelyanskiy et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

$$c_p^j = \sqrt{\frac{(\lambda_j + 2\mu_j)}{\rho_j}} = \sqrt{\frac{E_j(1 - \nu_j)}{\rho_j(1 + 2\nu_j)(1 - 2\nu_j)}} = \alpha|_j, \quad (3)$$

$$c_s^j = \sqrt{\mu_j/\rho_j} = \sqrt{E_j/2\rho_j(1 + \nu_j)} = \beta|_j, \quad (4)$$

$$k_p^j = \omega/c_p^j, k_s^j = \omega/c_s^j. \quad (5)$$

where material parameters $\alpha, \beta, \mu, \lambda E, \rho, \nu$ correspond to each layer $j = 1, 2, 3$.

The three-dimensional formulation assumes the continuity of displacements u_j and components of stress tensor σ_{ij} on the inner boundaries of the layers and allows for a quite general form of the boundary conditions. For the circular plate PZT sensors exiting dynamic stresses at a circular source could be considered symmetrical about its axis (0Z) and on the interfaces $x_3 = z_j, j = 1, 2, 3$.

Let's consider an infinite sandwich panel for theoretical modeling and square panel for FE simulation. The cylindrical coordinates (r, θ, z) are used for consideration where zeros of z-axis coincide with the midplane of the panel. In the simplest case due to axisymmetry the solution problem is two dimensional in coordinates (r, z) (Zakharov, 2008)

$$u_r^j = \left[-u^j B_n'(sr) + w^j \frac{n}{kr} B_n(sr) \right] \begin{Bmatrix} \cos n\theta \\ -\sin n\theta \end{Bmatrix}, \quad (6)$$

$$u_\theta^j = \left[u^j \frac{n}{kr} B_n(sr) - w^j B_n'(sr) \right] \begin{Bmatrix} \cos n\theta \\ -\sin n\theta \end{Bmatrix}, \quad (7)$$

$$u_z^j = v^j B_n(sr) \begin{Bmatrix} \cos n\theta \\ -\sin n\theta \end{Bmatrix}. \quad (8)$$

where the first or second term could be chosen in the French brackets, so they represent the terms in the trigonometrical Fourier series wrt θ . The terms $B_n = B_n(sr)$ are any of the appropriate Bessel function or Hankel function of the first or second kind and $B' = dB_n(sr)/d(sr)$,

$$\begin{bmatrix} u^j \\ v^j \end{bmatrix} = A_L^{+j} \begin{bmatrix} \cos C_\alpha z \\ \frac{C_\alpha^j}{k} \sin C_\alpha z \end{bmatrix} + A_L^{-j} \begin{bmatrix} \sin C_\alpha z \\ -\frac{C_\alpha^j}{k} \cos C_\alpha z \end{bmatrix} \\ + A_S^{+j} \begin{bmatrix} -\frac{C_\beta^j}{k} \cos C_\beta z \\ \sin C_\beta z \end{bmatrix} + A_S^{-j} \begin{bmatrix} \frac{C_\beta^j}{k} \sin C_\beta z \\ \cos C_\beta z \end{bmatrix}, \quad (9)$$

$$w^j = B_S^{+j} \cos C_\alpha z + B_S^{-j} \sin C_\beta z. \quad (10)$$

where $A_{L,S}^{\pm j}, B_S^{\pm j}$ are constants and displacement components are composed of symmetrical and anti-symmetrical terms according to $z = 0$, which are corresponding to symmetrical and anti-symmetrical modes, respectively. For the homogeneous material properties the general approach outlined above can be simplified allowing for further analytical treatment of the problem of finding dispersion relations of the sandwich panel (Zakharov, 2008), (Lowe, 1995). Accordingly, the Lamb wave dispersion relations are determined by

the determinant of square matrix of the 16x16 order

$$\det \begin{bmatrix} [D_{0b}] & [-D_{1t}] & & & & \\ & [D_{1b}] & & & & \\ & & [-D_{2t}] & & & \\ & & [D_{2b}] & & & \\ & & & [-D_{3t}] & & \\ & & & [D_{3b}] & & [-D_{0t}] \end{bmatrix} = 0, \quad (11)$$

where the D matrices for the top (index t) and bottom (index b) of a layer can be expressed, respectively, as

$$D_{jt} = \begin{bmatrix} s & sg_\alpha & C_\beta & -C_\beta g_\beta \\ C_\alpha & -C_\alpha g_\alpha & -s & -sg_\beta \\ i\rho B & i\rho B g_\alpha & -p_\beta C_\beta & p_\beta C_\beta g_\beta \\ p_\alpha C_\alpha & -p_\alpha C_\alpha g_\alpha & i\rho B & i\rho B g_\beta \end{bmatrix}_j, \quad (12)$$

$$D_{jb} = \begin{bmatrix} sg_\alpha & s & C_\beta g_\beta & -C_\beta \\ C_\alpha g_\alpha & -C_\alpha & -sg_\beta & -s \\ i\rho B g_\alpha & i\rho B & -p_\beta C_\beta g_\beta & p_\beta C_\beta \\ p_\alpha C_\alpha g_\alpha & -p_\alpha C_\alpha & i\rho B g_\beta & i\rho B \end{bmatrix}_j, \quad (13)$$

where $j = 0$ corresponds to air, s is the wave number in the direction of wave propagation,

$$p_\alpha = 2i\rho s\alpha^2, p_\beta = 2i\rho s\beta^2,$$

$$C_\alpha = (\omega^2/\alpha^2 - s^2)^{1/2}, C_\beta = (\omega^2/\beta^2 - s^2)^{1/2},$$

$$g_\alpha = e^{iC_\alpha z}, g_\beta = e^{iC_\beta z}, B = \omega^2 - 2\beta^2 s^2.$$

All matrices in (11) are 4×4 except D_{0b} and D_{0t} which are 4×2 .

Let us investigate dispersion curves for typical sandwich structure with soft core about 1in in thickness and carbon fibre reinforced plastic facesheet consisting of 14 ply. The velocities are determined by the geometry of the structure as well as longitudinal and shear velocities characterizing materials. In the simulations we used CFRP face sheets with Young modulus 60GPa, Poisson ratio 0.3 and density 1500kg/m³ and homogenized core with Young modulus $E=80$ MPa, the same Poisson ratio and density was 100kg/m³. As a result, longitudinal velocity was 7338m/s and 1038m/s and shear velocity 3922 and 555m/s, correspondingly. These velocities are plotted by green dashed lines. The spectrum of the waves (9),(10) is presented in Figure 1 and 2 for the case when the core of the sandwich ($j = c$) is much softer than for the facesheet. The vibration of the soft core is restricted by rigid surfaces of the face sheet and we have many local modes in the structure. The dispersion curves change very drastically if the sandwich core is viscoelastic. Many curves corresponding to honeycomb core just disappear (Figure 2) and for the case when the real and imaginary parts of the elastic module become comparable the propagation is determined by facesheet modes.

Analyzing dispersion plots for low velocities we can see that at low velocities dispersion curves are modulated by facesheet flexural velocity. At higher frequency these curves tend to shift to shear velocity of the core. For high velocities dispersion curves exhibit a set of vertical paths where phase

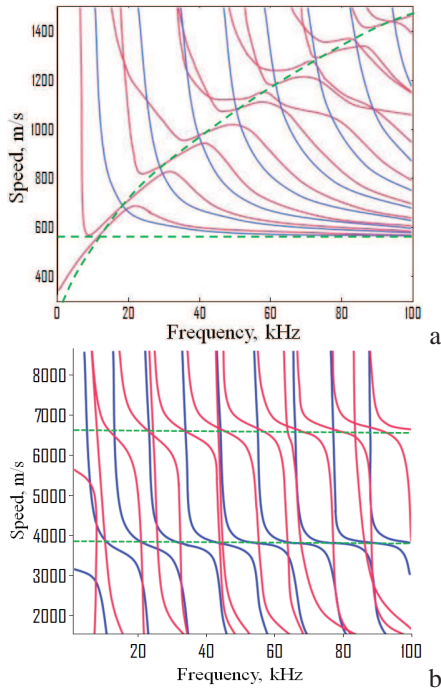


Figure 1: Dispersion curves of Lamb wave of phase velocities with respect to frequency for symmetric sandwich, (a) – low speed limit and (b) – high speed limit, blue lines - shear waves, lilac lines - Lamb waves).

velocity changes by a large value at the same frequency. The group velocity at these frequencies (vertical lines) is close to zero and this means waves are practically standing. This statement is confirmed by FE simulations.

If the core is viscoelastic we have coupling of different modes and the dispersion curves start to intersect, some of the modes vanish and some of them change their trend with increasing frequency (Fig. 2). With introducing viscoelasticity, as seen from the plots, high velocity Lamb wave modes tend to S0 modes of the facesheet and shear modes to shear velocity in the facesheet. At a low frequency the limit coupling is not so pronounced but the tendency is that we have two characteristic velocities here: shear velocity of the core and flexural velocity of the facesheet (green dashed line in the Fig. 1). These two modes mainly determine the form of the dispersion curves at the low velocity limit.

It should be noted that attenuation increases very sharply in the frequency range where coupling takes place. This can be seen from simulations presented in Figure 3 when red dashed lines vanish due to the interaction between modes associated with viscoelasticity. The small interval in 1-10kHz is plotted to see how coupling between S0 and A1 arises with small viscoelasticity ($E'' = 0.01E'$). In this case standing modes transform into propagating modes with high attenuation in the frequency region of coupling. As can be seen from Fig. 3

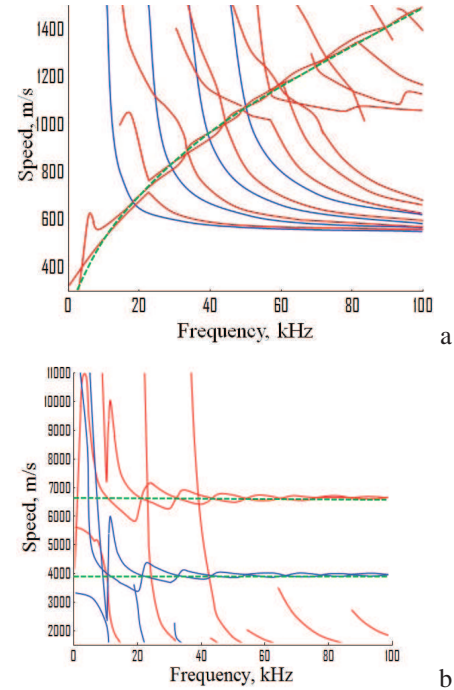


Figure 2: Dispersion curves of Lamb wave of phase velocities with respect to frequency for symmetric sandwich with viscoelastic core $\lambda'_c = 0.1\lambda_c$, $\mu'_c = 0.1\mu_c$; (a) – low speed limit, and (b) – high speed limit. Blue lines - shear waves, lilac lines - Lamb waves.

attenuation increases practically in the order at the frequency region of 7.8 kHz where coupling takes place (dark dashed line).

A more simplified approach to investigation of wave propagation for SHM in CSP lies in using averaged over thickness parameters of the structures since we obtain 2D model in contrast to 3D theory considered above. In many cases, such approach is sufficiently good since it makes it possible to find a simpler analytical solution for propagating waves than the solution described by formulas (9),(10). The next section is devoted to the review of the Mindlin plate theory and the application of this approach to wave propagation modeling.

3. MINDLIN PLATE THEORY FOR SANDWICH STRUCTURES

In the Mindlin plate theory the displacements of the plate in the transverse, radial, and tangential direction components are expressed as follows (Mindlin & Deresiewicz, 1954)

$$w = w(r, \theta, t), u = z\psi_r(r, \theta, t), v = z\psi_\theta(r, \theta, t),$$

where z is the coordinate defining points across the thickness of the plate ($z = 0$ is the neutral plane), w is the out-of-plane displacement of the wave, ψ_r and ψ_θ are the rotations of vertical lines perpendicular to the mid-plane.

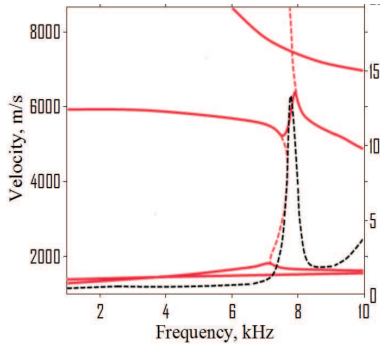


Figure 3: Dispersion curves of Lamb wave of phase velocities with respect to frequency for symmetric sandwich panel (red solid line corresponds to viscoelastic core $E' = 0.01E''$, dashed curve - elastic core $E'' = 0$, black dashed line is at attenuation of the coupled mode, $[Np/m]$).

The governing equations for the symmetric honeycomb panels in terms of moments and shear forces can be presented based on shell approximation by the following approach (Rose & Wang, 2004)

$$\frac{1}{r} \frac{\partial Q_\theta}{\partial \theta} + \frac{\partial}{\partial r} Q_r + \frac{1}{r} Q_r - Q_\theta = \rho \frac{\partial^2}{\partial t^2} w, \quad (14)$$

$$\frac{\partial M_{rr}}{\partial r} + \frac{1}{r} M_{rr} - \frac{1}{r} M_{\theta\theta} + \frac{1}{r} \frac{\partial}{\partial \theta} M_{r\theta} - Q_r = I \frac{\partial^2}{\partial t^2} \psi_r, \quad (15)$$

$$\frac{1}{r} \frac{\partial M_{r\theta}}{\partial r} + \frac{2}{r} M_{r\theta} + \frac{1}{r} \frac{\partial}{\partial \theta} M_{\theta\theta} - Q_\theta = I \frac{\partial^2}{\partial t^2} \psi_\theta, \quad (16)$$

where $\rho = \sum_{k=1}^3 \int_{a_k}^{b_k} \rho_k dz$, is the mass density per unit area of the plate, index k corresponds to the material layer, ρ_k is the density, $I = \sum_{k=1}^3 \int_{a_k}^{b_k} \rho_k z^2 dz$ is the mass moment of inertia. Each layer in the sandwich panel is bounded by the coordinates a_k and b_k in the thickness direction as shown in the Figure 4 (a). The stress resultants in terms of moments M_{rr} , $M_{\theta\theta}$, and $M_{r\theta}$, along with shear forces Q_r and Q_θ can be related to the transverse displacements and rotations as follows:

$$M_{rr} = \frac{D}{r} \left[r \frac{\partial \psi_r}{\partial r} + \nu (\psi_r + \frac{\partial \psi_\theta}{\partial \theta}) \right], \quad (17)$$

$$M_{r\theta} = \frac{D(1-\nu)}{2r} \left[\frac{\partial \psi_r}{\partial \theta} - \psi_\theta + r \frac{\partial \psi_\theta}{\partial r} \right], \quad (18)$$

$$M_{\theta\theta} = \frac{D}{r} \left[\nu r \frac{\partial \psi_r}{\partial r} + \psi_r + \frac{\partial \psi_\theta}{\partial \theta} \right], \quad (19)$$

$$Q_r = 2\kappa^2 G \left(\psi_r + \frac{\partial}{\partial r} w \right), \quad (20)$$

$$Q_\theta = 2\kappa^2 G \left(\psi_\theta + \frac{1}{r} \frac{\partial}{\partial \theta} w \right), \quad (21)$$

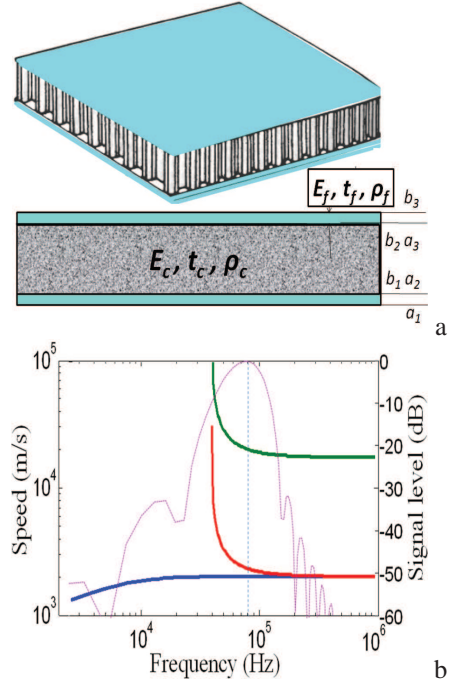


Figure 4: (a) - General view of the panel, (b) - Dispersion curves of the velocities with respect to frequency of symmetric sandwich panel in Mindlin approximation.

where $D = \frac{E_f t_f^3}{6} + \frac{E_c t_c^3}{12} + \frac{E_f t_f (t_f + t_c)^2}{4}$ - is the flexural stiffness, ν is the Poisson ratio, which for the sake of simplicity is taken as equal for each layer, E_f, E_c are the Young's modulus of the facesheet and the core, correspondingly, t_f, t_c - thicknesses of the facesheet and the core layers, G is the shear stiffness of the plate, κ is the shear correction factor ~ 1 .

The general solution of the acoustic waves propagation with cycling frequency ω is $(w, \psi) = Re [(W, \Psi) \exp(-i\omega t)]$ (Rose & Wang, 2004), where and throughout this paper $Re(\cdot)$ denotes the real part of the quantity appearing in parentheses, $\psi = (\psi_r, \psi_\theta)$. The variables W, Ψ are presented by expressions

$$W = W_1 + W_2,$$

$$\Psi = \xi_1 \nabla W_1 + \xi_2 \nabla W_2 - e_z \times \nabla V,$$

where e_z is a unit vector in z direction (the displacement is $u_z = w e_z$, normal stress and strain in the thickness direction of the plate are not included in Mindlin plate theory), W_1, W_2, V satisfy three Helmholtz equations

$$\Delta W_1 + k_1^2 W_1 = 0, \quad (22)$$

$$\Delta W_2 + k_2^2 W_2 = 0, \quad (23)$$

$$\Delta V + k_3^2 V = 0 \quad (24)$$

and Δ, ∇ - Laplace and Nabla operators, correspondingly. For isotropic sandwich layers

$$k_{1,2}^2 = \frac{1}{2}(k_p^2 + k_s^2) \pm \sqrt{k_f^4 + \frac{1}{4}(k_p^2 - k_s^2)^2}, \quad (25)$$

$$k_3^2 = \frac{k_1^2 k_2^2}{k_p^2} \quad (26)$$

$$k_s = \omega/c_s, k_p = \omega/c_p, k_f = (\rho\omega^2/D)^{1/4}, \\ c_s = (G/\rho)^{1/2}, c_p = (D/I)^{1/2}, \xi_j = (k_s/k_j)^2 - 1.$$

The dispersion curves for a typical sandwich panel corresponding to three branches are shown in the Figure 4 (b). The flexural wave corresponds to the real ω, k in whole ω domain. The second (and third) dilatation branch of (k, ω) dependence become real starting from the cutoff frequency.

For a similar 3D consideration we will consider a circular-patch actuator on the Mindlin plate generated by a surface traction plate waves in the form (Mindlin & Deresiewicz, 1954). In this case, the radially directed bending moments m_r , uniformly distributed along the ring of radius r_0 , can be described as follows:

$$m_r = \frac{1}{2}hp(t)\delta(r - r_0), \quad m_{r\theta} = 0,$$

where $p(t)$ is the amplitude of the force.

The source term for circular force leads to the solution for out-of-plane displacement

$$w(r, \omega) = s_1 H_0(k_1 r) + s_2 H_0(k_2 r), \quad (27)$$

where radius vector r is counted from the center of the actuator and coefficients s_1, s_2 are presented by

$$s_1 = \frac{i\pi hp(\omega)}{4D} \frac{k_1 r_0 J_1(k_1 r_0)}{k_1^2 - k_2^2}, \\ s_2 = -\frac{i\pi hp(\omega)}{4D} \frac{k_2 r_0 J_2(k_2 r_0)}{k_1^2 - k_2^2},$$

where J_n and H_n are the Bessel and the Hankel functions of the first kind, respectively. We will consider that the frequency f of the source is sufficiently high $\omega = 2\pi f > \omega_c$, where ω_c is the cutoff frequency $\omega_c = (G/I)^{1/2}$. As a result, the propagation spectrum is determined by two real wavenumbers k_1 and k_2 . Expressions for rotations $\psi = (\psi_r, \psi_\theta)$ and, consequently, $u(r, \omega), v(r, \omega)$ can be found in the article (Mindlin & Deresiewicz, 1954).

4. DYNAMICS. TRANSIENT SOLUTION

To study the transient wave propagation we consider that the plate is excited by a pulse of the load stimulated by a PZT sensor (Raghavan & Cesnik, 2007). The expression for the wave pulses in the plane (x, y) may be derived from the steady-state solution in the frequency domain by applying the Fourier transform technique.

Let us consider that any pulse of the wave can be expanded into the Fourier transform which represents pulse as a series of plane waves. If the Fourier spectrum $G(\omega)$ of the signal $g(t)$ is

$$g(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(\omega) e^{-i\omega t} d\omega, \quad (28)$$

then the final solution for mechanical fields in the time domain will be

$$\begin{bmatrix} w^j(r, t, \theta) \\ v^j(r, t, \theta) \\ w^j(r, t, \theta) \end{bmatrix} = Re \frac{1}{2\pi} \int_{-\infty}^{\infty} G(\omega) \begin{bmatrix} w^j(\omega) \\ v^j(\omega) \\ w^j(\omega) \end{bmatrix} e^{-i\omega t} d\omega \quad (29)$$

For Mindlin plate theory we omit indices j . Let us present the results obtained by (29) and compare them with the results of the direct computer simulation of the real honeycomb plate. We consider the Hanning type actuation signals which are usually used for fault detection in SHM (Raghavan & Cesnik, 2007). The Hanning signal can be presented in the form

$$g(t) = [\Theta(t) - \Theta(t - 2\pi N/\omega_0)] [1 - \cos(\omega_0 t/N)] \sin(\omega_0 t) \quad (30)$$

where N is a parameter of impulse, $f_0 = \omega_0/2\pi$ is a carrier frequency, $\Theta(t)$ is the Heaviside step function. The selection of the driving frequency f was made in the frequency range from 20 to 100 kHz and this selection is critical for Lamb waves generation and fault detection.

We compare the analytical results with the corresponding results obtained by the Finite Element simulation. The FE modeling for 2D Mindlin plate is presented in Fig. 6. and it fits well with the theoretical approach. The main difference between these signals is that the theoretical results are valid for the infinite plate, and the FE 2D Mindlin plate model provides the result which takes into account reflections from the boundaries. The comparison of the 3D modeling of sandwich composite panels with the theoretical result is considered in the last section.

4.1 Propagation of the signal

As an example, for simulation we used Hanning pulses with 3.5 windowed input waveform with different carrying frequency f . The Fourier transform of such signal is presented in Fig. 5 a) for $f = 100 \text{ kHz}$ (dashed red line). Dispersion curves are presented in Fig. 5 (b) calculated according to the characteristic equation (11). It can be seen that the main domain of narrowband 3.5 windowed input waveform taking part in wave propagation is sufficiently broad (the domain between two red lines in dispersion curves Fig. 5 (b), which is taken, for example, on the level of 10 db Fig. 5 (a)). The modes A_0 and S_0 of the facesheets are the leading ones in the formation of wave propagation through the structure. The

wave velocities $v = \omega/k$, where $\omega = 2\pi f$, k is a wavenumber, of the soft core are much lower than the facesheet velocities and that is why most dispersion curves have much lower slope than facesheet modes, except for the small vicinity of the facesheet modes A_0 and S_0 . As a result, the generation of the Hanning windowed signal with leading frequency f ($f = 100kHz$ on the plot) leads to generation of antisymmetric and symmetric zeros modes of the facesheets coupled with a large number of local modes of the soft core.

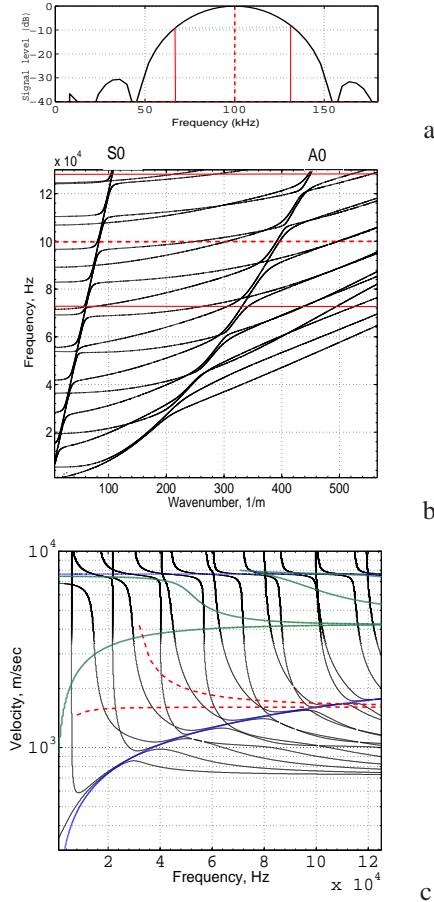


Figure 5: Dispersion relation for symmetric sandwich panel, (a) – Spectrum of the input signal for $f = 100kHz$, (b) – Honeycomb layered structure modes, (c) – Lamb velocities modes. (Blue lines - facesheet modes, red line - homogenized panel modes, dashed line - modes of Mindlin plate theory approach).

The phase velocities of the sandwich panel are presented in the Fig. 5 c (black lines) and correspond to the same dependencies $\omega = \omega(k)$ of the Fig. 5 (a). It can be easily seen that for the considered impute signal many modes propagate in the structures. If we consider the facesheet itself then only the S_0 mode and the A_0 mode can propagate in the considered frequency range below or at the order of $100kHz$ (blue

lines in the Fig. 5 (c). In this region, the S_0 mode is almost non-dispersive, and the A_0 mode is slightly dispersive. Considering the structure of the dispersion curves (Fig. 5 b) and velocity curves (Fig. 5 c) we can confirm that the majority of wavenumbers of the modes in dispersion curves are located in the small vicinity of A_0 and S_0 modes and in this case facesheet modes are much more sensitive to debond delamination defects than the core modes synchronizing vibration of the two facesheets.

For comparison, we will consider here the dispersion relations of Mindlin plate theory for symmetric sandwich structures (Dashed red line in Fig. 5 (c)). The theory of the sandwich panels is considered in (Zenkert, 1995). We used analytical formulas from these sources just to identify coefficients in Mindlin plate theory used for investigation of wave propagation. The Mindlin plate theory approach shows that dispersion curves in the vicinity of $f=100kHz$ are sufficiently close to the antisymmetric mode of the facesheet (Fig. 4).

Dispersion curves in coordinates (v, f) for homogenized plate are presented by green curves and they show quite different dispersion curves (Fig. 5c). In this case we can expect that a simplified approach can not completely describe wave propagation, and wave patterns in honeycomb structure are much more diverse. This is especially true for a high frequency excitation signal like $100kHz$.

5. FINITE ELEMENT MODEL

The SCP has two main components, namely two stiff facesheets and a soft core between them. In addition to these subcomponents, we will consider an adhesive layer binding facesheets with the core. The thickness of the adhesive layer is generally sufficiently small but this component is important for simulation of the debond origination and growth. We also consider PZT actuator and sensors mounted on the panel. As a result of simulation, electrical signals of the sensors were compared with the signals obtained experimentally. Such approach best fits the typical sketch we have in SHM when the measured signals are used for interpretation of changes in monitored panels.

The FE model of the honeycomb sandwich structures with a piezoelectric actuator/sensor distribution is shown in Fig. 6. The model consists of the honeycomb core and two laminated facesheets with an actuator and sensors attached to the top sheet.

5.1 Facesheet

The facesheet in Abaqus can be modeled using shell, continuum shell, or solid element types (Fig. 7). We have found that continuum shell element type for the facesheet provides performance that is close to optimal. The facesheets were made by graphite/epoxy with lay-up sequence of $[0/90]$. In all cases, the composite layup is modeled explicitly as shown

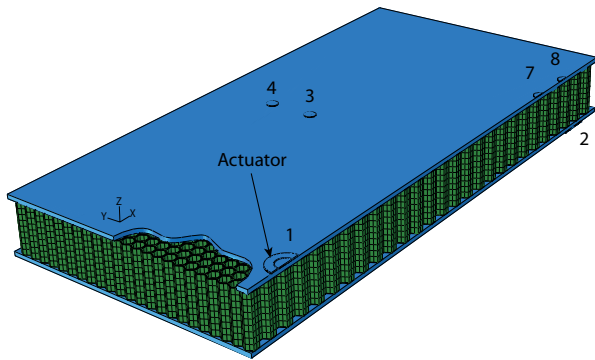


Figure 6: Finite element model of the sandwich honeycomb structure with a piezoelectric actuator (shown by an arrow) and a set of sensors (marked by the numbers).

Table 1: Parameters of the facesheet.

Ply elastic modulus E_{11}	16 Msi
Ply elastic modulus E_{22}	1.2 Msi
Ply Poisson's ratio ν_{12}	0.3
Ply shear modulus G_{12}	0.6 Msi
Ply thickness	6 mils
Laminate thickness	84 mils

in the figure 7. The parameters of the ply are shown in the table

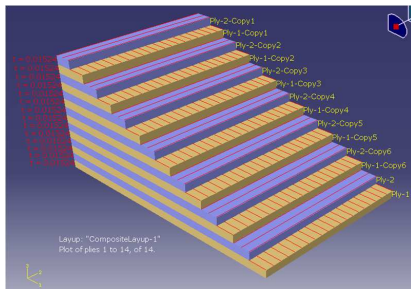


Figure 7: The composite layup of the facesheets consisting of 14 layers with orientation 0 and 90 degrees. The parameters of the lamina are shown in the Table1.

5.2 PZT sensors and actuators

Geometrical properties of PZT elements (Fig. 8) of the model are summarized in the Table2. The response of the PZT elements was determined by the piezoelectric stress matrix e and

Table 2: Parameters of the actuator and sensors.

Actuator diameter	0.709"
Actuator inner diameter	0.394"
Actuator thickness	20 mils
Sensor A diameter	0.354"
Sensor A thickness	20 mils
Sensor B diameter	0.250"
Sensor B thickness	10.5 mils

elasticity matrix c

$$[e] = \begin{bmatrix} 0 & 0 & -5.4 \\ 0 & 0 & -5.4 \\ 0 & 0 & 15.8 \\ 0 & 0 & 0.0 \\ 0 & 12.3 & 0.0 \\ 12.3 & 0 & 0.0 \end{bmatrix} [Cm^{-2}] \quad (31)$$

$$[c] = \begin{bmatrix} 12.1 & 7.54 & 7.52 & 0 & 0 & 0 \\ 7.54 & 12.1 & 7.52 & 0 & 0 & 0 \\ 7.52 & 7.52 & 11.1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2.26 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2.11 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2.11 \end{bmatrix} \times 10^{10} [Pa] \quad (32)$$

The dielectric matrix of the PZT material has the following diagonal elements $\epsilon_{11} = \epsilon_{22} = 8.11 \times 10^{-9} [C/V/m]$ and $\epsilon_{33} = 7.35 \times 10^{-9} [C/V/m]$. The density of the PZT material is $\rho_{PZT} = 7750 [kg/m^3]$.

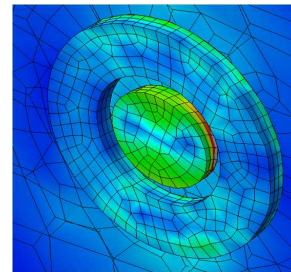


Figure 8: Snapshot of the actuator and sensor A during the simulations.

5.3 Honeycomb core

A special attention was paid to modeling of the detailed honeycomb structure (Fig. 9 a,b,c) including the difference in thickness for different walls of the structure and the presence of bending tips. The structure was built from a single strip shown in the Fig. 9(c). The bending tips were attached to the structure using the boolean operation on the mesh. The parameters characterizing mechanical properties of the honeycomb structure are listed in Tab. 3

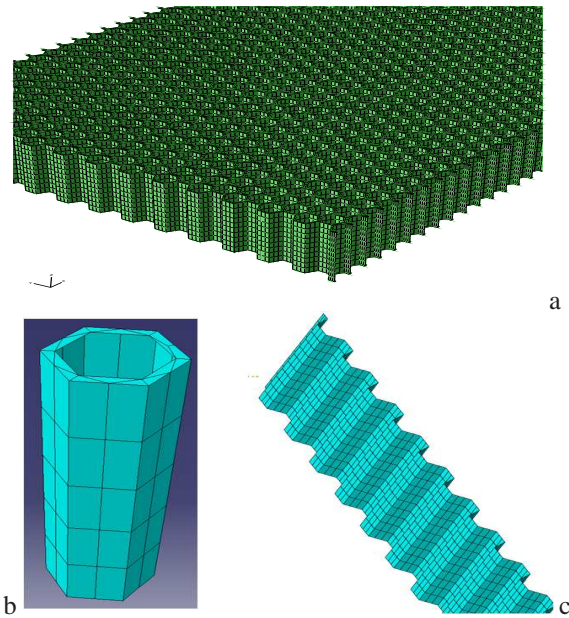


Figure 9: (a) Detailed view of the honeycomb structure. (b) A single cell of the structure with bending tips. (c) A single strip of the core used to build the structure.

Table 3: Parameters of the honeycomb core.

Cell size	0.25"
Shear modulus, ribbon direction (L)	70 ksi
Shear modulus, transverse direction (W)	40 ksi
Density	5.2 lbs/ft ³
Shear strength (L)	380 psi
Shear strength (W)	220 psi
Thickness	1"

The material properties of the Aluminum used to build the structure are the following: Young's modulus $E_{Alm} = 7.3084 \times 10^{10}$ [Pa], Poisson's ratio $\nu_{Alm} = 0.33$, Mass density $\rho_{Alm} = 2700$ [kg/m³].

5.4 Adhesive layers

An important property of the honeycomb sandwich structure is the presence of adhesive layers both between the actuator/sensors and the facesheet and between the facesheet and the honeycomb core (Fig. 10). Accordingly, the layer with the following properties (Young's modulus $E_{Adh} = 4.82 \times 10^9$ [Pa], Poisson ratio $\nu_{Adh} = 0.40$, and mass density $\rho_{Adh} = 1255$ [kg/m³]) was explicitly included into the finite element model.

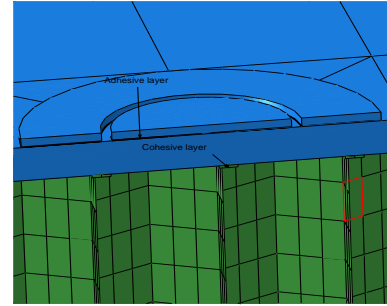


Figure 10: The location of the adhesive layer between PZT elements and facesheet and cohesive layer between facesheet and honeycomb.

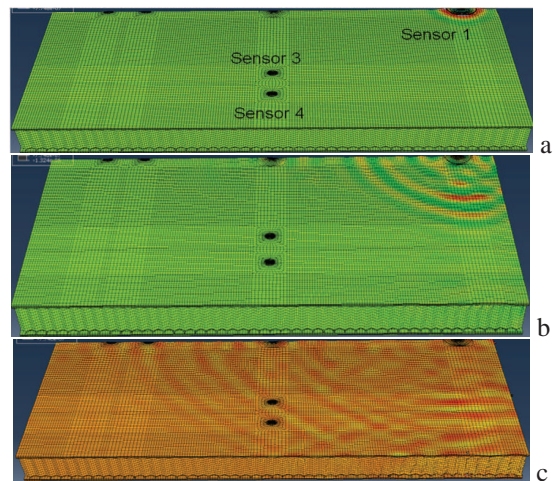


Figure 11: Finite element modeling of the sandwich honeycomb structure with a piezoelectric actuator. (a) – out-of-plane displacement for $t = 0.02ms$, (b) – $t = 0.06ms$, (c) – $t = 0.16ms$, (PZT sensors corresponding experimental layup are denoted black circles).

6. FE MODELING WAVE PROPAGATION

6.1 Numerical results

An experiment in Lamb wave propagation in a honeycomb sandwich panel was done by Metis Design Inc. in collaboration with ARC NASA. The sandwich panel fabricated for this test consisted of two 84-mil thick cross-ply carbon fiber composite laminates (bonded to a 1"- thick aluminum honeycomb core). The size of the panel was $1ft \times 1ft$. In the experiment, PZT sensors located on the facesheet of the honeycomb panel were used to determine the deformation at a different location (Fig. 6). The primary goal of this study is to model wave propagation field in 3D sandwich honeycomb panel to fit these results to SHM experimental data. Lamb wave tests were done over a frequency range from 20 to 100 kHz and 3.5-cycle Hanning windowed toneburst was used as

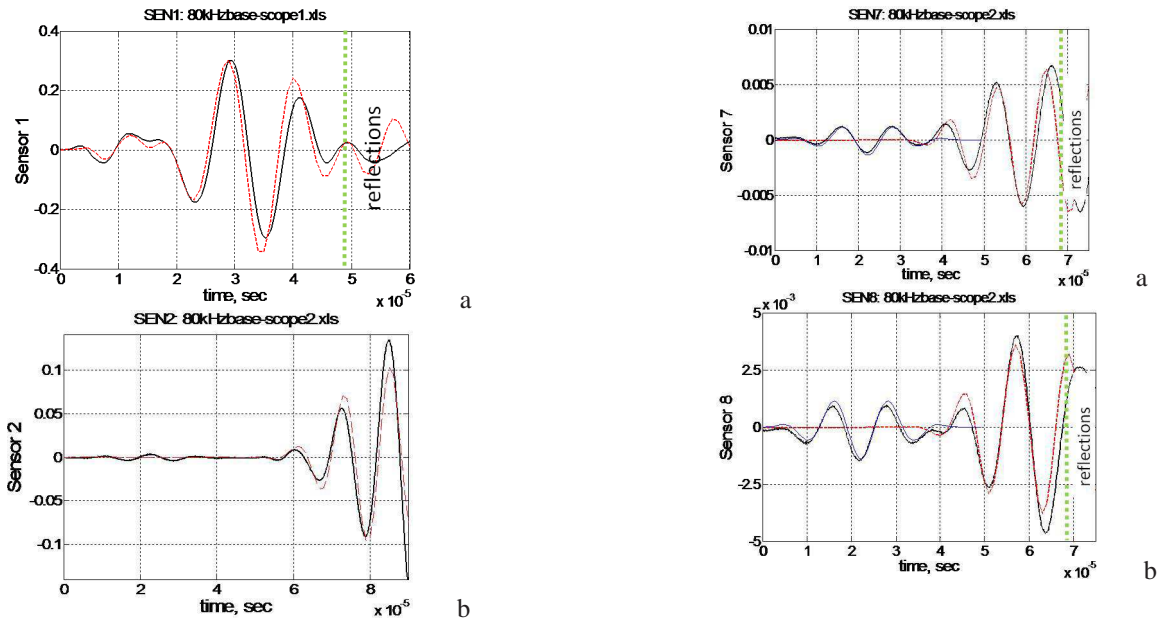


Figure 12: The results of the simulations of the wave propagation in honeycomb structure are shown in comparison with the experimental results measured for 80 kHz on (a) sensor # 1 embedded into actuator and (b) a similar sensor # 2 on the other side of the structure.

an actuation signal for SHM.

First, we simulated a generation of acoustic modes by annular PZT patch and studied wave propagation. Second, different voltage generating signals were used to obtain transient fields which generate an electrical signal in a set of PZT sensors mounted on the facesheet plane in a particular experiment. Finally, we compared an electrical signal in pitch catch and pulse-echo technique and showed a very good agreement of both the theoretical and experimental data. Typical view of the FE simulations results is presented in Fig. 11. It can be easily seen from Fig. 11 that magnified displacement actuated by transducer located at the top of right-hand side corner propagates through the structure and generates electric signal in PZT sensors mounted to the top facesheet.

Results for voltage measurement for different sensors mounted on the plate are presented in Figures 12 and 13 for $f = 80\text{kHz}$, which shows the voltage on the corresponding sensors as a function of time at two locations, $x = -0.135\text{m}$ ($y=6\text{in}$) and $x = 0.135\text{m}$ ($y=6\text{in}$), respectively, (the panel is centered at (0,0) and all sensors are positioned with respect to the center).

Theoretical and experimental results fit very well at the initial stage of wave propagation. The wave modes reflected from the boundaries of the plate lead to change in the phase of strain vibration and are not identical in the instance when we have reflections from the boundaries.

Figure 13: The results of the simulations of the wave propagation in honeycomb structure are shown in comparison with the experimental results measured for 80 kHz on (a) sensor # 7 and (b) sensor # 8.

The results of simulations of the wave propagation in a honeycomb sandwich structure are shown in comparison with the experimental results for $f = 100\text{kHz}$ for a much longer time period, and they are presented in Fig. 14. You can see that when the time is longer we have stronger discrepancy between the theoretical and the experimental results even for the pristine panel without any damage. This is probably due to imperfections in the panel manufacturing and nonperfect boundary conditions in experiment in contrast to perfect geometry we use in numerical simulation.

7. CONCLUSION

We have investigated the wave propagation in sandwich honeycomb panels. A narrowband excitation waveform is employed to study wave propagation and damage detection in CSP. The new detailed model of SCP is developed. Computer simulation of the wave propagation is performed and results of the strains are compared with those obtained by experimental testing. For this the PZT sensors mounted on a composite facesheet plate are used. It has been demonstrated that initial stages of the propagating pulse practically always fit each other. For much longer time intervals many reflections from the boundaries change the phases of the strain oscillations, and it is not always possible to fit theoretical and experimental signals well. The conducted analysis has shown that in thick (1in core) sandwich panels with Al honeycomb structure acoustic signal generated by the PZT actuator can be easily detected. Simulations have demonstrated practically the same response of the sensors we have in the experiment. The

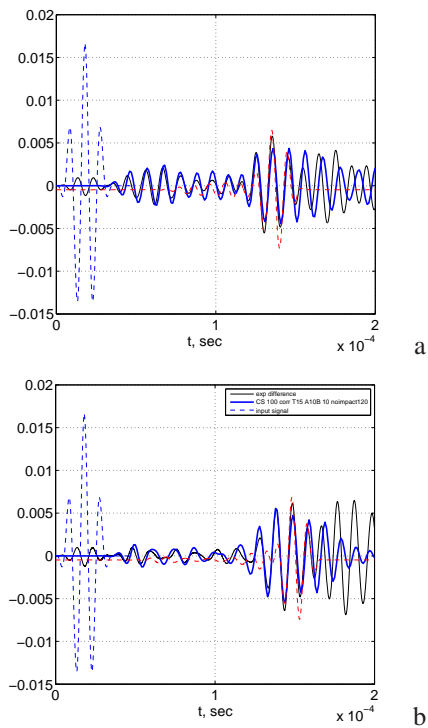


Figure 14: The results of the simulations of the wave propagation in honeycomb structure are shown in comparison with the experimental results measured for 100 kHz on (a) sensor # 3 and (b) sensor # 4. The blue line corresponds to FE modeling, Black one to experimental results and dashed red line to the modes calculated by Mindlin plate theory approach.

obtained results allow us to use FE methods for simulation of the acoustic waves propagating in the panel. The obtained results open up the prospect of the development of the SHM methods for advanced composite panels. This study makes it possible to deeper understand the physics based processes for the development of SHM methods. It should be pointed out that the FE model developed in this study has only been tested on one sample CSP and additional study will be necessary.

This paper addresses the different approaches to simulation of the guided wave propagation in sandwich structures with the emphasis given to the properties which can be used for SHM. The analytical investigation of dispersion curves, the plate wave using the Mindlin plate theory and the numerical simulations shows the main features we come across when developing real SHM methods. An analytical study is carried out to find the solution for transient wave propagation. The obtained analytical solutions are compared to the FE analysis as well as the experimental data.

REFERENCES

- Bednarczyk, B., Arnold, S., Collier, C., & Yarrington, P. (2007). Preliminary Structural Sizing and Alternative Material Trade Study for CEV Crew Module.
- Lowe, M. J. S. (1995). Matrix Techniques for Modeling Ultrasonic Waves in Multilayered Media. *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, 42(1), 154-171.
- Mindlin, R. D., & Deresiewicz, H. (1954). Thickness-shear and flexural vibrations of a circular disk. *Journal of applied physics*, 25(6), 1328-1332.
- Raghavan, A., & Cesnik, C. E. S. (2007). Review of guidedwave structural health monitoring. *The Shock and vibration digest*, 39(1), 91-114.
- Rose, R. F., & Wang, C. H. (2004). Mindlin plate theory for damage detection: Source solutions. *J. Acoust. Soc. Am.*, 116(1), 154-171.
- Zakharov, D. D. (2008). Orthogonality of 3D guided waves in viscoelastic laminates and far field evaluation to a local acoustic source. *International Journal of Solids and Structures*, 45(1), 1788-1803.
- Zenkert, D. (1995). *An Introduction to Sandwich Structures*.
- Zenkert, D. (1997). *Ed. The Handbook of Sandwich Construction*, EMAS Ltd, Warley, UK.

Online Estimation of Lithium-Ion Battery State-of-Charge and Capacity with a Multiscale Filtering Technique

Chao Hu¹, Byeng D. Youn^{2,*}, Taejin Kim² and Jaesik Chung³

¹*Department of Mechanical Engineering, University of Maryland, College Park, MD 20742, USA
huchaost@umd.edu*

²*School of Mechanical and Aerospace Engineering, Seoul National University, Seoul, 151-742, South Korea
bdyoun@snu.ac.kr*

³*PCTEST Engineering Laboratory, Columbia, Maryland 21045, USA
anto@pctestlab.com*

ABSTRACT

Real-time prediction of state-of-charge (SOC), state-of-health (SOH) and state-of-life (SOL) plays an essential role in many battery energy storage applications, such as electric vehicle (EV), hybrid electric vehicle (HEV) and smart power grid. However, among these three quantities, only the SOC has been thoroughly studied while there is still lack of rigorous research efforts on the other two quantities, SOH and SOL. Specially, real-time estimation of the SOH-relevant cell capacity by tracking readily available measurements (e.g., voltage, current and temperature) is still an open problem. Commonly used joint/dual extended Kalman filter (EKF) suffers from the lack of accuracy in the capacity estimation since (i) the cell voltage is the only measurable data for the SOC and capacity estimation and updates and (ii) the capacity is very weakly linked to the cell voltage. Furthermore, although the capacity is a slowly time-varying quantity that indicates cell state-of-health (SOH), the capacity estimation is generally performed on the same time-scale as the quickly time-varying SOC, resulting in high computational complexity. To resolve these difficulties, this paper proposes a multiscale framework with EKF for SOC and capacity estimation. The proposed framework comprises two ideas: (i) a multiscale framework to estimate SOC and capacity that exhibit time-scale separation and (ii) a state projection scheme for accurate and stable capacity estimation. Simulation and experimental results verify the effectiveness of our framework. †

* Corresponding author.

† This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

1. INTRODUCTION

As a battery cell ages, the cell capacity and resistance directly limit the pack performance through capacity and power fade, respectively (Plett, 2004a). These two degradation parameters are often used to quantify the cell state of health (SOH). Thus, it is important to accurately estimate these parameters to monitoring the present battery SOH and to predict the remaining useful life (RUL). Recent literature reports various approaches to estimate the SOH with a focus on the capacity estimation. Joint/dual extended Kalman filter (EKF) (Plett, 2004a) and unscented Kalman filter (Plett, 2006a) with an enhanced self-correcting model were proposed to simultaneously estimate the SOC, capacity and resistance. To improve the performance of joint/dual estimation, adaptive measurement noise models of the Kalman filter were recently developed to separate the sequence of SOC and capacity estimation (Lee et al., 2008). A physics-based single particle model was used to simulate the life cycling data of Li-ion cells and to study the physics of capacity fade (Zhang and White, 2008a; Zhang and White, 2008b). A Bayesian framework combining the relevance vector machine (RVM) and particle filter was proposed for prognostics (i.e., RUL prediction) of Li-ion battery cells (Saha et al., 2009). More recently, the particle filter with an empirical circuit model was used to predict the remaining useful lives for individual discharge cycles as well as for cycle life (Saha and Goebel, 2009).

Among these techniques, the joint/dual estimation technique is capable of real-time SOC and capacity estimation. Although it provides highly accurate SOC estimation, it suffers from the lack of accuracy in the capacity estimation since (i) the cell voltage is the only directly measurable data for the measurement updates in the SOC and capacity estimation (indirectly measurable data such as electrochemical impedance

require additional devices) and (ii) the capacity is very weakly linked to the cell voltage. Due to the strong correlation between the SOC and capacity, inaccurate capacity estimation may further lead to inaccurate SOC estimation and vice versa. Furthermore, although the capacity is a slowly time-varying quantity that indicates cell state-of-health (SOH), the capacity estimation is generally performed on the same time-scale as the quickly time-varying SOC, resulting in high computational complexity. To resolve these difficulties, this paper proposes a multiscale framework with EKF for SOC and capacity estimation. The proposed framework comprises two ideas: (i) a multiscale framework to estimate SOC and capacity that exhibit time-scale separation and (ii) a state projection scheme for accurate and stable capacity estimation. It is noted that the multiscale framework is generic since it can be used to achieve highly-confident health prognostics for any engineered system with multiple time-scales.

This paper is organized as follows. Section 2 describes the discrete-time state-space model of an engineered system with multiple time-scales. Section 3 reviews the numerical formulation and implementation of the dual EKF method. Section 4 presents the proposed multiscale framework with EKF and introduces the state projection scheme. The proposed ideas are applied to a Li-ion battery system to estimate SOC and capacity in Section 5. Section 6 contains simulation and experimental results of this application. The paper is concluded in Section 7.

2. SYSTEM DESCRIPTION

To make the discussion more concrete, we will use discrete-time state-space models with multiple time-scales. Without loss of generality, we assume the system has two time-scales: the macro and micro time-scales. System quantities on the macro time-scale tend to vary slowly over time while system quantities on the micro time-scale exhibit fast variation over time. The former are referred to as the model parameters of the system while the latter are called the states of the system. We then begin by defining the nonlinear state-space model considered in this work as

$$\begin{aligned} \text{Transition:} \quad \mathbf{x}_{k,l+1} &= \mathbf{F}(\mathbf{x}_{k,l}, \mathbf{u}_{k,l}, \boldsymbol{\theta}_k) + \mathbf{w}_{k,l}, \\ \boldsymbol{\theta}_{k+1} &= \boldsymbol{\theta}_k + \mathbf{r}_k, \end{aligned} \quad (1)$$

$$\text{Measurement:} \quad \mathbf{y}_{k,l} = \mathbf{G}(\mathbf{x}_{k,l}, \mathbf{u}_{k,l}, \boldsymbol{\theta}_k) + \mathbf{v}_{k,l}$$

where $\mathbf{x}_{k,l}$ is the vector of system states at the time $t_{k,l} = t_{k,0} + l \cdot T$, for $1 \leq l \leq L$, with T being a fixed time step between two adjacent measurement points, and k and l being the indices of macro and micro time-scales, respectively; $\boldsymbol{\theta}_k$ is the vector of system model parameters at the time $t_{k,0}$; $\mathbf{u}_{k,l}$ is the vector of observed exogenous inputs; $\mathbf{y}_{k,l}$ is the vector of system observations (or measurements); $\mathbf{w}_{k,l}$ and \mathbf{r}_k are the

vectors of process noise for states and model parameters, respectively; $\mathbf{v}_{k,l}$ is the vectors of measurement noise; $\mathbf{F}(\bullet, \bullet, \bullet)$ and $\mathbf{G}(\bullet, \bullet, \bullet)$ are the state transition and measurement functions, respectively. Note that L represents the level of time-scale separation and that $\mathbf{x}_{k,0} = \mathbf{x}_{k-1,L}$. With the system defined, we aim at estimating both the system states \mathbf{x} and model parameters $\boldsymbol{\theta}$ from the noisy observations \mathbf{y} .

Let's take the battery system as an example. In the battery system, the system state x refers to the SOC, which changes very rapidly and may transverse the entire range 100%-0% within minutes. The system model parameter θ represents the cell capacity which tends to vary very slowly and typically decreases 1.0% or less in a month with regular use. The state transition equation $F(\bullet, \bullet, \bullet)$ models the variation of SOC over time while the cell dynamic model $G(\bullet, \bullet, \bullet)$ relates the measured cell terminal voltage y with the unmeasured state (SOC) and model parameter (capacity) and the measured exogenous input u being the cell current. Given the system's state-space model in Eq. (1) and knowledge of the measured input/output signals (cell current/cell terminal voltage), we are interested in estimating the unmeasured state (SOC) and model parameter (capacity) in real-time and in a dynamic driving environment.

3. REVIEW OF DUAL EXTENDED KALMAN FILTER METHOD

The dual extended Kalman filter (EKF) method is a commonly used technique to simultaneously estimate the states and model parameters (Haykin, 2001). The essence of the dual EKF method is to combine the state and weight EKFs with the state EKF estimating the system states and the weight EKF estimating the system model parameters. In the algorithm, two EKFs are run concurrently and, at every time step when observations are available, the state EKF estimates the states using the current model parameter estimates from the weight EKF while the weight EKF estimates the model parameters using the current state estimates from the state EKF. This section gives a brief review of the dual EKF method. Section 3.1 presents the numerical formulation of the dual EKF method and the numerical implementation of the recursive derivative computation is described in Section 3.2.

3.1 Numerical Formulation: Dual Estimation

The algorithm of the dual EKF for the system described in Eq. (1) is summarized in Table 1. Since the dual EKF does not take into account the time-scale separation, $\boldsymbol{\theta}_k$ is estimated on the micro time-scale. To reflect this, we use the notations $\boldsymbol{\theta}_{k,l}$ and $\mathbf{r}_{k,l}$ to replace $\boldsymbol{\theta}_k$ and \mathbf{r}_k , respectively. Also note that, to be consistent with the system description in Eq. (1), we use two time

indices k and l to present the dual EKF algorithm and this presentation is equivalent to a simpler version in (Wan and Nelson, 2001) with only one time index l .

The algorithm is initialized by setting the model parameters θ and states \mathbf{x} to the best guesses based on the prior information. The covariance matrices Σ_0 and Σ_x of estimation errors are also initialized based on the prior information. At each measurement time step, the time- and measurement-updates of the state and weight EKFs are performed. In the time-update, the state and parameter estimates from the previous measurement time step are propagated forward in time according to the transition equations in Eq. (1). The current state and parameter estimates are set equal to these propagated estimates and the error uncertainties are increased due to the addition of process noise \mathbf{w} and \mathbf{r} . In the measurement update, the measurement at the current time step is compared with the predicted model outputs based on the current state and parameter estimates and the differences are used to adapt the current estimates.

3.2 Numerical Implementation: Recursive Derivative Computation

The dual EKF method, which adapts the states and parameters using two concurrently running EKFs, has a recursive architecture associated with the computation of $\mathbf{C}_{k,l}^0$ in the weight filter. The computation of $\mathbf{C}_{k,l}^0$ involves a total derivative of the measurement function with respect to the parameters θ as

$$\mathbf{C}_{k,l}^0 = \left. \frac{d\mathbf{G}(\hat{\mathbf{x}}_{k,l}^-, \mathbf{u}_{k,l}, \theta)}{d\theta} \right|_{\theta=\hat{\theta}_{k,l}}. \quad (2)$$

This computation requires a recursive routine similar to a real-time recursive learning (Williams and Zipser, 1989). Decomposing the total derivative into partial derivatives and propagating the states back in time results in the recursive equations

$$\frac{d\mathbf{G}(\hat{\mathbf{x}}_{k,l}^-, \mathbf{u}_{k,l}, \theta)}{d\theta} = \frac{\partial \mathbf{G}(\hat{\mathbf{x}}_{k,l}^-, \mathbf{u}_{k,l}, \theta)}{\partial \theta} + \frac{\partial \mathbf{G}(\hat{\mathbf{x}}_{k,l}^-, \mathbf{u}_{k,l}, \theta)}{\partial \hat{\mathbf{x}}_{k,l}^-} \frac{d\hat{\mathbf{x}}_{k,l}^-}{d\theta}, \quad (3)$$

$$\frac{d\hat{\mathbf{x}}_{k,l}^-}{d\theta} = \frac{\partial \mathbf{F}(\hat{\mathbf{x}}_{k,l-1}, \mathbf{u}_{k,l-1}, \theta)}{\partial \theta} + \frac{\partial \mathbf{F}(\hat{\mathbf{x}}_{k,l-1}, \mathbf{u}_{k,l-1}, \theta)}{\partial \hat{\mathbf{x}}_{k,l-1}} \frac{d\hat{\mathbf{x}}_{k,l-1}}{d\theta}, \quad (4)$$

$$\frac{d\hat{\mathbf{x}}_{k,l-1}}{d\theta} = \frac{d\hat{\mathbf{x}}_{k,l-1}^-}{d\theta} - \mathbf{K}_{k,l-1}^x \frac{d\mathbf{G}(\hat{\mathbf{x}}_{k,l-1}^-, \mathbf{u}_{k,l-1}, \theta)}{d\theta} + \frac{\partial \mathbf{K}_{k,l-1}^x}{\partial \theta} [\mathbf{y}_{k,l-1} - \mathbf{G}(\hat{\mathbf{x}}_{k,l-1}^-, \mathbf{u}_{k,l-1}, \theta)]. \quad (5)$$

The last term in Eq. (5) can be set to zero with the assumption that $\mathbf{K}_{k,l}^x$ is not dependent on θ . Indeed, since $\mathbf{K}_{k,l}^x$ is often very weakly dependent on θ , the extra computational effort to consider this dependence is not worth the improvement in performance. Therefore, we drop the last term in Eq. (5) in this study. Then the three total derivatives can be computed in a recursive manner with initial values set as zeros. It is noted that the partial derivatives of the transition and measurement functions with respect to the states \mathbf{x} and parameters θ can be easily computed with the explicitly given function forms.

4. A MULTISCALE FRAMEWORK WITH EXTENDED KALMAN FILTER

As discussed in Section 3, the dual EKF method estimates both the states and parameters on the same time-scale. However, for systems that exhibit the time-scale separation, it is natural and desirable to adapt the slowly time-varying parameters on the macro time-scale while keeping the estimation of the fast time-varying states on the micro time-scale. This section is dedicated to the discussion of this multiscale framework. Section 4.1 presents the numerical formulation of the multiscale framework and the numerical implementation of the recursive derivative computation is described in Section 4.2.

4.1 Numerical Formulation: Multiscale Estimation

As opposed to the dual estimation, we intend to derive a multiscale estimation which allows for a time-scale separation in the state and parameter estimations. More specifically, we aim at estimating the slowly time-varying model parameters on the macro time-scale and, at the same time, intend to keep the estimation of fast time-varying states on the micro time-scale to utilize all the measurements. For these purposes, we derive the so-called micro and macro EKFs running on the micro and macro time-scales, respectively. Note that, the micro time-scale here refers to the time-scale on which system states exhibit fast variation while the macro time-scale refers to the one on which system model parameters tend to vary slowly. For example, in the battery system, the SOC, as a system state, changes every second, which suggests the micro time-scale is approximately one second. In contrast, the cell capacity, as a system model parameter, typically decreases 1.0% or less in a month with regular use, resulting in the macro time-scale being approximately one day or so. In the micro EKF, similar to the state EKF in the dual estimation, the states are estimated based on measurements \mathbf{y} . In the macro EKF, the measurements used to adapt the model parameters are the estimated states from the micro EKF.

Table 1 Algorithm of dual extended Kalman filter (Wan and Nelson, 2001)

Initialization

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{0,0} &= E[\hat{\boldsymbol{\theta}}_{0,0}], \quad \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{k,l}} = E\left[(\boldsymbol{\theta}_{0,0} - \hat{\boldsymbol{\theta}}_{0,0})(\boldsymbol{\theta}_{0,0} - \hat{\boldsymbol{\theta}}_{0,0})^T\right], \\ \hat{\mathbf{x}}_{0,0} &= E[\mathbf{x}_{0,0}], \quad \boldsymbol{\Sigma}_{\mathbf{x}_{k,l}} = E\left[(\mathbf{x}_{0,0} - \hat{\mathbf{x}}_{0,0})(\mathbf{x}_{0,0} - \hat{\mathbf{x}}_{0,0})^T\right].\end{aligned}\quad (6)$$

For $k \in \{1, \dots, \infty\}$, $l \in \{1, \dots, L\}$, compute

Time-update equations for the weight filter

$$\hat{\boldsymbol{\theta}}_{k,l}^- = \hat{\boldsymbol{\theta}}_{k,l-1}^-, \quad \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{k,l}}^- = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{k,l-1}}^- + \boldsymbol{\Sigma}_{\mathbf{r}_{k,l-1}}. \quad (7)$$

Time-update equations for the state filter

$$\hat{\mathbf{x}}_{k,l}^- = \mathbf{F}(\hat{\mathbf{x}}_{k,l-1}^-, \mathbf{u}_{k,l-1}, \hat{\boldsymbol{\theta}}_{k,l-1}^-), \quad \boldsymbol{\Sigma}_{\mathbf{x}_{k,l}}^- = \mathbf{A}_{k,l-1} \boldsymbol{\Sigma}_{\mathbf{x}_{k,l-1}}^- \mathbf{A}_{k,l-1}^T + \boldsymbol{\Sigma}_{\mathbf{w}_{k,l-1}}. \quad (8)$$

Measurement-update equations for the state filter

$$\mathbf{K}_{k,l}^{\mathbf{x}} = \boldsymbol{\Sigma}_{\mathbf{x}_{k,l}}^- (\mathbf{C}_{k,l}^{\mathbf{x}})^T \left[\mathbf{C}_{k,l}^{\mathbf{x}} \boldsymbol{\Sigma}_{\mathbf{x}_{k,l}}^- (\mathbf{C}_{k,l}^{\mathbf{x}})^T + \boldsymbol{\Sigma}_{\mathbf{v}_{k,l}} \right]^{-1}. \quad (9)$$

$$\hat{\mathbf{x}}_{k,l} = \hat{\mathbf{x}}_{k,l}^- + \mathbf{K}_{k,l}^{\mathbf{x}} \left[\mathbf{y}_{k,l} - \mathbf{G}(\hat{\mathbf{x}}_{k,l}^-, \mathbf{u}_{k,l}, \hat{\boldsymbol{\theta}}_{k,l}^-) \right], \quad \boldsymbol{\Sigma}_{\mathbf{x}_{k,l}} = (\mathbf{I} - \mathbf{K}_{k,l}^{\mathbf{x}} \mathbf{C}_{k,l}^{\mathbf{x}}) \boldsymbol{\Sigma}_{\mathbf{x}_{k,l}}^-. \quad (10)$$

Measurement-update equations for the weight filter

$$\mathbf{K}_{k,l}^{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{k,l}}^- (\mathbf{C}_{k,l}^{\boldsymbol{\theta}})^T \left[\mathbf{C}_{k,l}^{\boldsymbol{\theta}} \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{k,l}}^- (\mathbf{C}_{k,l}^{\boldsymbol{\theta}})^T + \boldsymbol{\Sigma}_{\mathbf{n}_{k,l}} \right]^{-1}. \quad (11)$$

$$\hat{\boldsymbol{\theta}}_{k,l} = \hat{\boldsymbol{\theta}}_{k,l}^- + \mathbf{K}_{k,l}^{\boldsymbol{\theta}} \left[\mathbf{y}_{k,l} - \mathbf{G}(\hat{\mathbf{x}}_{k,l}^-, \mathbf{u}_{k,l}, \hat{\boldsymbol{\theta}}_{k,l}^-) \right], \quad \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{k,l}} = (\mathbf{I} - \mathbf{K}_{k,l}^{\boldsymbol{\theta}} \mathbf{C}_{k,l}^{\boldsymbol{\theta}}) \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{k,l}}^-. \quad (12)$$

where

$$\mathbf{A}_{k,l-1} = \left. \frac{\partial \mathbf{F}(\mathbf{x}, \mathbf{u}_{k,l-1}, \hat{\boldsymbol{\theta}}_{k,l}^-)}{\partial \mathbf{x}} \right|_{\mathbf{x}=\hat{\mathbf{x}}_{k,l-1}^-}, \quad \mathbf{C}_{k,l}^{\mathbf{x}} = \left. \frac{\partial \mathbf{G}(\mathbf{x}, \mathbf{u}_{k,l}, \hat{\boldsymbol{\theta}}_{k,l}^-)}{\partial \mathbf{x}} \right|_{\mathbf{x}=\hat{\mathbf{x}}_{k,l}^-}, \quad \mathbf{C}_{k,l}^{\boldsymbol{\theta}} = \left. \frac{d\mathbf{G}(\hat{\mathbf{x}}_{k,l}^-, \mathbf{u}_{k,l}, \boldsymbol{\theta})}{d\boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{k,l}^-}. \quad (13)$$

A state projection scheme is introduced to project the state through the macro time step, expressed as

$$\mathbf{x}_{k-1,L} = \mathbf{F}_{0 \rightarrow L}(\mathbf{x}_{k-1,0}, \mathbf{u}_{k-1,0:L-1}, \boldsymbol{\theta}_{k-1}) \quad (14)$$

where the state projection function $\mathbf{F}_{0 \rightarrow L}(\bullet, \bullet, \bullet)$ can be expressed as a nested form of the state transition function $\mathbf{F}(\bullet, \bullet, \bullet)$. It is noted that the computational effort involved in computing $\mathbf{F}_{0 \rightarrow L}(\bullet, \bullet, \bullet)$ is negligible compared to the time- and measurement-updates conducted in L micro time steps.

The algorithm of the multiscale framework for the system described in Eq. (1) is summarized in Table 2. Note that, in contrast to the dual EKF algorithms in Table 1, we only use the macro time-scale index k to present the macro EKF since it estimates the parameters on the macro time-scale. The algorithm is initialized by setting the model parameters $\boldsymbol{\theta}$ and states

\mathbf{x} to the best guesses based on the prior information. The covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\mathbf{x}}$ of estimation errors are also initialized based on the prior information. At each time step on the macro time-scale, the time- and measurement-updates of the macro EKF is performed while, at each time step on the micro time-scale, the time- and measurement-updates of the micro EKF is performed. In the measurement-update of the macro EKF, the state estimate at the previous macro time step from the micro EKF is projected through the macro time step according to the state projection equation in Eq. (14). Then the state estimates at the current macro time step from the micro EKF are compared with the projected estimates and the differences are used to adapt the current parameter estimates.

Table 2 Algorithm of a multiscale framework with extended Kalman filter

Initialization

$$\begin{aligned}\hat{\boldsymbol{\theta}}_0 &= E[\hat{\boldsymbol{\theta}}_0], & \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{k,J}} &= E\left[(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_0)(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_0)^T\right], \\ \hat{\mathbf{x}}_{0,0} &= E[\mathbf{x}_{0,0}], & \boldsymbol{\Sigma}_{\mathbf{x}_{k,J}} &= E\left[(\mathbf{x}_{0,0} - \hat{\mathbf{x}}_{0,0})(\mathbf{x}_{0,0} - \hat{\mathbf{x}}_{0,0})^T\right].\end{aligned}\quad (15)$$

For $k \in \{1, \dots, \infty\}$, compute

Time-update equations for the macro EKF

$$\hat{\boldsymbol{\theta}}_k^- = \hat{\boldsymbol{\theta}}_{k-1}, \quad \boldsymbol{\Sigma}_{\boldsymbol{\theta}_k}^- = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{k-1}} + \boldsymbol{\Sigma}_{\mathbf{r}_{k-1}}. \quad (16)$$

State projection equation for the macro EKF

$$\tilde{\mathbf{x}}_{k-1,L} = \mathbf{F}_{0 \rightarrow L}(\hat{\mathbf{x}}_{k-1,0}, \mathbf{u}_{k-1,0:L-1}, \hat{\boldsymbol{\theta}}_k^-). \quad (17)$$

Measurement-update equations for the macro EKF

$$\mathbf{K}_k^0 = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_k}^- (\mathbf{C}_k^0)^T \left[\mathbf{C}_k^0 \boldsymbol{\Sigma}_{\boldsymbol{\theta}_k}^- (\mathbf{C}_k^0)^T + \boldsymbol{\Sigma}_{\mathbf{n}_k} \right]^{-1}. \quad (18)$$

$$\hat{\boldsymbol{\theta}}_k = \hat{\boldsymbol{\theta}}_k^- + \mathbf{K}_k^0 [\hat{\mathbf{x}}_{k-1,L} - \tilde{\mathbf{x}}_{k-1,L}], \quad \boldsymbol{\Sigma}_{\boldsymbol{\theta}_k} = (\mathbf{I} - \mathbf{K}_k^0 \mathbf{C}_k^0) \boldsymbol{\Sigma}_{\boldsymbol{\theta}_k}^-. \quad (19)$$

For $l \in \{1, \dots, L\}$, compute

Time-update equations for the micro EKF

$$\hat{\mathbf{x}}_{k,l}^- = \mathbf{F}(\hat{\mathbf{x}}_{k,l-1}, \mathbf{u}_{k,l-1}, \hat{\boldsymbol{\theta}}_{k-1}), \quad \boldsymbol{\Sigma}_{\mathbf{x}_{k,l}}^- = \mathbf{A}_{k,l-1} \boldsymbol{\Sigma}_{\mathbf{x}_{k,l-1}} \mathbf{A}_{k,l-1}^T + \boldsymbol{\Sigma}_{\mathbf{w}_{k,l-1}}. \quad (20)$$

Measurement-update equations for the micro EKF

$$\mathbf{K}_{k,l}^x = \boldsymbol{\Sigma}_{\mathbf{x}_{k,l}}^- (\mathbf{C}_{k,l}^x)^T \left[\mathbf{C}_{k,l}^x \boldsymbol{\Sigma}_{\mathbf{x}_{k,l}}^- (\mathbf{C}_{k,l}^x)^T + \boldsymbol{\Sigma}_{\mathbf{v}_{k,l}} \right]^{-1}. \quad (21)$$

$$\hat{\mathbf{x}}_{k,l} = \hat{\mathbf{x}}_{k,l}^- + \mathbf{K}_{k,l}^x [\mathbf{y}_{k,l} - \mathbf{G}(\hat{\mathbf{x}}_{k,l}^-, \mathbf{u}_{k,l}, \hat{\boldsymbol{\theta}}_{k-1})], \quad \boldsymbol{\Sigma}_{\mathbf{x}_{k,l}} = (\mathbf{I} - \mathbf{K}_{k,l}^x \mathbf{C}_{k,l}^x) \boldsymbol{\Sigma}_{\mathbf{x}_{k,l}}^-. \quad (22)$$

where

$$\mathbf{A}_{k,l-1} = \left. \frac{\partial \mathbf{F}(\mathbf{x}, \mathbf{u}_{k,l-1}, \hat{\boldsymbol{\theta}}_{k-1})}{\partial \mathbf{x}} \right|_{\mathbf{x}=\hat{\mathbf{x}}_{k,l-1}^-}, \quad \mathbf{C}_{k,l}^x = \left. \frac{\partial \mathbf{G}(\mathbf{x}, \mathbf{u}_{k,l}, \hat{\boldsymbol{\theta}}_{k-1})}{\partial \mathbf{x}} \right|_{\mathbf{x}=\hat{\mathbf{x}}_{k,l}^-}, \quad \mathbf{C}_k^0 = \left. \frac{d\mathbf{F}_{0 \rightarrow L}(\hat{\mathbf{x}}_{k-1,0}, \mathbf{u}_{k-1,0:L-1}, \boldsymbol{\theta})}{d\boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_k^-}. \quad (23)$$

4.2 Numerical Implementation: Recursive Derivative Computation

In the multiscale framework, the computation of \mathbf{C}_k^0 in the macro EKF involves a total derivative of the state projection function with respect to the parameters $\boldsymbol{\theta}$ as

$$\mathbf{C}_k^0 = \left. \frac{d\mathbf{F}_{0 \rightarrow L}(\hat{\mathbf{x}}_{k-1,0}, \mathbf{u}_{k-1,0:L-1}, \boldsymbol{\theta})}{d\boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_k^-}. \quad (24)$$

Similar to the total derivative in Eq. (2), this computation also requires a recursive routine.

Decomposing the total derivative into partial derivatives, we then obtain the following equation

$$\begin{aligned}\frac{d\mathbf{F}_{0 \rightarrow L}(\hat{\mathbf{x}}_{k-1,0}, \mathbf{u}_{k-1,0:L-1}, \boldsymbol{\theta})}{d\boldsymbol{\theta}} &= \frac{\partial \mathbf{F}_{0 \rightarrow L}(\hat{\mathbf{x}}_{k-1,0}, \mathbf{u}_{k-1,0:L-1}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &+ \frac{\partial \mathbf{F}_{0 \rightarrow L}(\hat{\mathbf{x}}_{k-1,0}, \mathbf{u}_{k-1,0:L-1}, \boldsymbol{\theta})}{\partial \hat{\mathbf{x}}_{k-1,0}} \frac{d\hat{\mathbf{x}}_{k-1,0}}{d\boldsymbol{\theta}}.\end{aligned}\quad (25)$$

The total derivative in the last term can be obtained by using the recursive equations Eqs. (3)-(5).

5. APPLICATION TO LI-ION BATTERY SYSTEM

In this section, we use the proposed framework to estimate the SOC and capacity in a Li-ion battery system. When applied to the battery system, the multiscale framework can be treated as a hybrid of coulomb counting and adaptive filtering techniques and comprises two new ideas: (i) a multiscale framework to estimate SOC and capacity that exhibit time-scale separation and (ii) a state projection scheme for accurate and stable capacity estimation. Section 5.1 presents the discrete-time cell dynamic model used in this study. Section 5.2 presents the multiscale estimation of SOC and capacity.

5.1 Discrete-Time Cell Dynamic Model

In order to execute the time-update in the micro and macro EKF, we need a state transition model that propagate the SOC forward in time. In order to execute the measurement-update in the micro-EKF, we need a “discrete-time cell dynamic model” that relates the SOC to the cell voltage. Here we employ the enhanced self-correcting (ESC) model which considers the effects of open circuit voltage (OCV), internal resistance, voltage time constants and hysteresis (Plett, 2004a). The effects of voltage time constants and hysteresis in the ESC model can be expressed as (Plett, 2004a)

$$\begin{bmatrix} f_{k,l+1} \\ h_{k,l+1} \end{bmatrix} = \begin{bmatrix} \text{diag}(\boldsymbol{\alpha}) & 0 \\ 0 & \varphi(i_{k,l+1}) \end{bmatrix} \begin{bmatrix} f_{k,l} \\ h_{k,l} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 - \varphi(i_{k,l+1}) \end{bmatrix} \begin{bmatrix} i_{k,l} \\ M(x, \dot{x}) \end{bmatrix} \quad (26)$$

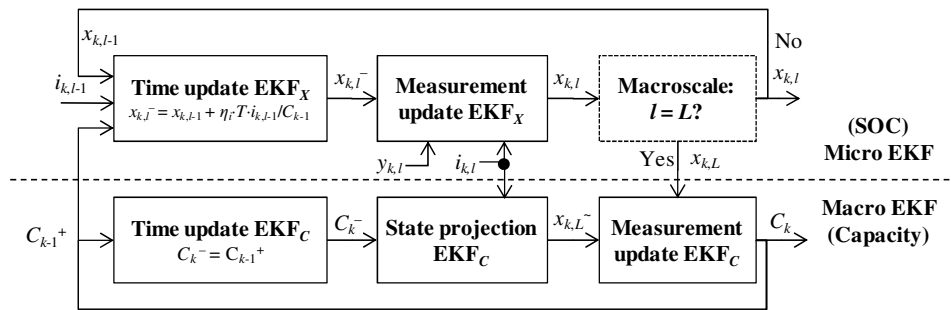


Figure 1: Flowchart of a multiscale framework with EKF for battery SOC and capacity estimation.

The framework consists of two EKFs running in parallel: the top one (micro EKF) adapting the SOC in the micro time-scale and the bottom one (macro EKF) adapting the capacity in the macro time-scale. The micro EKF sends the SOC estimate to the macro EKF and receives the capacity estimate from the macro EKF. In what follows, we intend to elaborate on the key

$$\varphi(i_{k,l+1}) = \exp\left(-\left|\frac{\eta_i \cdot i_{k,l} \cdot \gamma \cdot T}{C_k}\right|\right) \quad (27)$$

where x is the SOC, f the filter state, h the hysteresis voltage, $\boldsymbol{\alpha}$ the vector of filter pole locations, γ the hysteresis rate constant, i the current, $M(\cdot, \cdot)$ maximum hysteresis, η_i the Coulombic efficiency, T the length of measurement interval, C the nominal capacity. We then obtain the state transition and measurement equations as

$$\begin{aligned} x_{k,l+1} &= F(x_{k,l}, i_{k,l}, C_k) \\ &= x_{k,l} - \frac{\eta_i \cdot T \cdot i_{k,l}}{C_k}, \\ y_{k,l+1} &= G(x_{k,l}, i_{k,l}, C_k) \\ &= \text{OCV}(z_k) - i_{k,l} \cdot R + h_{k,l+1} + S \cdot f_{k,l+1}. \end{aligned} \quad (28)$$

where OCV is the open circuit voltage, y_k the predicted cell terminal voltage, R the cell resistance, S a vector of constants that blend the time constant states together in the output.

5.2 Multiscale Estimation of SOC and Capacity

We then begin to introduce the multiscale framework with EKF for the Li-ion battery system by drawing a flowchart in Figure 1, where T is a fixed time step between two adjacent measurement points, $x_{k,l}$ is the SOC estimate at the time $t_{k,l} = t_{k,0} + l \cdot T$, for $1 \leq l \leq L$ (k and l are the indices of macro and micro time-scales, respectively), y and i are the cell voltage and current measurements, and C is the cell capacity estimate.

technical component of the multiscale framework, the macro EKF, which consists of the following recursively executed procedures (see Figure 2):

Step 1: At the macro time step k , the capacity transition step, also known as the time update step, computes the expected capacity and its variance based

on the updated estimates at the time step $k - 1$, expressed as

$$C_k^- = C_{k-1}^+, \quad \Sigma_{C_k}^- = \Sigma_{C_{k-1}}^+ + \Sigma_{i_{k-1}}. \quad (29)$$

For a stable system, the capacity variance term tends to decrease over time with the measurement update to be detailed in the subsequent step. However, the process noise term always increases the uncertainty of the capacity estimate due to the addition of unpredictable process noise. To clearly illustrate the idea, we intend to classify the capacity estimates into three cases (see Figure 1): a larger estimate $C_{k-1}^{(L)}$, an accurate estimate $C_{k-1}^{(N)}$, and a smaller estimate $C_{k-1}^{(S)}$.

Step 2: Based on the capacity estimate C_k^- , the state projection scheme projects the SOC through the macro time step, expressed as a state projection equation derived from Eqs. (14) and (28)

$$x_{k,L} = x_{k,0} + \frac{\eta \cdot T}{C_k^-} \cdot \sum_{j=0}^{L-1} i_{k,j}. \quad (30)$$

As can be seen in Figure 2:, the projected SOC exhibits large deviations from their true value (from micro EKF), which suggests a magnified effect of the capacity on the SOC.

Step 3: Following the state projection step, the difference between the projected SOC and the estimated SOC by the micro EKF is used to update the capacity estimate, known as the measurement update. It is noted that the measurement update requires accurate SOC estimates which can be obtained from the micro EKF. The updated capacity estimate equals the predicted capacity estimate in *Step 1* plus a correction factor, expressed as

$$C_k^+ = C_k^- + K_k^C [\hat{x}_{k,L} - \tilde{x}_{k,L}], \quad (31)$$

$$\Sigma_{C_k}^+ = (1 - K_k^C C_k^C) \Sigma_{C_k}^-.$$

where the Kalman gain K_k^C and the total derivative C_k^C can be estimated using Eqs. (18) and (23), respectively.

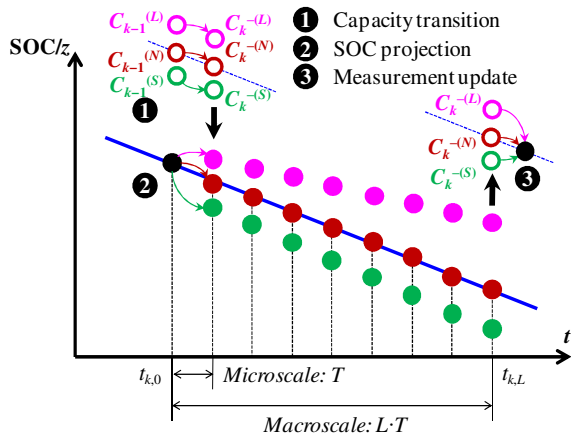


Figure 2: Procedures of capacity estimation in macro EKF.

5.3 Remarks on Multiscale Framework

We note that the proposed framework decouples the SOC and capacity estimation in terms of both the measurement and time-scale, with an aim to avoid the concurrent SOC and capacity estimation relying on the only measurement (cell terminal voltage) in the dual EKF (Plett, 2004a). In fact, the very motivation of this work lies in the fact that the coupled estimation in the dual EKF falls short in the way of achieving stable capacity estimation, precisely because it is difficult to distinguish the effects of two states (SOC and capacity) on the only measurement (cell terminal voltage), especially in the case of the micro time-scale where the capacity only has a very small influence on the SOC. Regarding the measurement decoupling, the multiscale framework uses the cell terminal voltage exclusively as the measurement for adapting the SOC (micro EKF) which in turn serves as the measurement to adapt the capacity (macro EKF). Regarding the time-scale decoupling, the state projection using the coulomb counting in Eq. (30) significantly magnifies the effect of the capacity on the SOC, i.e., that the capacity affects the SOC projected on the macro time-scale ($L \cdot T$) more significantly than that projected on the micro time-scale (T). The larger influence of the capacity on the SOC leads to the possibility of more stable capacity estimation, and that is precisely the main technical characteristic that distinguishes our approach from the dual EKF.

6. SIMULATION AND EXPERIMENTAL RESULTS

The verification of the proposed multiscale framework was accomplished by conducting an extensive urban dynamometer drive schedule (UDDS) test. In Section 6.1, the synthetic data using a valid dynamic model of a high power LiPB cell are used to verify the effectiveness of the multiscale framework. Section 6.2 reports the results of UDDS cycle life test on Li-ion prismatic cells.

6.1 SOC and Capacity Estimation with Synthetic Data of High Power Cell

Synthetic Data Generation

In order to evaluate the performance of our proposed approach, we generated the synthetic data ($T = 1s$) using an ESC model of a prototype LiPB cell with a nominal capacity of 7.5Ah (Plett, 2006b). The root-mean-square (RMS) modeling error compared to cell tests was reported to be less than 10mV (Plett, 2004b). A sequence of 15 urban dynamometer driving schedule (UDDS) cycles (see Figure 3a), separated by 30A constant current discharge and 5min rest, result in the spread of SOC over the 100%-4% range (see Figure

3b). To account for the measurement error, the current and voltage data were contaminated by zero mean Gaussian noise with standard deviations 200mA and 10mV, respectively.

Capacity Estimation Results

To test the performance of the dual EKF and the multiscale framework with EKF, we intentionally offset the initial capacity value (7.0Ah) from the true value (7.5Ah). The results of capacity estimations by these two methods are summarized in Figure 3c and 3d, respectively, from which three important observations can be made. First of all, both methods produced converged capacity estimates with identical similar convergence rate. Indeed, the convergence rate can be adjusted by varying the process and measurement noise covariances which, respectively, represent the process uncertainty resulting from the model inaccuracy and the measurement uncertainty resulting from external disturbance that corrupts the measurement data. Secondly, the dual EKF yielded inaccurate and noisy capacity estimation (see Figure 3c) while the multiscale framework ($L = 100$) with EKF produced more accurate and stable capacity estimation (see Figure 3d). This can be attributed to the fact that the state projection in Eq. (30) magnifies the effect of the capacity on the SOC as well as removes to some extent

the measurement noise. To minimize the effect of randomness in measurement noise, we repeated this simulation process ten times and obtained average RMS capacity estimation errors after convergence (at $t = 200$ mins) to be 0.048Ah (relative error 0.640%) and 0.033Ah (relative error 0.440%) for the dual EKF and the multiscale framework with EKF, respectively. Thirdly, it is observed that, although the multiscale framework with EKF produced stable capacity estimation, the estimate still exhibits small fluctuation over time. It is fair to say, however, that the small noise does not really affect the practical use of this estimate.

Computational Efficiency

In the previous subsection, we have demonstrated that the proposed multiscale framework yielded higher accuracy than the dual EKF. In this subsection, we compare the two methods in terms of computational efficiency. To minimize the effect of randomness in measurement noise, we employed the ten synthetic data sets with each being executed ten times. Our computations were carried out on a processor Intel Core i5 760 CPU 2.8GHz and 4 GByte RAM. The codes for both methods were self-devised hand-optimized MATLAB codes running in Matlab environment (MATLAB Version 7.11.0.584, The MathWorks, Inc., Natick, MA USA).

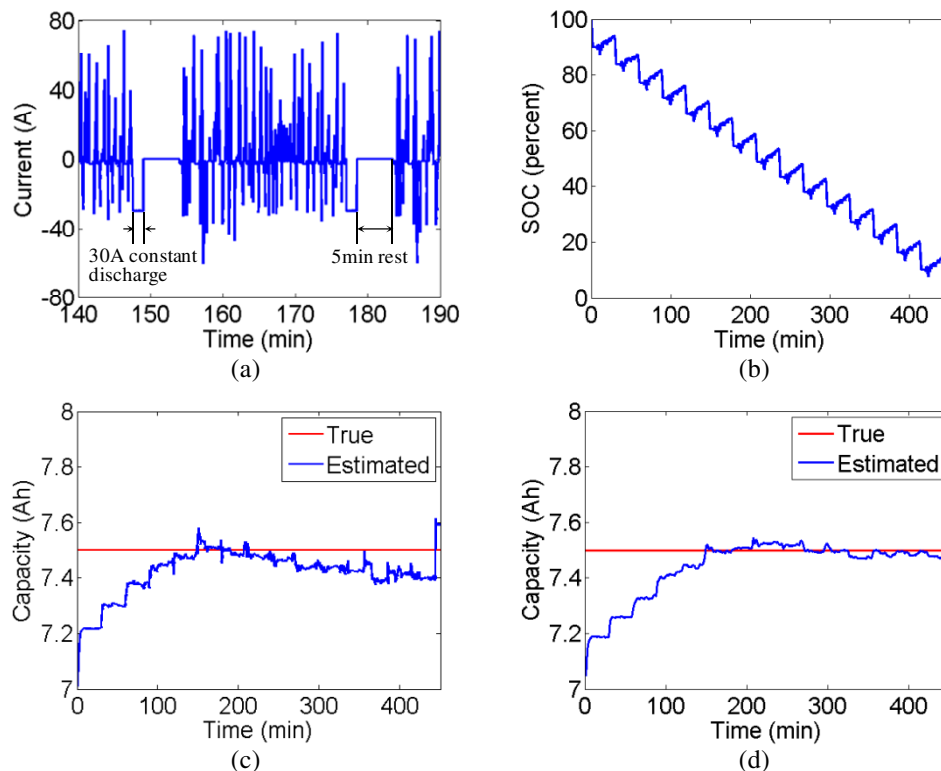


Figure 3: Synthetic data and results of capacity estimation. Figure (a) plots the rate profile for one UDDS cycle and (b) plots the SOC profile; (c) and (d) plot the results of capacity estimation by dual EKF and multiscale framework with EKF, respectively.

To make our comparison of general use to other engineering systems, we ruled out the computational time required to execute the ESC model in this study. In fact, the measurement functions of two engineered systems may exhibit a large difference in the level of complexity, resulting in different amounts of computational time. Thus, we intend to minimize the effect of system-to-system variation and focus on the general functions in an EKF by assuming a negligibly small amount of time for the execution of the system-specific measurement function (ESC model).

Table 3 summarizes the mean computational times. It is observed that the multiscale framework with EKF requires a smaller amount of computational time of 1.456s for the sequence of 15 UDDS cycles, a 34.145% reduction over the dual EKF whose computational time is 2.210s. Note that the percent of improvement is less than 50%. This can be attributed to the following two reasons: (i) from the standpoint of computations on the micro time-scale, it is noted that, in addition to the time- and measurement-update computations for SOC estimation, both methods also require the recursive derivative computation which, to some extent, reduces their efficiency gap; and (ii) from the standpoint of computations on the macro time-scale, although the macro-EKF is executed only upon the completion of $L = 100$ executions of the micro-EKF, it still requires a certain amount of time to compute the time- and measurement-updates for capacity estimation. In spite of these points, it is fair to say, however, that the proposed method achieves considerable improvement over the dual EKF in terms of computational efficiency. This improvement is critical to alleviating the

computational burden imposed on the hardware and thus enhancing the feasibility of applications.

Table 3 Comparison results of computation efficiency with ten synthetic data sets

Method	Computational time (s)	Improvement (%)
Dual EKF	2.210	---
Mutiscale EKF	1.456	34.145

6.2 SOC and Capacity Estimation with UDDS Cycle Life Test of a Prismatic Cell

Description of Test Procedure

In addition to the numerical study using synthetic data, we also conducted the UDDS cycle test to verify the effectiveness of the multiscale framework. The cycle test data were extracted from an accelerated life test (ALT) that is currently being performed on sixteen 1500-mAh Li-ion prismatic cells. We set up a UDDS test system (see Figure 4) which comprises of an MACCOR Series 4000 cycle tester with a data acquisition device, an Espec SH-241 temperature chamber and a test jig as a connector holder for prismatic cells. Sixteen prismatic cells were placed in the temperature chamber and held by the test jig throughout the test.

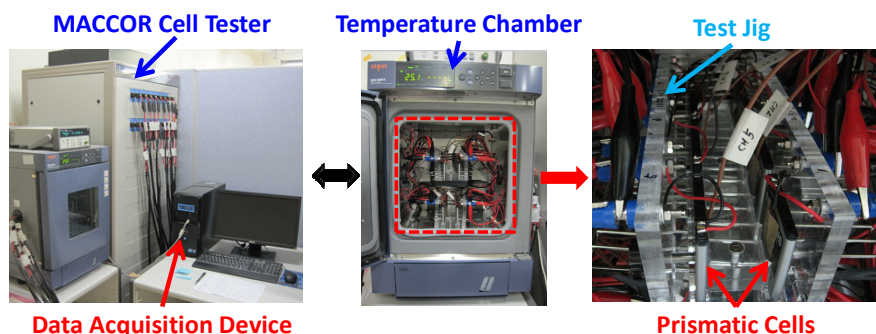


Figure 4: Experiment setup – UDDS cycle life test system.

All cycling experiments were performed at a constant room temperature, i.e., 25°C. A two-level design of experiment (DOE) was used to study the effects of charging and discharging conditions on the health degradation. With two levels for charging conditions (1.0C and 1.5C) and discharging conditions (1.0C and 2.0C), we have four experimental settings as shown in Table 4. Based on the cell degradation data

obtained from tests, we will develop real-time SOH and SOL prediction algorithms. Figure 5 shows the detailed test procedure. After every 10 charging and discharging cycles with specified rates in Table 4, cells are tested with 10 urban dynamometer drive schedule (UDDS) cycles for algorithm verification, followed by a small rate (0.05C) constant discharge for capacity check.

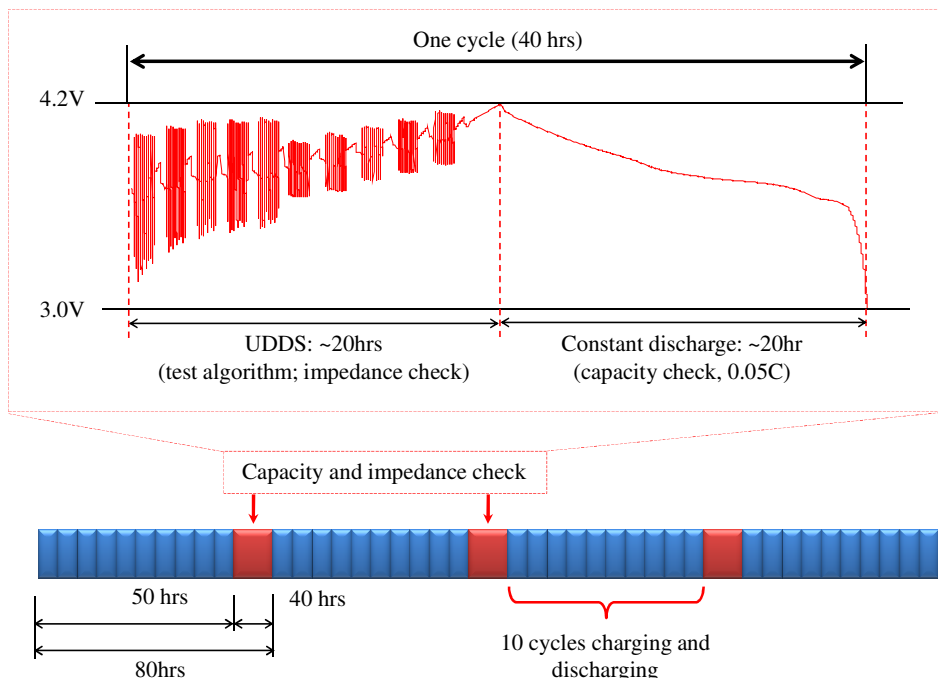


Figure 5: Detailed test procedure.

Table 4 Experiment settings

Charging Rate	Discharging Rate	Number of Cells
1.0C	1.0C	4
1.5C	1.0C	4
1.0C	2.0C	4
1.5C	2.0C	4

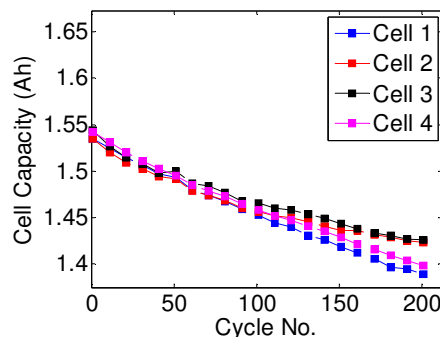
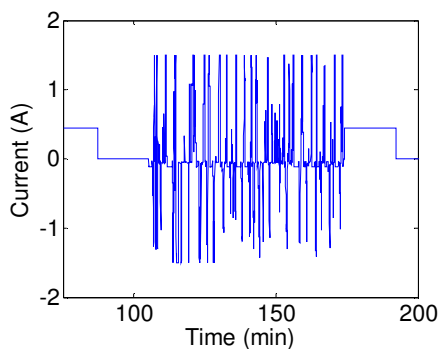


Figure 6: Capacity degradation under 1.0C charging rate and 1.0C discharging rate.

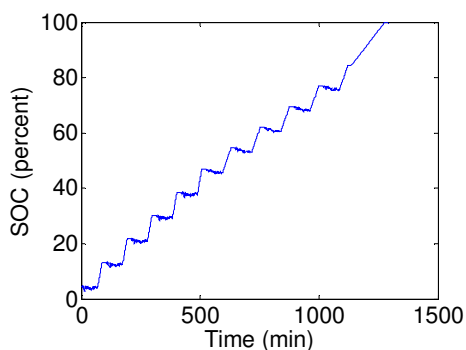
The capacity degradation of the 1500-mAh prismatic Li-ion cells under the first cycling condition in Table 3 is plotted in Figure 6. Under this condition, the cell capacity exhibits a linear relationship with the number of cycles and decreases by about 0.1Ah (6.5%) after 200 charging and discharging cycles. In what follows, we do not intend to investigate how to utilize this degradation behavior for SOL prediction but to employ the UDDS cycle test data before the cycling (1.0C charging, 1.0C discharging) from the first two cells to verify effectiveness of the proposed multiscale framework. The cycle test (see Figure 7a) is composed of 10 UDDS cycles, separated by 1C constant charge for 18 min and 18 min rest. This test profile resulted in the spread of SOC over the 4%-100% range. The SOC profile for 10 UDDS cycles is plotted in Figure 7b, where the SOC increases by about 9% during each charge period between cycles.

Training of ESC Cell Model

The current and voltage measurements of Cell 1 were used to train the ESC model (Plett, 2004a) while Cell 2 was treated as the testing cell. We followed the procedures described in (Plett, 2005) to obtain the open circuit voltage (OCV) curve. Through numerical optimization, optimum ESC model parameters were obtained which minimize the root mean squared (RMS) error of cell terminal voltage. The numerical optimization was performed using with a sequential quadratic programming (SQP) method. We employed a nominal capacity of 1.5Ah, a measurement interval of $T \approx 1s$, and four filter states $n_f = 4$. The voltage modeling results for one UDDS cycle are shown in Figure 8, where a good agreement can be observed.



(a)



(b)

Figure 7: SOC profile and one cycle rate profile for UDDS cycle test. Figure (a) plots the rate profile for one UDDS cycle and (b) plots the SOC profile for 10 UDDS cycles.

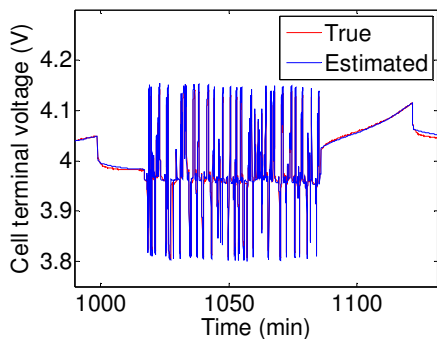


Figure 8: Modeled and measured cell terminal voltage for one UDDS cycle.

SOC and Capacity Estimation Results

The SOC estimation results for the training cell are shown in Figure 9, where we observe accurate SOC estimation produced by the multiscale framework ($L = 1200$). Table 5 summarizes the SOC estimation errors under two different settings of the initial SOC. Here, the RMS and maximum errors take into account the initial offset in the case of an incorrect initial SOC and are formulated as

$$\begin{aligned} \mathcal{E}_{RMS} &= \sqrt{\frac{1}{nm} \sum_{k,j} (\hat{x}_{k,l} - x_{k,l})^2}, \\ \mathcal{E}_{Max} &= \max_{k,j} |\hat{x}_{k,l} - x_{k,l}|. \end{aligned} \tag{32}$$

where nm is the number of measurements and reads 69,173 (about 1290mins) in this study; and $x_{k,l}$ is the true SOC at the time $t_{k,l}$ estimated with the coulomb counting technique. It is observed that the RMS SOC estimation errors produced by the multiscale framework are less than 4.00%, regardless of initial values of the SOC. As expected, the SOC estimation with incorrect initial SOC (20%) shows larger errors than those with correct initial SOC (4.84% and 4.77% for Cells 1 and 2, respectively). However, the RMS SOC estimation errors with incorrect initial SOC (20%) are still less than 4.00% since the multiscale framework produced converged SOC estimate for both cases.

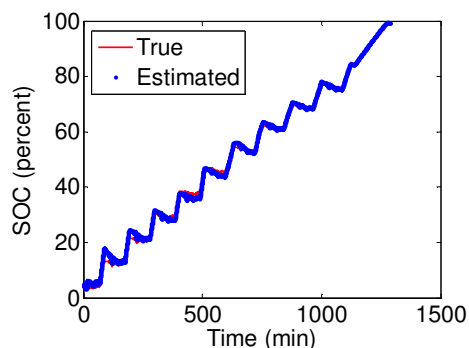


Figure 9: Estimated and true SOC estimate for 10 UDDS cycles.

Table 5 SOC estimation results under different settings of initial SOC and capacity

Initial SOC	SOC errors	Cell 1	Cell 2
Correct (4.84% and 4.77% for Cells 1 and 2)	RMS (%)	1.21	1.22
	Max (%)	4.58	4.95
Incorrect (20%)	RMS (%)	3.79	3.65
	Max (%)	15.16	15.23

Regarding the capacity estimation, both results with initial values smaller than the true value (see Figure10a) and larger than the real value (see Figure10b) for all the two cells exhibit convergence to the true capacity within an error range of around 5%. The noise in the capacity estimate is due to the SOC estimation error. We note that the time-scale separation in the SOC and capacity estimation enables converged capacity estimation in spite of SOC estimation error.

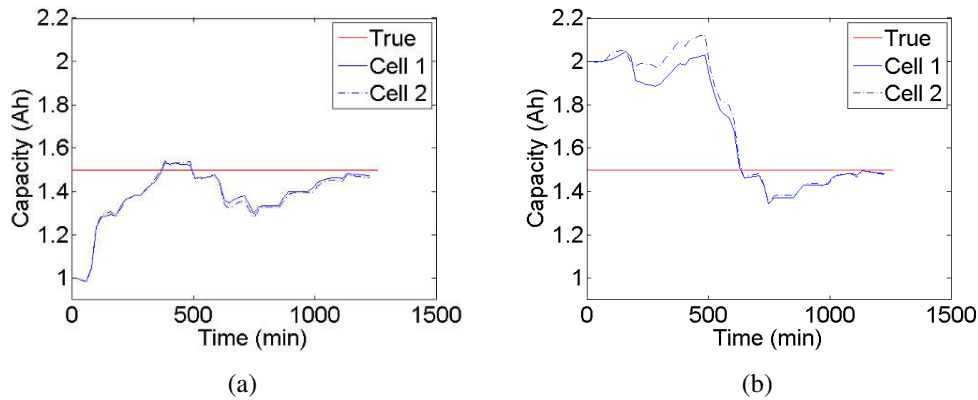


Figure 10: Capacity estimation results for UDDS cycle test. Figures (a) and (b) plot capacity estimation results by the multiscale framework with the initial values smaller than and larger than the true value, respectively.

7. CONCLUSION

This paper presents a multiscale framework with EKF to efficiently and accurately estimate state and parameter for engineered systems that exhibit time-scale separation. We applied the proposed framework applied to the Li-ion battery system for SOC (state) and capacity (parameter) estimation. The main contribution of this paper lies in the decoupling of the SOC and capacity estimation from two perspectives, namely the measurement and time-scale, through the construction of a multiscale computational scheme. The resulting benefits are the significant reduction of the computational time as well as the increase of the accuracy in the capacity estimation. The former benefit makes the proposed methodology more attractive than the dual EKF for onboard estimation devices where the computational efficiency is the key aspect for practical use. Results from UDDS simulation and testing verify the effectiveness of the proposed framework for SOC and capacity estimation. As mentioned in Section 6.2, we are currently conducting ALTs (cell aging tests) on 16 Li-ion prismatic batteries. Based on the upcoming testing results, we aim to extend the proposed multiscale framework for efficient and accurate SOL prediction based on readily available measurements in a dynamic environment (e.g., UDDS cycling).

ACKNOWLEDGMENT

The authors gratefully acknowledge PCTEST Engineering Laboratory Inc. for providing testing facilities and Prof. Gregory L. Plett for providing the UDDS profile for this research.

NOMENCLATURE

C	cell capacity
F	state transition function
G	state measurement function
i	current
L	number of micro steps in a macro time step
\mathbf{r}	vector of process noise for model parameters
T	time between micro time step
x	cell state of charge
y	cell terminal voltage
\mathbf{u}	vector of observed exogenous inputs
\mathbf{v}	vector of measurement noise
\mathbf{w}	vectors of process noise for states
η	columbic efficiency
EKF	extended Kalman filter
HEV	hybrid electric vehicle
SOC	state of charge
SOH	state of health
SOL	state of life

REFERENCES

- Plett, G.L. (2004a). Extended Kalman filtering for battery management systems of LiPB-based HEV battery packs Part 3. State and parameter estimation, *Journal of Power Sources*, vol. 134, no. 2, pp. 277–292.
- Plett, G.L. (2006a). Sigma-point Kalman filtering for battery management systems of LiPB-based HEV battery packs Part 2: Simultaneous state and parameter estimation, *Journal of Power Sources*, vol. 161, no. 2, pp. 1369–1384.
- Lee, S., Kim, J., Lee, J. & Cho, B.H. (2008). State-of-charge and capacity estimation of lithium-ion battery using a new open-circuit voltage versus state-of-charge, *Journal of Power Sources*, vol. 185, no. 2, pp. 1367–1373.
- Zhang, Q. & White, R.E. (2008a). Capacity fade analysis of a lithium ion cell, *Journal of Power Sources*, vol. 179, no. 2, pp. 793–798.
- Zhang, Q. & White, R.E. (2008b). Calendar life study of Li-ion pouch cells Part 2: Simulation, *Journal of Power Sources*, vol. 179, no. 2, pp. 785–792.
- Saha, B., Goebel, K., Poll, S. & Christophersen J. (2009). Prognostics methods for battery health monitoring using a Bayesian framework, *IEEE Transaction on Instrumentation and Measurement*, vol. 58, no. 2, pp. 291–296.
- Saha, B. & Goebel, K. (2009). Modeling Li-ion battery capacity depletion in a particle filtering framework, *In Proceedings of Annual Conference of the PHM Society*, San Diego, CA, Sep. 27-Oct. 1.
- Plett, G.L. (2004b). Extended Kalman filtering for battery management systems of LiPB-based HEV battery packs Part 2. Modeling and identification, *Journal of Power Sources*, vol. 134, no. 2, pp. 262–276.
- Plett, G.L. (2006b). Sigma-point Kalman filtering for battery management systems of LiPB-based HEV battery packs Part 1: Introduction and state estimation, *Journal of Power Sources*, vol. 161, no. 2, pp. 1356–1368.
- Haykin, S. (2001). *Kalman Filtering and Neural Networks*, Wiley/Inter-Science, New York.
- Wan, E. & Nelson, A. (2001). *Dual extended Kalman filter methods*, in: Haykin S. (Ed.), *Kalman Filtering and Neural Networks*, Wiley/Inter-Science, New York, p123–174.
- Williams, R.J. & Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks, *Neural Computation*, vol. 1, no. 2, pp. 270–280.
- Plett, G. (2005). Results of temperature-dependent LiPB cell modeling for HEV SOC estimation, *In Proceedings of the 21st Electric Vehicle Symposium (EVS21)*, Monaco, April 2-6.
- Chao Hu:** Mr. Hu received his B.E. degree in Engineering Physics from Tsinghua University (Beijing, China) in 2003. He is currently pursuing the Ph.D. degree in mechanical engineering at The University of Maryland, College Park (Maryland, USA). His research interests are system reliability analysis, prognostics and health management (PHM), and battery power and health management of Li-ion battery system.
- Byeng D. Youn:** Dr. Byeng D. Youn is currently an Assistant Professor in the School of Mechanical and Aerospace Engineering at Seoul National University in South Korea. Dr. Youn is dedicated to well-balanced experimental and simulation studies of system analysis and design and is currently exploring three research avenues: (1) system risk-based design, (2) prognostics and health management (PHM), and (3) energy harvester design. Dr. Youn's research and educational portfolio includes: (i) *six notable awards*, including the ISSMO/Springer Prize for the Best Young Scientist in 2005 from the International Society of Structural and Multidisciplinary Optimization (ISSMO), (ii) *over one hundred publications* in the area of system risk assessment and design and PHM. His primary applications include Li-ion batteries, consumer electronics, and large-scale engineered systems (e.g., automobiles).
- Taejin Kim:** Mr. Kim received his B.E. degree in the School of Mechanical and Aerospace Engineering at Seoul National University (Seoul, Korea) in 2011. He is currently pursuing the M.S. degree in mechanical engineering at Seoul National University (Seoul, Korea). His research interests are prognostics and health management (PHM) for smart plant, and battery thermal and health management of Li-ion battery system.
- Jaesik Chung:** Dr. Chung, CTO of PCTEST, has managed the battery safety and reliability laboratory in PCTEST since 2007 in the business area of battery test, certification, safety project and R&D project relate to the battery safety and reliability. Before he joined PCTEST, he had worked for nineteen year in the area of electrochemical systems and Li-ion battery safety and reliability. He has managed a battery pack R&D team for his last eight years and developed smart battery algorithms and battery packs for mobile phones, notebook PCs, power tools, and mobility products and performed cell and battery pack safety designs for Apple, Dell, HP, IBM, Nokia, Motorola, Sony and other system makers.

Optimization of fatigue maintenance strategies based on prognosis results

Yibing Xiang and Yongming Liu

Department of Civil & Environment Engineering, Clarkson University, Potsdam, NY 13699-5710, USA
xiangyi@clarkson.edu
yliu@clarkson.edu

ABSTRACT

A general approach to determine the optimal set of maintenance alternatives for fatigue safety is introduced in this paper. The optimal maintenance alternatives are the solutions to maximize the fatigue reliability of aircrafts fleet subject to maintenance budget. A novel equivalent stress transformation model and the first-order-reliability method (FORM) are adopted to determine the failure probability or reliability associated with future fatigue loading. The equivalent stress transformation model is capable to transform future random loading to an equivalent constant loading, and does not require cycle-by-cycle simulation. First-order-reliability-method can resolve the computational complexity. Optimal maintenance solution can be efficiently found considering the future fatigue loading. Numerical examples are performed to demonstrate the application of the proposed approach.

1 INTRODUCTION

Most structures and components, e.g. aircrafts and rotorcrafts, are experiencing cyclic loading throughout their service life. These cyclic loading results in many failure modes, and fatigue failure is one of most common failure modes. There is an increasing interest to enhance the durability, reliability and safety of the structures with limit budget. Scheduling of inspection and repair activities can effectively mitigate the fatigue detrimental effects (Y. Garbatov & C. Guedes Soares, 2001) (D. Straub & M. H. Faber, 2005).

To obtain a reasonable future maintenance plan, first of all, very good diagnostic techniques are required. There exist several non-destructive inspection (NDI) techniques, e.g. shearography (Y. Y. Hung, 1996), thermography (M. Koruk & M. Kilic, 2009), ultrasonics (R. Kazys & L. Svilainis, 1997), X-ray CT (G. Nicoletto, G. Anzelotti & R. Konecny) and so on. Furthermore, structures experience different loading spectrums during entire fatigue life. The applied fatigue

Yibing Xiang et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

cyclic loading (S. Pommier, 2003) is stochastic in nature. It is well-known that different loading sequences may induce different load-interaction effects (S. Mikheevskiy & G. Glinka), such as the overload retardation effect and underload acceleration effect. Due to the complicated and nonlinear nature of random loading interaction, a cycle-by-cycle simulation is generally required for each different loading history. Hence this approach is computationally expensive for fatigue safety optimization, which usually requires a large number of Monte Carlo simulations.

Prediction will provide valuable information for decision making in prognostics and health management (PHM). The most difficult part is how to accurately and effectively estimate the future health status of aircraft fleet. This estimation should be built on an efficient fatigue damage prognosis procedure. A novel equivalent stress transformation (Y. Xiang & Y. Liu, 2010) and reliability method have been adopted to reduce the complexity of fatigue damage prognosis. This equivalent stress transformation is using the statistical description of the random loading, such as the probabilistic distribution of applied stress range and stress ratio. The future variable amplitude loading problem is reduced to an equivalent constant amplitude problem, which greatly facilitates the integration for crack length prediction. The FORM have been developed and used for the reliability-based design optimization problem (A. Der Kiureghian, Y. Zhang & C.-C. Li, 1994).

This paper is organized as follows. First, basic problem for optimal maintenance alternatives will be formulated, and some key parts will be pointed out. Following this, the equivalent stress transformation is briefly discussed. After that, the first-order-reliability method will be introduced. Numerical example is used to demonstrate the application of the proposed method. Parametric study has been performed to investigate the effects of some important parameters. Finally, some conclusions and future work are given based on the current investigation.

2 Problem formulation

It is well-known that structures experience fatigue cyclic loading during their service life. Crack may

propagate until parts of some components fail. The structures may break suddenly in a few cycles, or survive for a long period of time. Hence, difference exists in fatigue duration due to the uncertainties. An appropriate fatigue maintenance plan is required to optimize the condition status.

First of all, very good diagnostic techniques are required to detect the current damage stage. Several advanced diagnostic techniques are available. The current diagnostic results are regarded as the baselines for future fatigue damage prognosis. This paper mainly focuses on the prognosis techniques, and diagnostic techniques are beyond the scope of this paper.

The future loading is a critical problem in fatigue maintenance alternatives optimization. The fatigue loading is usually stochastic in nature, and the loading sequence effects are big challenges in fatigue prognosis. Traditional fatigue prognosis models focus on different explanation of crack growth mechanism, and require cycle-by-cycle simulation. These models require a large number of Monte-Carlo simulation, and is computational expensive for fatigue maintenance optimization. An equivalent stress transformation (Y. Xiang & Y. Liu, 2010) has been proposed based on the statistical description of the random loading. The variable amplitude loading problem is reduced to an equivalent constant amplitude problem. Detailed derivation and explanation will be discussed in Section 3.

Apparently, fatigue prognosis will provide valuable information for decision making in PHM. Maintenance optimization under uncertainty can be formulated as a reliability problem. Therefore, some of the developed algorithms can be applied (e.g., FORM, subset simulation, etc.) In the current study, First-order-reliability-method will be applied to find the fatigue reliability of structures. Comprehensive derivation will be discussed in Section 4.

Fatigue maintenance problem can be formulated in different ways, e.g., minimizing the total cost subjected to reliability constraints and performance constraints. This kind of problem is quite common in real engineering application, since the best condition stage of structures are desired with least cost. There is another way to formulate the problem, such like maximizing the performance reliability subject to budget constraints (e.g., annual budget for maintenance is fixed). Basically the budget is limited and the desirable condition stage of structures is required. This paper is mainly focusing on fatigue performance maximization.

The first step in the fatigue performance optimization is to define several categories depending on the crack length. For example, the fatigue performance can be divided into six stages, excellent condition, very good

condition, good condition, fair condition, poor condition and very poor condition (or failure condition).

Secondly, diagnostic methods are used to determine the fatigue damage in the current stage. The performance transition matrix can be formulated using some existing fatigue prognosis models (Equivalent stress level mode) and diagnostic results.

Thirdly, a maintenance decision matrix should be defined to specify the maintenance method for each performance category. Then the cost function can be calculated associated with each category using different maintenance alternatives.

At last, the maximization of the performance under the budget constraints can be formulated. This maintenance optimization under uncertainty can be formulated as a reliability problem. Some of the developed algorithms can be applied (e.g., FORM, subset simulation, etc.)

In the maintenance optimization problem, some variables need to be clarified:

G = number of facility groups Group of aircrafts

T = number of missions in the planning horizon

Q_g = total quantity of facilities in group g

S = number of performance condition states;

M_g = number of possible maintenance alternatives for facilities in group g

C_{gm} = cost vector ($s \times 1$) of group g and maintenance alternative m

$D_{gt} = [d_{gt}^1, d_{gt}^2, d_{gt}^3, d_{gt}^4, \dots, d_{gt}^S]$, condition vector of group g at beginning of mission t , each term represents the percentage. d_{gt}^i is the element on the diagonal of an $S \times S$ matrix.

D_{total} = the summation of the condition elements for G groups from the condition 1 to condition S , after T missions

$X_{gmt} = [x_{gmt}^1, x_{gmt}^2, x_{gmt}^3, x_{gmt}^4, \dots, x_{gmt}^S]$, x_{gmt}^s is the maintenance decision matrix, percentage of facilities in group g and condition state s that had maintenance m in year t .

P_{gm} = transition probability matrix ($S \times S$) of group g when the maintenance m is implemented (from model prediction or existing database)

$$P_{gm} = \begin{bmatrix} P_{gm}^{1,1} & P_{gm}^{1,2} & \dots & P_{gm}^{1,S} \\ P_{gm}^{2,1} & P_{gm}^{2,2} & \dots & P_{gm}^{2,S} \\ \dots & \dots & \dots & \dots \\ P_{gm}^{S,1} & P_{gm}^{S,2} & \dots & P_{gm}^{S,S} \end{bmatrix}$$

The condition of facilities from group g at year t can be predicted using previous information.

$$D_{gt} = \sum_{m=1}^{m=M_g} X_{gm(t-1)} D_{g(t-1)} P_{gm} \quad (1)$$

The total cost function can be formulated as:

$$Cost = Q_g \sum_{m=1}^{m=Mg} X_{gmt} D_{gt} C_{gm} \quad (2)$$

From above derivation, the maximization problem can be easily built as:

$$D_{total} = \sum_{s=1}^{s=S} S \sum_{g=1}^{g=G} \sum_{t=1}^{t=T} d_{gt} \quad (3)$$

For a certain mission, the budget $Budget_i$ is limited after each year i , and may be different from one year to another. The total budget $Budget_{total}$ during year t is also limited. And the budget constraints can be built as Eq. (4) and Eq. (5):

$$\sum_{g=1}^{g=G} Q_g \sum_{m=1}^{m=Mg} X_{gmt} D_{gt} C_{gm} \leq Budget_t \quad (4)$$

$$\sum_{t=1}^{t=T} \sum_{g=1}^{g=G} Q_g \sum_{m=1}^{m=Mg} X_{gmt} D_{gt} C_{gm} \leq Budget_{total} \quad (5)$$

For some cases, the reliability constraints are required. For example, the percentage of facilities in condition s should be less than a value R^s . The reliability constraints can be formulated as Eq. (6) :

$$d_{gt}^s \leq R^s \quad (6)$$

Following the above procedures, the fatigue maintenance problem can be easily formulated. However, there are some problems existing: first, the transition probability matrix reliability of each future mission is complex problem, due to measurement uncertainties (NDI testing) modeling uncertainties. The future loading dominates the transition probability matrix. The Equivalent stress transformation is proposed for the future loading. First order reliability method (FORM) can be used to calculate the probability transformation matrix

3 Equivalent stress level

Fatigue cyclic loading is a random process in nature. Proper inclusion of loading interaction effects is a big challenge, and is very important for future mission reliability. Traditional models focus on different explanation of fatigue crack growth mechanism, and require cycle-by-cycle simulation. Therefore, a large number of simulations is required and is time-consuming.

Equivalent stress transformation model has been proposed transformation (Y. Xiang & Y. Liu, 2010). This objective is to transform a random loading to an equivalent constant loading, which does not require a cycle-by-cycle simulation and can facilitate the integration. The basic idea of the equivalent stress level can be shown as below:

For an arbitrary random future loading, the statistics of stress range and stress ratio can be obtained. After a

series of calculation, the random loading can be transformed to an equivalent constant loading, which can be directly used for fatigue damage prognosis. Because this transformation is not the focus of this study, only the brief idea is illustrated. The details of the derivation and validation of the equivalent stress transformation can be found in the referred paper (Y. Xiang & Y. Liu, 2010).

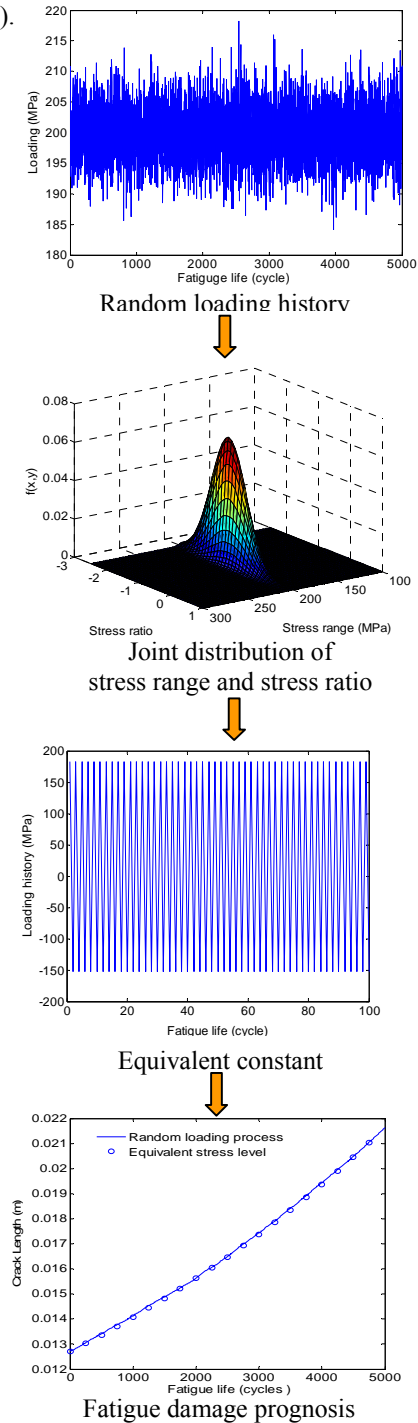


Figure. 1 basic principle of equivalent stress level

There are several different fatigue crack growth models, such as Forman's model (N. E. Dowling, 2007), Nasgro model, and EIFS-based fatigue crack growth model (Y. Liu & S. Mahadevan, 2009b). Different models focus on different aspects and will give different predictions. A generic function of crack growth rate curve can be expressed as

$$da/dN = f(\Delta\sigma, R, a) \quad (7)$$

Eq. (7) can be reformulated as

$$dN = \frac{1}{f(\Delta\sigma, R, a)} da \quad (8)$$

The total fatigue life N under arbitrary random loading history is the summation of N_i and can be written as

$$N_{total} = \sum_{i=0}^n N_i = \sum_{i=0}^n \int_{a_i}^{a_{i+1}} \frac{1}{f(\Delta\sigma_i, R_i, a)} da \quad (9)$$

where a_0 is the initial crack size and a_n is crack length at fatigue cycle N .

In this ideal crack growth process, the stress level is constant and is the proposed equivalent stress level (ESL). The equivalent stress level can be expressed as

$$N_{total} = \int_{a_0}^{a_{n+1}} \frac{1}{f(\Delta\sigma_{eq}, R_{eq}, a)} da \quad (10)$$

The equivalent stress level can be obtained by equating Eq. (9) and Eq. (10) as

$$\int_{a_0}^{a_{n+1}} \frac{1}{f(\Delta\sigma_{eq}, R_{eq}, a)} da = \sum_{i=0}^n \int_{a_i}^{a_{i+1}} \frac{1}{f(\Delta\sigma_i, R_i, a)} da \quad (11)$$

Eq. (12) is the proposed equivalent stress level calculation and it can be applied to different types of crack growth models. For any arbitrary functions of $f()$. The analytical solution is not apparent and discussions of some special cases are given below.

In the current study, the simple Paris' equation is used as the $f()$. For a general case where both of stress range and stress ratio are random variables, a joint distribution of them is required for the derivation. The general equivalent stress can be expressed as

$$\begin{aligned} \Delta\sigma_{eq} &= \left(\sum_1^n \frac{N_i}{N_{total}} \frac{g(R_i)}{g(0)} \Delta\sigma_i^m \right)^{\frac{1}{m}} \\ &= \left(\sum_1^n p_i(R_i, \Delta\sigma_i) \frac{g(R_i)}{g(0)} \Delta\sigma_i^m \right)^{\frac{1}{m}} \end{aligned} \quad (12)$$

where $p_i(R_i, \Delta\sigma_i)$ is the joint distribution of stress range and stress ratio. $g()$ is a function of stress ratio. Eq. (12) is the generalized equivalent stress level expression without considering the loading interaction effect.

The above discussion did not consider the load interaction. It is well known that the "memory" effect exists for fatigue crack growth and coupling effect has to be considered. In this section, the previous developed equivalent stress model is extended to include the load interaction effect, such as the overload

retardation and underload acceleration. The modification is based on a recently developed small time scale formulation of fatigue crack growth and a load interaction correction function. The details of the small time scale model has been developed by Lu and Liu (Z. Lu & Y. Liu). This method is based on the incremental crack growth at any time instant within a cycle, and is different from the classical reversal-based fatigue analysis.

The equivalent stress level consider load interaction effect is defined as

$$\Delta\sigma_{eq}^* = \eta \Delta\sigma_{eq} \quad (13)$$

where $\Delta\sigma_{eq}^*$ is the equivalent stress level considering the load interaction effect and $\Delta\sigma_{eq}$ is calculated using Eq. (12) without considering the load interaction term. η is the coefficient for the load interaction effect and the details of derivation can be found in (Y. Xiang & Y. Liu, 2010).

4 FORM methodology

The first-order reliability method is a widely used numerical technique to calculate the reliability or failure probability of various engineering problems (J. Cheng & Q. S. Li, 2009; S. Thorndahl & P. Willems, 2008; D. V. Val, M. G. Stewart & R. E. Melchers, 1998). Unlike the FORM method (A. Haldar & S. Mahadevan, 2000; Y. Liu, Mahadevan, S, 2009), the inverse FORM method tries to solve the unknown parameters under a specified reliability or failure probability level, which is more suitable for probabilistic life prediction (i.e., remaining life estimation corresponding to a target reliability level).

Limit state function is required for the analytical reliability method. A generic limit state function is expressed as Eq. (14a) as a function of two sets of variables x and y . x is the random variable vector and represents material properties, loadings, and environmental factors, etc. y is the index variable vector, e.g., time and spatial coordinates. The limit state function is defined in the standard normal space in Eq. (14a). The limit state function definition is similar to the classical FORM method (A. Haldar & S. Mahadevan, 2000). The solution for the unknown parameters needs to satisfy the reliability constraints, which are described in Eq. 14b) and Eq. (14c). β is the reliability index, which is defined as the distance from origin to the most probable point (MPP) in the standard normal space. The failure probability P_f can be calculated using the cumulative distribution function (CDF) Φ of the standard Gaussian distribution. Numerical search is required to find the optimum solution, which satisfies the limit state function (Eq. (14d)). Details of the general FORM method and

concept can be found in (A. Der Kiureghian et al., 1994).

$$\begin{cases} (a) : g(x, y) = 0 \\ (b) : \|x\| = \beta_{target} \\ (c) : p_f = \Phi(-\beta_{target}) \\ (d) : \begin{cases} (1) x + \frac{\|x\|}{\|\nabla_x g(x, y)\|} \nabla_x g(x, y) = 0 \quad (P_f < 50\%) \\ (2) x - \frac{\|x\|}{\|\nabla_x g(x, y)\|} \nabla_x g(x, y) = 0 \quad (P_f \geq 50\%) \end{cases} \end{cases} \quad (14)$$

The overall objective of the FORM method is to find a non-negative function satisfying all constraint conditions specified in Eq. (14). Thus, the numerical search algorithm can be used to find the solutions of the unknown parameters. Numerical search algorithm is developed to iteratively solve the Eq. (14). The search algorithm is expressed as Eq. (15) after k iterations.

$$\begin{cases} X_{k+1} \\ y_{k+1} \end{cases} = \begin{cases} X_k \\ y_k \end{cases} + f_k = \begin{cases} X_k \\ y_k \end{cases} + (a_1 f^1_k + a_2 f^2_k) \quad (15)$$

where f^1_k and f^2_k are the search directions corresponding to different merit functions.

The convergence criterion for the numerical search algorithm is

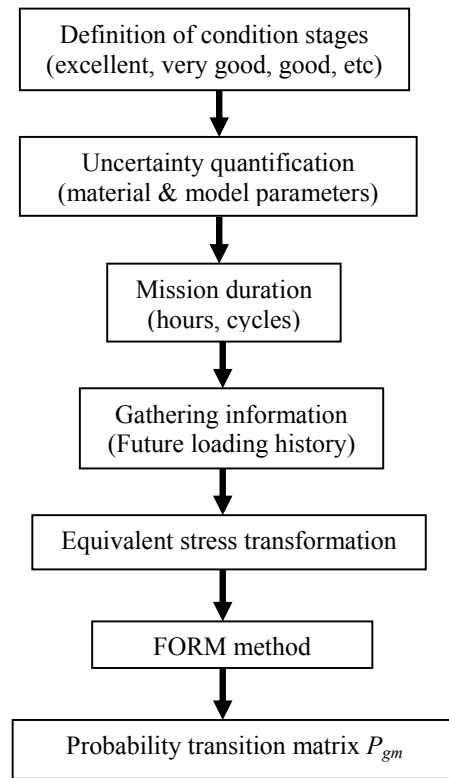
$$\frac{(\|x_{k+1} - x_k\|^2 + \|y_{k+1} - y_k\|^2)^{\frac{1}{2}}}{(\|x_{k+1}\|^2 + \|y_{k+1}\|^2)^{\frac{1}{2}}} \leq \varepsilon \quad (16)$$

where ε is a small value and indicates that the relative difference between two numerical solutions is small enough to ensure the convergence.

5 Transition probability matrix

Transition probability matrix is used to determine the future condition stage, based on the current observed fatigue damage. Calculation of the transition probability matrix P_{gm} is the key point in the fatigue safety optimization.

The general procedures to calculate P_{gm} is shown in flowchar.1. The first step is to define the condition stages, such as excellent, very good, good, etc. Following this, quantify the uncertainties in the current fatigue problem. Then obtain the information about future loading history, for example, the joint distribution of stress range and stress ratio. After equivalent stress transformation, the obtained equivalent constant loading can be directly used. The mission duration (cycles or hours) can be obtained. This information is the input data to the FORM method. The probability transition matrix can be directly calculated using FORM method.



Flowchart. 1 General procedure to calculate P_{gm}

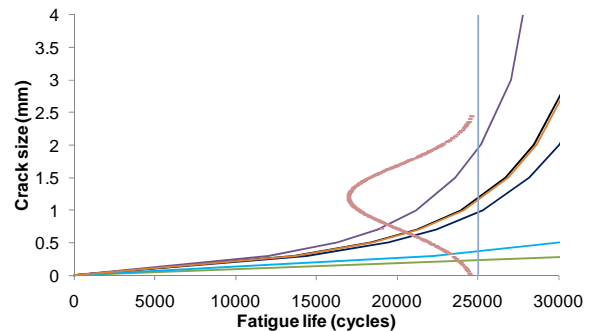


Figure. 2 Fatigue crack growth prognosis

Table 1. Statistics of random variables

Material	stress ratio	parameter	Mean	Std.
Al 7075-T6	R = -1	Mc	1.64E-10	3.86E-11
		Mm	2.3398	0.3122
		ai	0.05mm	0.006mm

Huge uncertainties exist in the fatigue damage prognosis model, e.g., the model properties C , m and the initial crack size a_i . To determine the future condition status, these three parameters are random variables and are assumed to follow log-normal

distribution. With an initial fatigue damage (a_i) around 0.05mm, after some mission (e.g. 25000 cycles), the probability can be calculated from excellent stage to the other stages. The statistics of these three random variables are shown in Table 1. Suppose there exists a constant loading history with $S_{max}=150$ MPa, $S_{min}=15$ MPa. With different combinations of the three random variables, unlimited fatigue crack growth curves can be drawn, as shown in Fig. 2.

It can be easily observed that, after a certain mission (25000 cycles) the cracks reach different conditions: very few remain the same level, and most of them increase between 0.5~2 mm. The transition probability matrix from initial condition stage (around 0.05 mm) to other condition stages can be easily obtained. The above discussion is a simple case. In this case, there are only three random variables and a constant amplitude loading history. A numerical example has been discussed for more general cases in Section 6.

6 Numerical example and parametric study

A numerical example is demonstrated in this section. In this example, there are 10, 9 and 9 aircrafts in three different groups *A*, *B* and *C* respectively. The total number of future mission is 10.

Firstly, the condition stages are defined into 6 stages: excellent (crack<0.05), very good (0.5<crack<0.6), good (0.6<crack<0.8), fair(0.8<crack<1.2), poor (1.2<crack<1.5) and very poor (crack>1.5). Three maintenance alternatives are available: do nothing, repair method I, repair method II. The cost of maintenance alternatives are shown as:

<i>conditionstate</i>	1	2	3	4	5	6
<i>do_nothing</i>	0	0	0	0	0	0
<i>repair_method_I</i>	0	400	600	800	1600	1800
<i>repair_method_II</i>	0	0	1600	1600	3000	3000

At the initial stage, the initial condition stage D_{gt} should be defined. In this numerical example, the initial condition stage is randomly generated and just for demonstration. In really engineering cases, the condition stage needs to be defined with help of advanced diagnostic techniques.

[0.1638 0.1896 0.1900 0.1627 0.1090 0.1849;
0.2618 0.1326 0.2039 0.0498 0.2056 0.1462;
0.0293 0.3242 0.3518 0.0086 0.2609 0.0251;]

Secondly, the uncertainties are quantified. The material used in the structure is Al-7075 and the random variables are calibrated from experimental fatigue crack growth data shown in Fig. 3. A summary of the properties for the collected experimental data are listed in Table 2& 3.

Thirdly, the mission duration needs to be clarified. Normally, the mission duration of a flight is about

10,000 fatigue cycles, which is used in the current study.

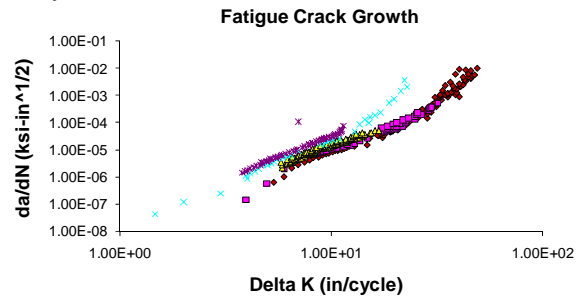


Figure. 3 Fatigue crack growth for Al-7075 under different stress ratios

Table 2 Stochastic coefficient of *a* and fatigue limit

Material	stress ratio	parameter	mean	std.
Al 7075-T6	0.03	M_C	7.72E-10	1.82E-10
		K_c	50	5
	0.05	M_C	7.96E-10	1.88E-10
		K_c	50	5

Table 3 Geometry and material properties of plate specimens

Specimen material	7075-T6
Ultimate strength σ_u (MPa)	575
Yield strength σ_y (MPa)	520
Modulus of elasticity <i>E</i> (MPa)	69600
Plate width (mm)	305
Plate thickness (mm)	4.1

The fourth step is gathering information about the future loading. Two blocks loading spectrum are used as the future loading in this numerical example. A schematic illustration of the f loading is shown in Fig. 4. *p* and *n* in Fig. 4 controls the number of cycles at the high amplitude (400MPa) and the low amplitude (250 MPa), respectively. $p=10$ and $n=50$ in the current study.

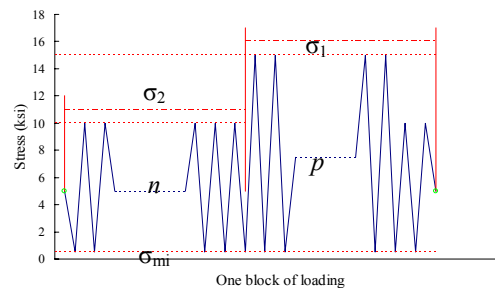


Figure. 4 Schematic illustration of the two blocks loading

It is assumed that, after some repair, the current condition stages (crack size) can be partially or fully changed to an ideal station. In another word, the fatigue crack size may follow a bi-normal distribution, for example:

$$a_i \sim A \times \log N(0.05, 0.006) + B \times \log N(0.25, 0.03) \quad (18)$$

A, B are two parameters. For different repair method I and repair method II, A and B take different value as shown in table 4.

Table 4 Model parameters in a bi-normal distribution

	Repair I	Repair II
A	0.7	0.3
B	0.9	0.1

The distribution of crack size after repair can be easily calculated using above information.

After the equivalent stress transformation, the above information can be inputted into FORM method. For example, the elements on the first row of P_{gm}^1 can be calculated by setting up the indexing vectors as condition limits in each condition stage in FORM method. The transition probability is shown below for three different maintenance alternatives:

$$P_{gm}^1 = \begin{bmatrix} 0.9956 & 0.0043 & 0.0001 & 0 & 0 & 0 \\ 0 & 0.8818 & 0.1065 & 0.0112 & 0.0004 & 0.0001 \\ 0 & 0 & 0.9246 & 0.0307 & 0.0253 & 0.0194 \\ 0 & 0 & 0 & 0.9161 & 0.0532 & 0.0307 \\ 0 & 0 & 0 & 0 & 0.9238 & 0.0762 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$P_{gm}^2 = \begin{bmatrix} 0.8325 & 0.0875 & 0.0373 & 0.0344 & 0.0073 & 0.0010 \\ 0.8325 & 0.0875 & 0.0373 & 0.0344 & 0.0073 & 0.0010 \\ 0.8325 & 0.0875 & 0.0373 & 0.0344 & 0.0073 & 0.0010 \\ 0.8325 & 0.0875 & 0.0373 & 0.0344 & 0.0073 & 0.0010 \\ 0.8325 & 0.0875 & 0.0373 & 0.0344 & 0.0073 & 0.0010 \\ 0.8325 & 0.0875 & 0.0373 & 0.0344 & 0.0073 & 0.0010 \end{bmatrix}$$

$$P_{gm}^3 = \begin{bmatrix} 0.9406 & 0.0370 & 0.0114 & 0.0099 & 0.0010 & 0.0001 \\ 0.9406 & 0.0370 & 0.0114 & 0.0099 & 0.0010 & 0.0001 \\ 0.9406 & 0.0370 & 0.0114 & 0.0099 & 0.0010 & 0.0001 \\ 0.9406 & 0.0370 & 0.0114 & 0.0099 & 0.0010 & 0.0001 \\ 0.9406 & 0.0370 & 0.0114 & 0.0099 & 0.0010 & 0.0001 \\ 0.9406 & 0.0370 & 0.0114 & 0.0099 & 0.0010 & 0.0001 \end{bmatrix}$$

The budget constraints in each mission are shown as below:

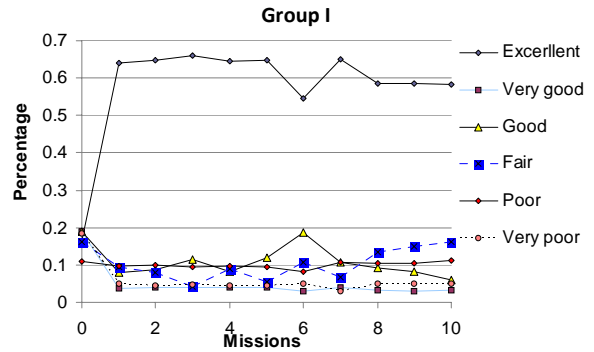
Budget=[10000 8000 9000 12000 10000 8000 9000 8000 8000 9000];

The total budget = \$65000.

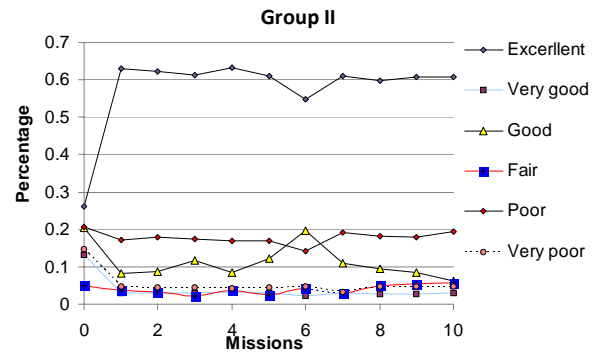
The reliability constraint is built as: the percentage of aircrafts in very poor condition is no more than 5%.

The fatigue maintenance problem is to optimize the maintenance design (X_{gmt}) to maximize the condition state, and satisfy the budget limits in each year, the

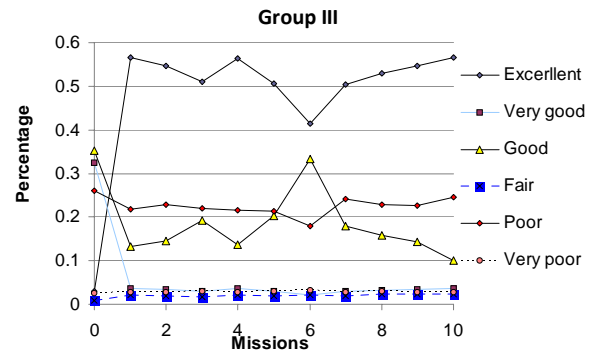
total budget for whole mission process, as well as the reliability constraints.



(a)



(b)



(c)

Figure. 5 optimal results after each mission

The optimal results are shown in Fig. 5 for three different groups. At the very beginning, only about 16% aircrafts are in excellent condition. To maximize the total condition, more money should spend to repair as many aircrafts as possible, subjected to the first year budget. It can be easily observed that more than 60% aircrafts are in excellent condition after the first mission. And those in excellent condition remain at very high level throughout the 10 missions. Those aircrafts in very poor condition takes less than 5%. In

another word, the money spent on very poor condition does not change much. The real cost and maintenance budget limit for each mission is shown in Fig. 6. The cost at each mission is less than the budget limit and satisfies the budget constraint.

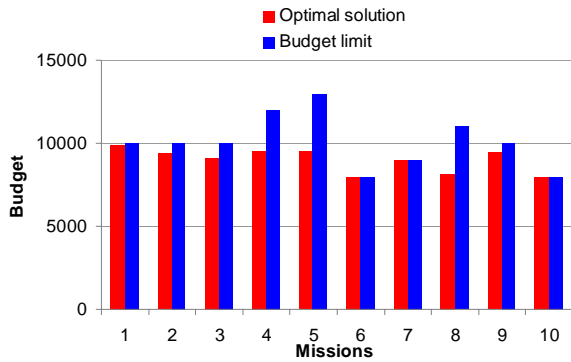


Figure. 6 Cost Vs Budget for each mission

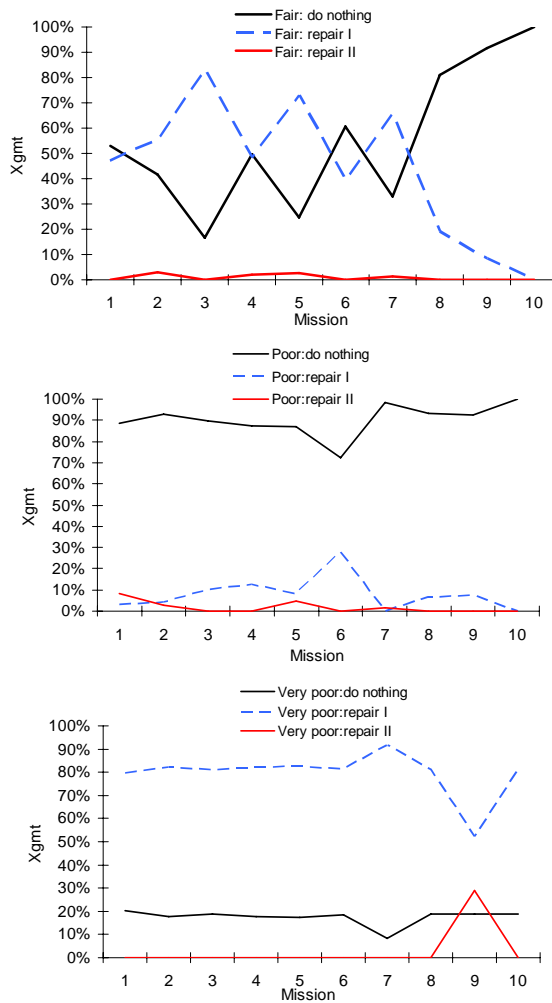
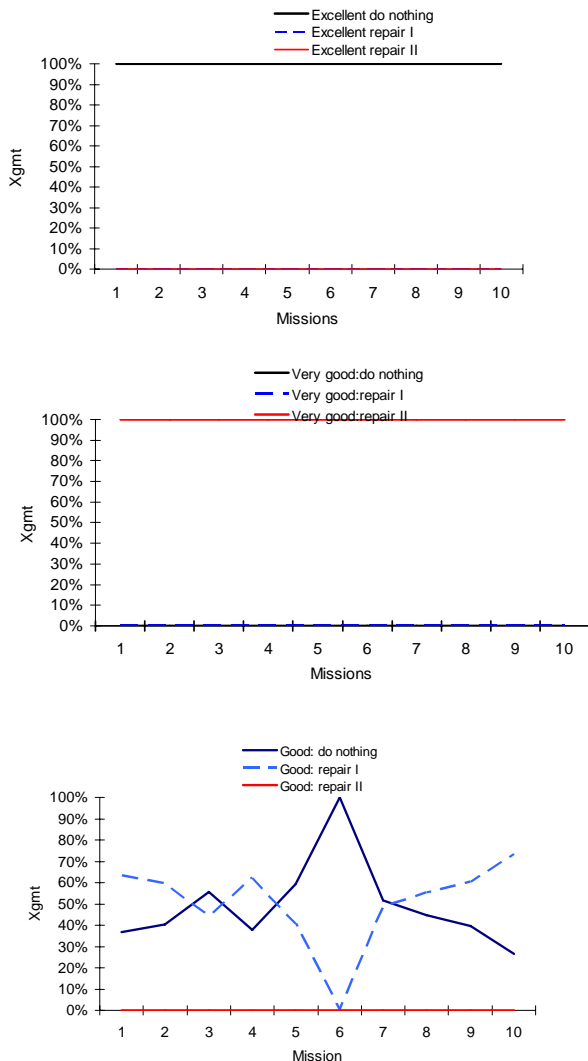


Figure. 7 Optimal solutions for maintenance alternatives

The optimal solution for maintenance alternatives are displayed in Fig. 7. For the excellent condition, no maintenance is required, which is reasonable. All the aircrafts in very good condition should take repair method II. For good and fair conditions, the aircrafts takes different combinations of maintenance alternatives. The best choice for those in poor condition is do nothing, but for those in very poor condition, repair method I is absolutely necessary.

Parametric study has been done to investigate effects caused by the variance of parameter C. In this case, the variance of parameter C takes four different values, 0.05, 0.1, 0.3, 0.5. From Fig. 8, it can be concluded that, as the variance increases, the total maximum condition value decrease. The best choice to maintain the maximum condition is to reduce the uncertainties materials properties.

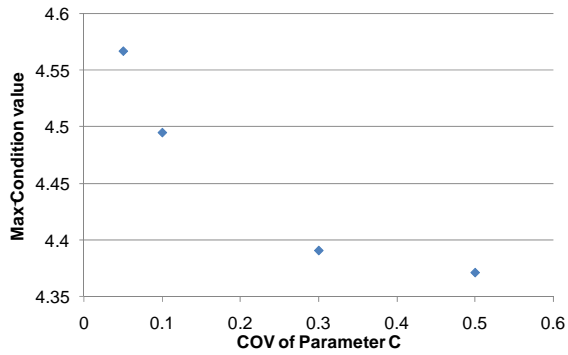


Figure. 8 Effects of the variance of parameter C

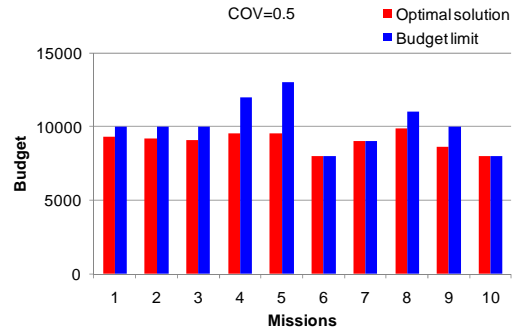
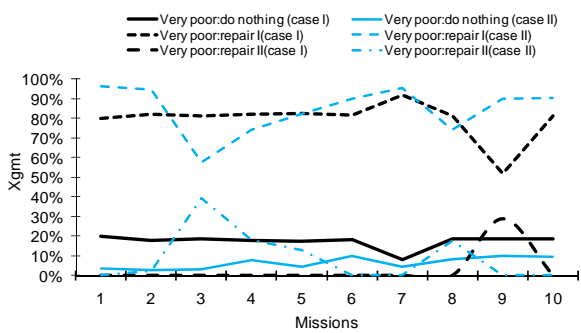
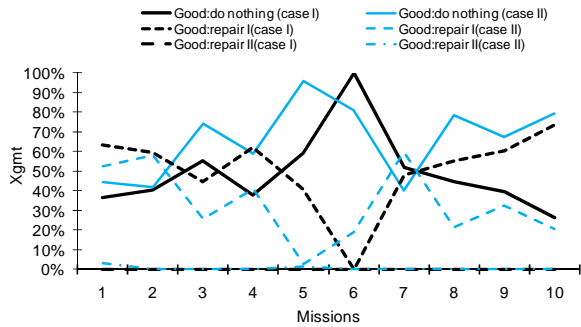
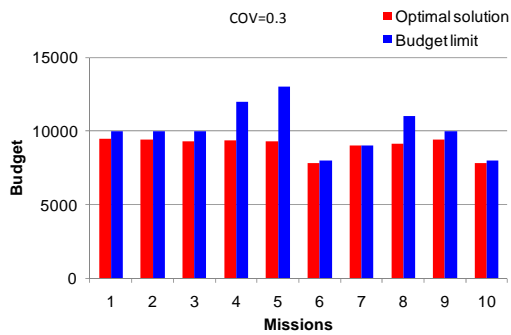
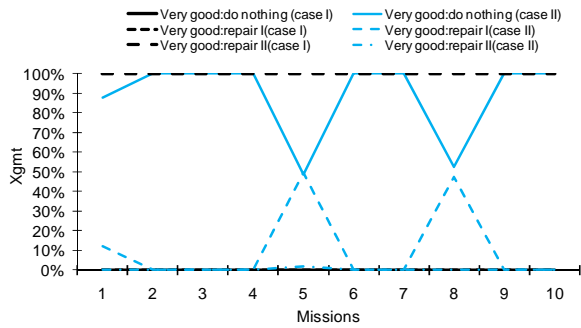
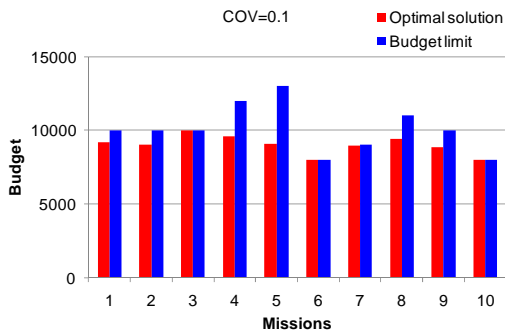
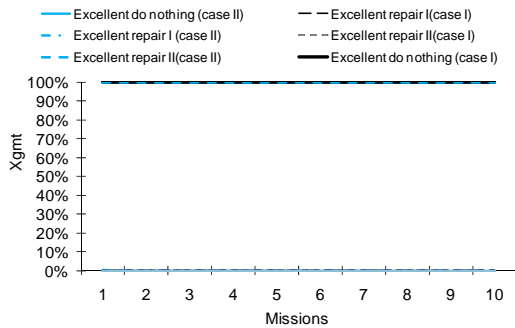
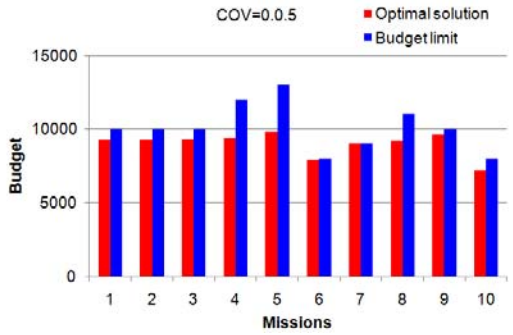


Figure. 9 Cost Vs Budget for different variance of parameter C



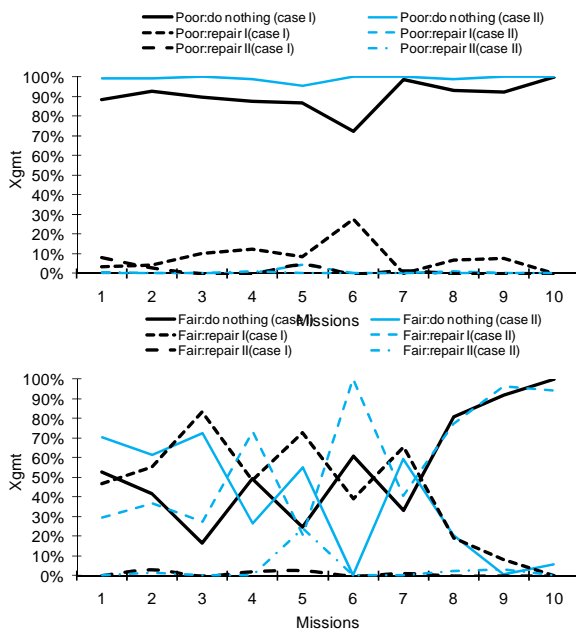


Figure. 10 Optimal solutions for maintenance alternatives

Fig. 9 shows the cost vs budget for each different variance of parameter C . No big difference can be observed. Fig. 10 displays the optimal maintenance alternatives for two cases, variance equaling to 0.05 and 0.5. Slightly difference can be observed for these two cases.

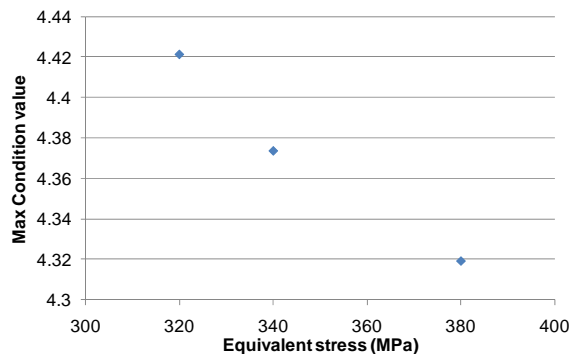


Figure. 11 Effect of equivalent stress level

The other parametric study investigates the effect of equivalent stress level. The maximum condition value decreases steadily with increase of equivalent stress level. This phenomenon is almost the same as expectation.

7 Conclusion

In this paper, a maintenance optimization framework using prognosis results is formulated. The proposed approach is based on a novel prognostic model. This

prognostic model is the combination of equivalent stress transformation and the FORM method. It is able to deal with the uncertainties in future loading. Optimization problem has been formulated based on the performance maximization under budget constraints and reliability constraint. An example with three group of facilities are considered. Parametric study has been done to investigate the effects of parameter C as well as the equivalent stress level. The results meet the expectation. Reliability constraints and other uncertainty effects are being investigated in the future study. More complicated case study is required.

ACKNOWLEDGMENT

The research reported in this paper was supported in part by the NASA ARMD/AvSP IVHM project under NRA NNX09AY54A. The support is gratefully acknowledged.

REFERENCES

- Cheng, J., & Li, Q. S. (2009) Reliability analysis of a long span steel arch bridge against wind-induced stability failure during construction. *Journal of Constructional Steel Research*, 65, 552-558.
- Cizelj, L., Mavko, B., & Riesch-Oppermann, H. (1994) Application of first and second order reliability methods in the safety assessment of cracked steam generator tubing. *Nuclear Engineering and Design*, 147, 359-368.
- Der Kiureghian, A., Zhang, Y., & Li, C.-C. (1994) Inverse reliability problem. *Journal of Engineering Mechanics, ASCE*, 120(5).
- Dowling, N. E. (2007). *Mechanical behavior of materials : engineering methods for deformation, fracture and fatigue*. Upper Saddle River, NJ, London: Pearson Prentice Hall ; Pearson Education.
- Garbatov, Y., & Guedes Soares, C. (2001) Cost and reliability based strategies for fatigue maintenance planning of floating structures. *Reliability Engineering & System Safety*, 73, 293-301.
- Haldar, A., & Mahadevan, S. (2000). *Probability, reliability, and statistical methods in engineering design*. New York ; Chichester [England]: John Wiley.
- Hung, Y. Y. (1996) Shearography for non-destructive evaluation of composite structures. *Optics and Lasers in Engineering*, 24, 161-182.
- Kam, T. Y., Chu, K. H., & Tsai, S. Y. (1998) Fatigue reliability evaluation for composite laminates via a direct numerical integration technique. *International Journal of Solids and Structures*, 35, 1411-1423.
- Kazys, R., & Svilainis, L. (1997) Ultrasonic detection and characterization of delaminations in thin composite plates using signal processing techniques. *Ultrasonics*,

35, 367-383.

Koruk, M., & Kilic, M. (2009) The usage of IR thermography for the temperature measurements inside an automobile cabin. *International Communications in Heat and Mass Transfer*, 36, 872-877.

Liao, M., Xu, X., & Yang, Q.-X. (1995) Cumulative fatigue damage dynamic interference statistical model. *International Journal of Fatigue*, 17, 559-566.

Liu, Y., & Mahadevan, S. (2007) Stochastic fatigue damage modeling under variable amplitude loading. *International Journal of Fatigue*, 29, 1149-1161.

Liu, Y., & Mahadevan, S. (2009a) Efficient methods for time-dependent fatigue reliability analysis. *AIAA Journal*, 47, 494-504.

Liu, Y., & Mahadevan, S. (2009b) Probabilistic fatigue life prediction using an equivalent initial flaw size distribution. *International Journal of Fatigue*, 31, 476-487.

Liu, Y., Mahadevan, S (2009) Efficient methods for time-dependent fatigue reliability analysis. *AIAA Journal*, 47, 494-504.

Lu, Z., & Liu, Y. Small time scale fatigue crack growth analysis. *International Journal of Fatigue*, 32, 1306-1321.

Mikheevskiy, S., & Glinka, G. Elastic-plastic fatigue crack growth analysis under variable amplitude loading spectra. *International Journal of Fatigue*, 31, 1828-1836.

Nicoletto, G., Anzelotti, G., & Konecn, R. X-ray computed tomography vs. metallography for pore sizing and fatigue of cast Al-alloys. *Procedia Engineering*, 2, 547-554.

Pommier, S. (2003) Cyclic plasticity and variable amplitude fatigue. *International Journal of Fatigue*, 25, 983-997.

Rackwitz, R. a. F., B (1978) Structural Reliability Under Combined Random Load Sequences. *Computers & Structures*, 9, 484-494.

Rackwitz, R. a. F., B (June 1976) Note on Discrete Safety Checking When Using Non-Normal Stochastic Models for Basic Variables. *Load Project Working Session, MIT, Cambridge, MA*.

Skaggs, T. H., & Barry, D. A. (1996) Assessing uncertainty in subsurface solute transport: efficient first-order reliability methods. *Environmental Software*, 11, 179-184.

Straub, D., & Faber, M. H. (2005) Risk based inspection planning for structural systems. *Structural Safety*, 27, 335-355.

Thorndahl, S., & Willems, P. (2008) Probabilistic modelling of overflow, surcharge and flooding in urban drainage using the first-order reliability method and parameterization of local rain series. *Water Research*, 42, 455-466.

Val, D. V., Stewart, M. G., & Melchers, R. E. (1998) Effect of reinforcement corrosion on reliability of

highway bridges. *Engineering Structures*, 20, 1010-1019.

Xiang, Y., & Liu, Y. (2010) Efficient probabilistic methods for real - time fatigue damage prognosis. PHM 2010, Portland.

Xiang, Y., & Liu, Y. (2010 (accepted)) Inverse first-order reliability method for probabilistic fatigue life prediction of composite laminates under multiaxial loading. *ASCE Journal of Aerospace Engineering*.

Yibing Xiang: a graduate research assistant in department of civil engineering at Clarkson University. He received his B.S. degree in civil Engineering from Tongji University in China in 2003, and then he worked as a structural analyst in Shanghai Xiandai Arch Design Company. Since 2007, he started his study at Clarkson University and got his M.S. degree in 2009, and continued his PHD degree. His research interests are probabilistic prognosis, reliability analysis, and system reliability.

Yongming Liu: an assistant Professor in the department of civil and environmental engineering. His research interests include fatigue and fracture analysis of metals and composite materials, probabilistic methods, computational mechanics, and risk management. He completed his PhD at Vanderbilt University, and obtained his Bachelors' and Masters' degrees from Tongji University in China. Dr. Liu is a member of ASCE and AIAA and serves on several technical committees on probabilistic methods and advanced materials.

Physics Based Prognostic Health Management for Thermal Barrier Coating System

Amar Kumar¹, Bhavaye Saxena², Alka Srivastava¹, Alok Goel³

¹Tecsis Corporation, 210 Colonnade Road, Ottawa, ON, K2E 7L5
amar, alka@tecsis.ca

²University of Ottawa, 800 King Edward Street, Ottawa, ON, K1N 6N5
bsaxe011@uottawa.ca

³OMTEC Inc., 170 Bristol Road East, Mississauga, ON, L4Z 3V3
goals@sympatico.ca

ABSTRACT

Reliable prognostic of thermal barrier coating systems (TBCs) as applied to hot section engine components is a challenging task. Physics based approach is made here involving both experimental physical damage signature analysis and thermal cycle simulations. Thermally grown oxides (TGO) and the developing cracks in TBCs increase with thermal exposures. An exponential relationship is observed between the two parameters. Significant variations in size and characteristics of the damage signatures are observed depending on the four typical cycle profiles considered. In this paper, fourth order Runge-Kutta method is used for the numerical analysis of the differential equation for TGO growth analysis. Damage tolerance approach considering fracture mechanics based stress intensity factor is used to determine the crack tolerance level and remaining useful life. Our earlier fracture mechanical model for composite TBCs is modified assuming the crack to nucleate and grow within the TBC and not inside TGO. An overview of the PHM solution is presented.

1. INTRODUCTION

Combined effects of high temperature and operational stress causes accelerated damage of monolithic hot-section

* This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

parts in aeroengine and drastically shortens the useful life as compared to the life of cold section components (Chin, 2005). Common metallurgical failure mechanisms of the monolithic alloys include low and high cycle fatigue, creep/rupture, oxidation, foreign object damage and corrosion (Wood, 2000; Christodoulou & Larsen, 2005). Thermal barrier coating (TBC) systems are now widely used for aero-propulsion and power generation. TBCs applied to hot section monolithic parts provide thermal insulation to gas turbine and other high temperature engine components. By lowering the temperature of the metallic substrates improves the life and performance of the components subjected to creep, fatigue, environmental attack and thermal fatigue (Shillington & Clarke, 1999; Evans et.al., 2001). A TBC system is a two layered coating consisting of 8% yttria stabilized zirconia (YSZ), and a bond coat (BC) enriched in aluminium over a Ni base superalloy substrate. During the operation, a layer of oxidation product known as thermally grown oxide (TGO- α Al₂O₃) form and grow with time in between YSZ (top coat) and bond coat (BC) layer under the influence of mechanical and thermal stress cycles. A TBC system is truly a composite structure with TBC as the insulating top layer, TGO provides the oxidation protection, BC provides adherence of TBC on superalloy while the alloy supports the structural load. Though the current state of development of TBCs meets most of the industrial needs, yet further enhancement of stability, durability and performance of TBCs providing thermal insulation to high temperature for aero-propulsion hot-section components is the pressing industrial needs.

A great deal efforts have been made over last decade to understand the damage and failure mechanisms of TBC that

form the basis of the development of physics model, which in turn is the backbone of PHM (Shillington & Clarke, 1999; Evans et al., 2001; Kumar et al., 2007; Karlsson, Xu, & Evans, 2002; Chen et al., 2005; Clarke, Levi, & Evans, 2006). The predominant failure mechanisms for TBCs include microcrack nucleation at the TBC/BC interface layer and coalescence and propagation of cracks to cause buckling / delamination over a large part and finally to spalling. Both tensile and compressive stresses of large magnitude (up to 6 GPa) develop due to growth and thermal expansion misfit of TGO. Crack formation is facilitated by the presence of small geometric imperfections at the interface regions. Such micro/sub size defects are expected to be present / formed in coat layers due to foreign object damage, processing of coatings and thermo-mechanical operations under aggressive environments. However, none of the proposed models could explain the wide scatter in the coat life and failure mechanism of TBC system. Some of the major factors contributing to the failure and scatter in TGO life include morphology, types, oxidation rate, surface treatments, alloy and phase, bond coat roughness, TGO thickness etc. (Christodoulou & Larsen, 2005; Shillington & Clarke, 1999; Kumar et al., 2007; He, Hutchinson & Evans, 2003). A critical thickness of TGO results in critical stress to cause crack initiation inside TGO and at the interface layer.

Traditional engine health management relies on the tracking of operating hours and cycles, material properties including fatigue behavior and worst case usage data. The safe life consideration ensures component safety by limiting the probable damage that can accumulate in the material long before failure indications arise. On the other hand, prognostic health management (PHM) approach maximizes the useful life of components with enhanced safety and reliability than the conventional time based engine maintenance approach. Prognostic is the ability to assess the current condition of an engine part and predict into the future for a fixed time horizon or predict the time to failure. Continuous monitoring and analysis of engine health data and usage parameters are integrated with physics based models for diagnostic and prognostics capabilities. Other advantages of the PHM solutions are increased mission availability, minimizing maintenance and life cycle cost. A good number of commercial diagnostic PHM (DPHM) systems are now available for industrial usage and being deployed for structural health monitoring and life assessment (Intellistart⁺, Altair avionics/Pratt & Whitney, SignalProTM, Impact technologies, NormNetPHM, Frontier Technology). While the technology is nearing maturization, still the products lack in certain aspects. Major limitations include accuracy in diagnostic outcome, reliability by reducing false signal, specific applicability, offer no probabilistic confidence level and uncertainty.

The PHM technology developed so far are for monolithic systems and not applicable for TBCs which

degrade differently during engine cycle operation. The primary objectives of the present work are to address the development of models and methodology appropriate for the PHM solutions for TBCs. As opposed to safe life approach, a damage tolerant (DT) approach that accounts for crack growth is assumed to be more appropriate for the TBC applications. The DT method recognizes the fact that materials and manufacturing defects whatever minute in size it may be are present in components. The material must have high fracture resistance to be damage tolerant, even with growing cracks during operation. A fracture mechanistic (FM) approach employing stress intensity factor as the crack driving parameter forms the basis for DT approach.

2. TBC DAMAGE MODEL

Earlier studies have focused on the progressive structural damage in TBCs leading to failure by cracking and spallation (Shillington & Clarke, 1999; Evans et al., 2001; Chen et al., 2005; Clarke, Levi & Evans, 2006; LeMieux, 2003; Kumar et al., 2010; Kumar, 2009). Analytical, experimental and simulation studies have identified the formation and growth of TGO leads to damage in TBCs. TGO is essentially a thermodynamic and diffusion controlled phenomenon and is a function of both time and temperature. Both of these factors were varied in our simulated experiments in order to allow appreciable TGO growth and failure. Figure 1 displays the microstructures illustrating the formation and growth of the TGO layer in YSZ TBC subjected to thermal cycling. Details of simulated experiment, thermal cycling and damage quantification are reported elsewhere (Kumar et al., 2010; Kumar et al., 2009).

2.1 Damage Signatures

Two physical damage signatures, namely TGO thickness and crack size are identified for the complete characterization of the TBC failures. Between the two, however, TGO growth is considered as the primary damage and is strongly influenced by the exposed temperature cycles. Cracks develop at specific locations and geometry of TGO profile due to stress and strain environments and are considered as the secondary damage source. The cracking and interface separation mechanisms are influenced by shape and curvatures at the TGO boundaries as may be seen in Figure 1(b). Cracks are mostly observed at and around sharp locations and ridges as at these locations, stresses must be higher and deformability may be less compatible between the adjacent phases with growing TGO (Kumar et al., 2010; Kumar et al., 2009).The

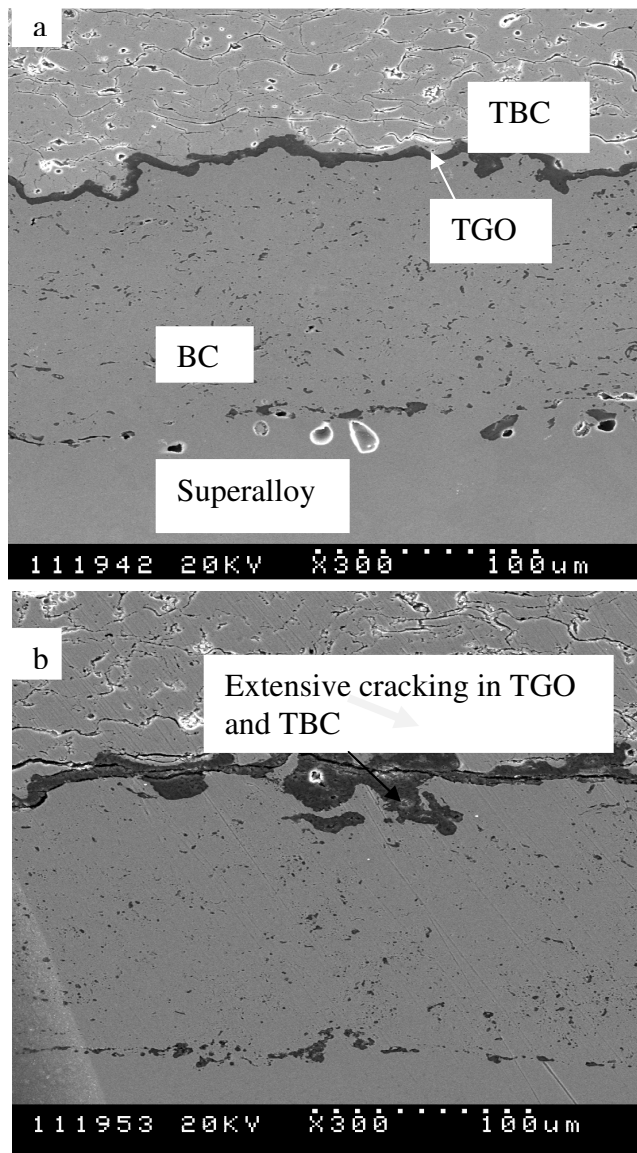


Figure 1: Microstructural demonstration of structural damages and progressive degradation in TBC; a) after 400 thermal cycles and b) after failure (around 1500 cycles).

zigzag profile of the BC-YSZ interface is held responsible for stress concentration to cause cracking at the interface and in YSZ. Experimental details for observations and quantification of damage signatures are documented elsewhere (Kumar et. al., 2010; Kumar et. al., 2009).

2.2 TGO-Crack Growth Relation

A functional relationship between the two forms of physical damages, namely crack damage and TGO growth is shown in Figure 2. Clearly and consistently, an exponential dependency between statistical mean of crack size in

different samples and the mean TGO thickness data is evident. The two plots for samples M07 (treated in normal atmospheric condition) and M08 (treated in vacuum) lie fairly close to each other demonstrating no significant effects of oxygen pressure on the physical damage size in TBC system as mentioned in earlier section. A linear relation between equivalent TGO thickness and maximum crack length was reported earlier for same TBC system (Chen et. al., 2006). An alternative and more meaningful dependency between the two parameters may also emerge by considering the first three points for both the classes having a linear function. The fourth data obtained at failures appears to follow exponentially function. This trend suggests two stage TGO growths kinetic from early oxidation until the failure time. During early stage of TBC life (up to 430 cycles), the cracking mechanisms in TBC is predominantly crack nucleation and opening controlled, while the mechanisms changes to predominant propagation controlled mode towards the later stage (exceeding 430 cycles) and until TBC failure. Two stage kinetics of oxide growth- cracking relation can be related to the changes in oxide phase changes. Alumina formation changes to mixed oxide formation consisting of Cr, Ni, Co, Al due to depletion in aluminium. Growth rate for mixed oxide is higher than alumina (Carlsson, 2007). It is confirmed that thermal cycling up to 500 cycles TGO is made of pure alumina while the internal oxide in BC consists of both alumina and spinels. However, some uncertainties are reported to be always involved in the oxide scale measurements because of higher instability of spinels as compared to alumina.

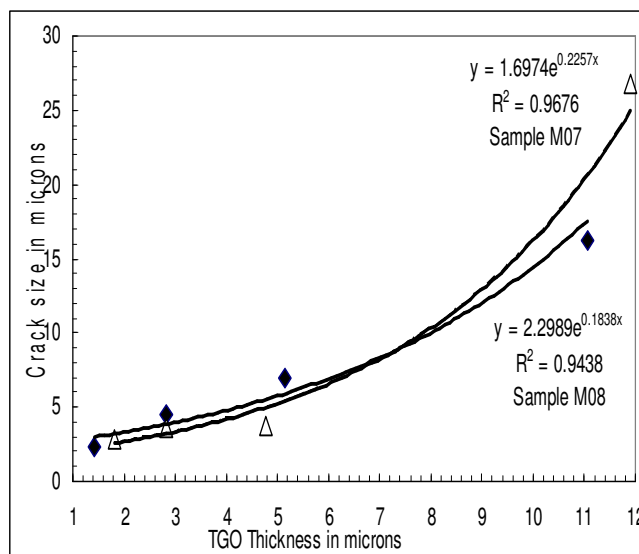


Figure 2: Exponential relationships are evident between two forms of physical damage signatures associated with thermal cycling of TBC; M07 samples were heattreated in normal air and M08 samples were heattreated in low pressure oxygen.

3. DAMAGE TOLERANCE MODEL

A damage tolerance model considering fracture mechanical approach was developed earlier for conventional three layer TBC system (Kumar et. al, 2007). The model establishes relation among various fracture critical factors, namely crack driving force, applied stress, crack size, and layer dimensions. The model considers isostrain behavior of the layers in TBC system based on the balance of elastic energy between the externally applied force and the localized stress fields around the cracks. The normalized stress intensity factor (SIF), K_I/K_0 is found to be an effective parameter in evaluating the fracture resistance of the interfaces in TBCs. The SIF ratio is represented as

$$\begin{aligned} \frac{K_I}{K_0} &= 1/\left[(m^2 - n^2)^{1/2} + \left(\frac{E_2}{E_1}\right)(1 - n^2)^{1/2} - \left(\frac{E_2}{E_1}\right)(m^2 - n^2)^{1/2} \right] \\ &= 1/\left[\left(1 - \frac{E_2}{E_1}\right)(m^2 - n^2)^{1/2} + \left(\frac{E_2}{E_1}\right)(1 - n^2)^{1/2} \right] \end{aligned} \quad (1)$$

where m is the crack depth to width ratio and n is the ratio of two adjacent layer widths. For the interface between TBC layer and TGO layer, n should be considered as the ratio of W_1 to W and for the other interface i.e. between TGO layer and BC, n should be ratio between W_2 to W (layer width W ,s are defined in Figure 3). E_1 and E_2 are the elastic moduli. Eq. (1) has been used to compute the normalized SIF for various situations that can be exploited for redesigning BC-TBC interface by FGM optimization. The damage signature data and the damage tolerance models are further exploited for prognostic health management solution for the thermal barrier coating applications. Various thermal cycles are designed and simulated to find PHM solutions as discussed in the following sections. DT model ensures that the stress intensity factor at the crack tip for a growing crack must be sufficiently smaller than the fracture toughness values for the material and loading conditions i.e. $K_{I,applied} \ll K_{IC}$. The ratio between the two K_I values gives a measure of safety factor against the linear elastic controlled fracture. However, on-line crack size measurement during thermal cycling of gas turbine blades and vanes are not simple and accurate. An indirect method based on physical damage developed in TBC for estimation of crack growth is considered.

3.1 Modification of DT Model

In the original model, the developing crack was assumed to be occurring within the TGO layer under thermal cycling. However, it is clear that the crack can grow significantly larger than the corresponding TGO thickness, especially at the later stages. This situation is illustrated in Figure 1. In order to keep the value of the Stress Intensity a real one, the crack size (a) cannot exceed the TGO thickness or the total width of the adjacent layers. If it does exceed, the value of the SIF simply becomes a complex number, and

cannot be used to represent a physical quantity in space. It is now assumed that cracks are present in the TBC layer, hence the Stress Intensity Factor is derived only for the TBC/TGO interface. The composite layer arrangement in TBCs is schematically represented in Figure 3. Details of assumptions, derivations and model are given in an earlier paper (Kumar et. al., 2007).

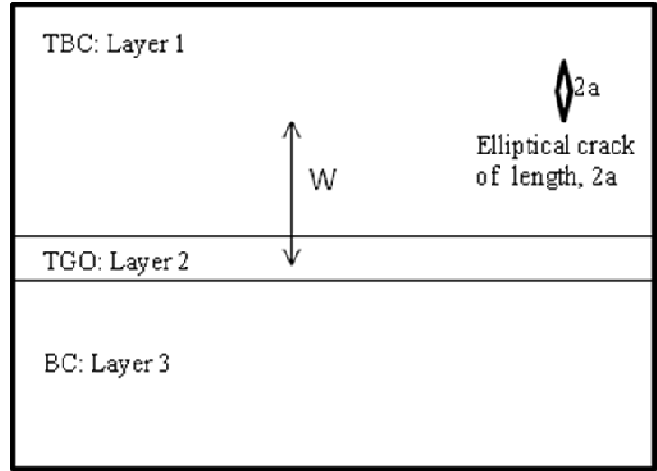


Figure 3: Cross-sectional view of the three layer layout in TBC system used for the model analysis. A thorough crack of size $2a$ is shown in the TBC layer and subjected to transverse stress, σ . Typical layer thicknesses are 300 microns for TBC ($2W_1$), 150 microns for BC ($2W_2$) and up to 15 microns for TGO layers ($2W_3$).

Westergaard's stress function and applied stress are assumed to be same as in the original model (Anderson, 1995). E_1 and E_2 are the elastic modulus of the adjacent and respective layers as in Figure 3.

$$\frac{F}{\sigma_2} = I = \int_a^{W_1} \frac{x}{(x^2 - a^2)^{1/2}} dx + \frac{E_2}{E_1} \int_{W_1}^{W_1+W_2} \frac{x}{(x^2 - a^2)^{1/2}} dx \quad (2)$$

Where E_1 (=40GPa) and E_2 (=380GPa) are the elastic modulus respectively for TBC and TGO.

$$I = (W_1^2 - a^2)^{1/2} + \frac{E_2}{E_1} \left[((W_1 + W_2)^2 - a^2)^{1/2} - (W_1^2 - a^2)^{1/2} \right] \quad (3)$$

Hence, the SIF is as follows:

$$K_I = \frac{1.12(\sigma_{applied} x (W_2 + W_1)(\pi a)^{1/2})}{I} \quad (4)$$

$$K_I = \frac{1.12(\sigma_{applied} x (W_2 + W_1)(\pi n)^{1/2})}{\left[(W_1^2 - a^2)^{1/2} + \frac{E_2}{E_1} \left[((W_1 + W_2)^2 - a^2)^{1/2} - (W_1^2 - a^2)^{1/2} \right] \right]} \quad (5)$$

W_1 represents the thickness of the TBC layer, which we assume remains constant throughout the evolution of the system with respect to time. Hence by dividing the numerator and denominator of Eq.(5) by W_1 , the following expression is obtained.

$$K_I = \frac{1.12(\sigma_{applied} x W_1^{3/2}(1+m)(\pi n)^{1/2})}{\left[(1-n^2)^{1/2} + \frac{E_2}{E_1} \left[((1+m)^2 - n^2)^{1/2} - (1-n^2)^{1/2} \right] \right]} \quad (6)$$

$$K_I = \frac{1.12(\sigma_{applied} x W_1^{3/2}(1+m)(\pi n)^{1/2})}{\left[\left(1 - \frac{E_2}{E_1}\right)(1-n^2)^{1/2} + ((1+m)^2 - n^2)^{1/2} \right]} \quad (7)$$

Where $n = a/W_1$ and $m = W_2/W_1$

Defining a new parameter, K_0 ,

$$K_0 = 1.12 \cdot \sigma_{applied} \cdot W_1^{3/2} \quad (8)$$

Hence, the normalized SIF can be represented as:

$$\frac{K_I}{K_0} = \frac{(1+m)(\pi n)^{1/2}}{\left[\left(1 - \frac{E_2}{E_1}\right)(1-n^2)^{1/2} + \frac{E_2}{E_1} \left((1+m)^2 - n^2 \right)^{1/2} \right]} \quad (9)$$

Hence, using Eq.(9), the result will always be a real number for the SIF determination since $(1+m)$ will always be greater than n for all real values of m and n , and n will always be less than one (assuming that the crack size is always less than the thickness to the TBC layer for realistic TBC life).

4. SIMULATION

This section describes the simulation work with temperature profile and estimation of RUL (remaining useful life). The objective of the simulation work is to make predictions of TGO as well as crack growth in the TBC system under various thermal cycles that the TBCs are likely

to be exposed. The TGO growth as mentioned earlier is strongly related to temperature (T). The TGO growth predicted with time will yield the crack size and this will be used to compute normalized SIF as shown in Eq. (9). The temperature data is required to be monitored continuously as the hot-section structural parts with TBC are operational.

4.1 TGO Estimation

The differential equation that describes the growth of the TGO layer is represented as (Mao et. al., 2006):

$$\frac{dh}{dt} = \gamma_1 \cdot 10^{\left(\frac{a_0}{T(t)} + b_0\right) \gamma_1} \cdot h^{\left(1 - \frac{1}{\gamma_1}\right)} \quad (10)$$

Where h is the TGO thickness; $T(t)$ is the temperature profile as a function of time, t ; a_0 , b_0 and γ_0 are the fitting constants. Eq. (10) can only be solved if the temperature function is defined properly. Even if the function is known, there is not a guarantee that an integration method exists for the equation. Clearly, for time-varying temperature functions, we see that numerical analysis is to be applied. To obtain the best practical compromise between accuracy and the computational effort of the numerical analysis, the fourth order Runge-Kutta method is used.

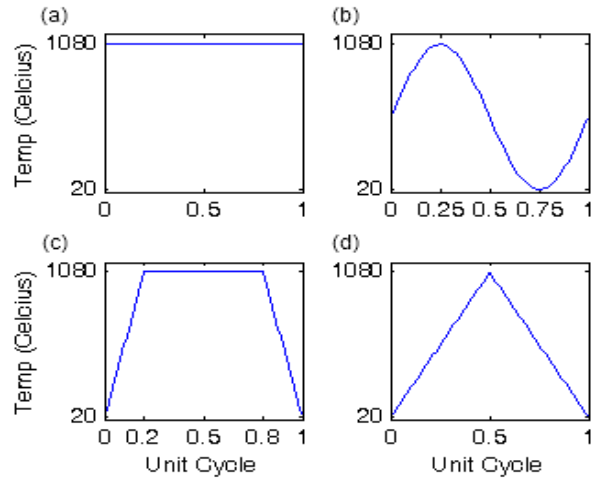


Figure 4: Schematic illustrations of four thermal cycles considered for simulation studies and experimental work. The functional relationship for each is also given. a) Isothermal case; $T(t) = 1080^{\circ}C$; b) Sinusoidal case; c) Trapezoidal cycle and d) Triangular cycle

To solve the generalized problem numerically, we have

$$\frac{dh}{dt} = f(h, t); h(0) = h_0; \tag{11}$$

$h(0)$ represent the initial value of the thickness h at time $t=0$
 The time of interest is split into smaller intervals of duration which is written as: Δt . Thus, the time t now becomes:

$$t = n\Delta t; \forall n : 1, 2, 3, \dots, N; \tag{12}$$

Hence, the Runge-Kutta Method is known as the following iterative method:

$$k_1 = \Delta t f(h_n, n\Delta t); \tag{13}$$

$$k_2 = \Delta t f\left(h_n + \frac{k_1}{2}, \left(n + \frac{1}{2}\right)\Delta t\right); \tag{14}$$

$$k_3 = \Delta t f\left(h_n + \frac{k_2}{2}, \left(n + \frac{1}{2}\right)\Delta t\right); \tag{15}$$

$$k_4 = \Delta t f(h_n + k_2, (n+1)\Delta t); \tag{16}$$

$$h_{n+1} = h_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) + O(\Delta t)^5 \tag{17}$$

$O(\Delta t)^5$ represents the error in the numerical method which is of fifth order with respect to the time step taken.

Four different thermal cycles are considered for simulation studies for prognostics and these are schematically displayed in Figure 4. During the flight operation the hot section parts with TBC are likely to be exposed to different thermal cycles. The maximum and minimum temperatures in all cases are maintained between 1080°C and 20°C respectively. The temperature cycle that was used in our simulated experimental research was trapezoidal temperature profile as shown in Figure 4c and functionally can be represented by the following periodic equations.

$$T(t) = \begin{cases} 1060(t)/0.2+20; & \text{for } 0 \leq t \leq 0.2\tau \\ 1080; & \text{for } 0.2\tau \leq t < 0.8\tau \text{ and} \\ -1060(t-0.8) / 0.2 + 20; & \text{for } 0.8\tau \leq t < \tau \end{cases}$$

$T(t)$ = Temperature at time t (°C)

τ = Duration of one cycle (=1 hr.)

Distinctly, the extent of thermal exposure of TBCs will vary with thermal cycles (Figure 4) and influence TGO growth and cracking. A comparison between the empirical relations and the actual experimental TGO growth is made in Figure 5.

There is no significant difference in the derived values and the experimentally observed data. Also, time period of the thermal cycling is negligible compared to the evolution of the TGO thickness for this case to a certain degree. The periodic equation is sufficient enough for the purpose of approximating the rise of the TGO thickness with respect to time under thermal stress. The empirical

formula can now be used to determine the TGO growth with respect to time, and eventually the values can be used for crack growth estimation as discussed in section 4.2.

4.2 Crack Growth and SIF

Till date, determining the crack size within the system has been quite a challenge. In order to predict a crack in a system, one initially needs to acquire the amount of thermal stress and the mechanical stress that will be induced to the TBC system. The process of measuring the amount of stress in the system or predicting it through the use of an algorithm has not yet been feasible, hence making it difficult for us to determine the crack size with relation to stress. This issue can be resolved by assuming a relation with respect to another variable. Since the TGO growth is the only other variable available from this experiment, a comparison between the TGO and crack size at their corresponding thermal cycle duration was conducted. The results are shown in the following section. Discussion on experimental details and quantitative estimation of damage signatures like crack size and TGO growth is beyond the scope of discussion here. Damage data was obtained by thermal cycling of TBC samples, metallography, scanning microscopy, rigorous measurements and analysis (Kumar et. al., 2010; Kumar et. al, 2009). Both Figures 2 and 5 demonstrate the relationship is of exponential type and not linear.

The relation between the thickness of the TGO and the Crack size for different number of thermal cycles appears to be exponential. The following equation for the growth of the crack size, a_c is obtained from experimental data as shown in Figure 2.

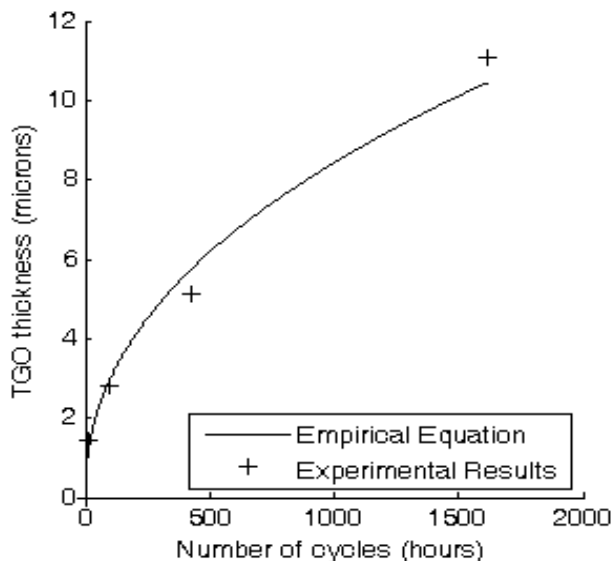


Figure 5: Comparison of simulated TGO growth in TBC with experimental data for identical thermal cycles.

$$a_t = (1.7 - 2.3) \exp(0.20 * TGO_t) \quad (18)$$

The function TGO_t is the value of the TGO thickness that will be determined from the eqs (1) to (5) which are mentioned in the previous section. As the crack size becomes available, Eq. 9 will yield the stress intensity factor as a function of time, t while the other layer dimensions and material data remaining constant.

5. PHM SOLUTION

An overview of the PHM solution for TBCs is shown in Figure 6 describing the algorithm implemented into the MATLAB™ program. The program is designed to determine the TGO thickness, crack Size, normalized SIF in real-time, trend analysis for RUL and plot new results after each calculation. A crack tolerance limit for TBCs is set in the program to allow maximum SIF that the system is supposed to withstand. In case the values overpass the margin, the plotting will carry on, but a warning sign will be indicated to the user.

computations.

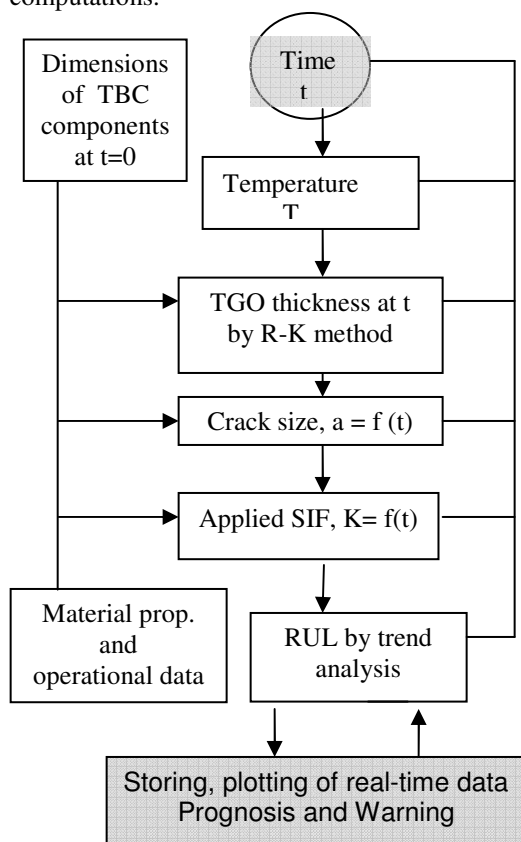


Figure 6: Overview of the proposed physics based PHM for the thermal barrier coating system as applied to hot-section components in aeroengine

As the temperature of the TBC system is measured, the timer activates the program. The TGO thickness and the crack size are determined. Now, using the TGO thickness and crack size, along with some predefined parameters such as the elastic modulus, speed, pressure and the dimensions of the TBC and BC layers, the program is able to compute the normalized SIF. All these values are then stored as an array and plotted for the user. All these operations are completed within the interval of time set for the timer, so that it does not interfere with monitored data and

5.1 RUL

Proposed PHM solution aims to obtain remaining useful life (RUL) based on the current health status of TBCs. As mentioned in earlier section that the SIF provides the current crack tolerance ability at a given time and configuration. Figure 7 illustrates the RUL estimation using SIF data and least square polynomial regression analysis. With operational cycle the TGO and cracks grow and so SIF level, prior to the attainment of threshold SIF level. Nonlinear polynomial regression through data points leads to the estimated RUL. However, the RUL will continue to vary as more new SIF data points are obtained and regression coefficients are like to change significantly. Some more discussion is relevant here with regard to TBCs crack tolerance behavior.

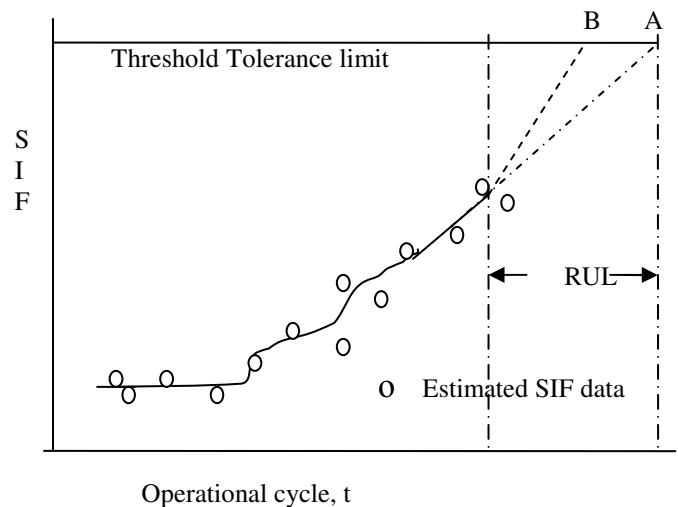


Figure 7 Illustration of RUL estimation at any point of time during operational life; points A and B signify the estimated RUL through data regression without and with some degree of conservatism.

The higher cracking and separation tendencies at TGO-TBC interface are primarily because of lower fracture resistance as compared to the BC-TGO interface. The fracture energy of the TBC-TGO interface is around only 2-4 $Mpa.m^{1/2}$ after 100 hrs. of oxidation. The fracture energy for BC-TGO layers has not been reported so far, but it is

likely to be much higher because BC layer toughness is more than $60 \text{ Mpa.m}^{1/2}$. During the early stage of growth, the separation and cracking mostly remain at and around the TGO-TBC interfaces. Under isostrain condition, an applied stress of magnitude around 100 Mpa results in K_I values of 2 to 4 $\text{Mpa.m}^{1/2}$ which is just enough for initiating a crack from a defect size of 2 microns. The stress level required for crack driving force (K_I / G_I) to be around and larger than the fracture resistance of the TBC component materials (TBC, TGO) are found to be in the range of 50 to 500 Mpa for crack size exceeding the critical length (> 1 micron) in TGO. The approximate stress levels as required for both crack initiation and propagation stages and computed from two models are consistent. The experimental work confirms the presence of cracks in the TBC system in the size range of 2-3 microns even at the onset of thermal cycling (Kumar et. al., 2010; Kumar et. al., 2009). An increasing tendency for stress intensity factor from 2 to 3.2 $\text{Mpa.m}^{1/2}$ was reported earlier with TGO growth from 4 to 8 microns (Tzimas et. al., 2000). It may be mentioned here that the research emphasized on the damage evolution and analysis for the TBC, rather than the method of prognostic analysis. However, standard regression analysis has been tried with damage signature data.

6. RESULTS

The simulation results for TGO growth and SIF estimation are given in Figures 8 and 9 for the thermal cycles considered. A smooth rise in the estimation are observed as shown in Figures 8b and 9b, while the actual pattern of TGO and SIF change can be seen at enlarger scale (up to 4 cycles) in Figures 8a and 9a respectively. Wide variations among the four are also evident in TGO, SIF and so will be in RUL as the TBC are exposed. The highest and continuous TGO growth and SIF increase are seen for isothermal temperature cycle and so the RUL may be expected to be shortest as compared to others. This is because of long uninterrupted thermal exposure at highest temperature of 1080°C. However, stepwise discontinuous changes in TGO and SIF are evident reflecting the nature of thermal cycles in other cases (Figures 4). The lowest TGO and SIF for any number of thermal cycles are obtained for triangular case as the TBCs are exposed to highest temperature momentarily. The Sinusoidal profile maintains high temperature longer than a triangular profile, thus having a faster TGO growth (Figure 8).

The other noteworthy issue affecting the RUL is that initial steep slopes of the plots tend to flatten with thermal exposure as the aluminium depletion in BC continues reducing the driving force for diffusion. The formation of other bulk mixed oxides, e.g NiO, $(\text{Cr, Al})_2\text{O}_3$, Fe_2O_3 , $(\text{Ni,CrAl})_2\text{O}_4$ etc. (Chen, Wu, Marple & Patnaik, 2005;

Chen, Wu, Marple & Patnaik, 2006; Sidhu & Prakash, 2005) also reduce the kinetics of oxidation process. Though maximum temperature has the major effect on RUL, but nature of oxidation and damage state depending upon the thermal cycle also determines the life time. Reducing the temperature from 1177°C to 1130°C is reported to increase sample lifetime by a factor of 2.4, though the damage state is observed to be same irrespective of temperature profile as long as the peak temperature remains constant (Nusier, Newas & Chaudhury, 2000).

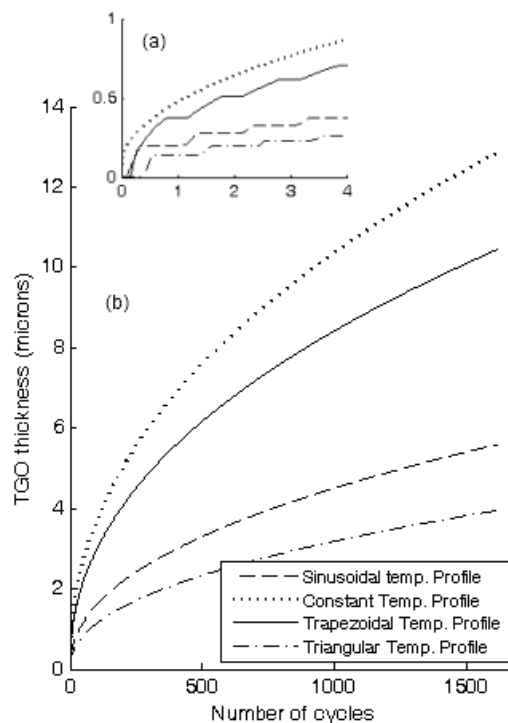


Figure 8: TGO growth rate characteristics in TBCs exposed to different thermal cycles; a) at magnified scale showing actual pattern of growth and b) at reduced scale indicating smooth rise for TGO thickness.

The other noteworthy issue affecting the RUL is that initial steep slopes of the plots tend to flatten with thermal exposure as the aluminium depletion in BC continues reducing the driving force for diffusion. The formation of other bulk mixed oxides, e.g NiO, $(\text{Cr, Al})_2\text{O}_3$, Fe_2O_3 , $(\text{Ni,CrAl})_2\text{O}_4$ etc. (Chen, Wu, Marple & Patnaik, 2005; Chen, Wu, Marple & Patnaik, 2006; Sidhu & Prakash, 2005) also reduce the kinetics of oxidation process. Though maximum temperature has the major effect on RUL, but nature of oxidation and damage state depending upon the thermal cycle also determines the life time. Reducing the temperature from 1177°C to 1130°C is reported to increase sample lifetime by a factor of 2.4, though the damage state is observed to be same irrespective of temperature profile as

long as the peak temperature remains constant (Nusier, Newas & Chaudhury, 2000). However, further experimental studies on the nature of oxidation and damage and cracking mechanisms under different thermal cycles are required to substantiate the results.

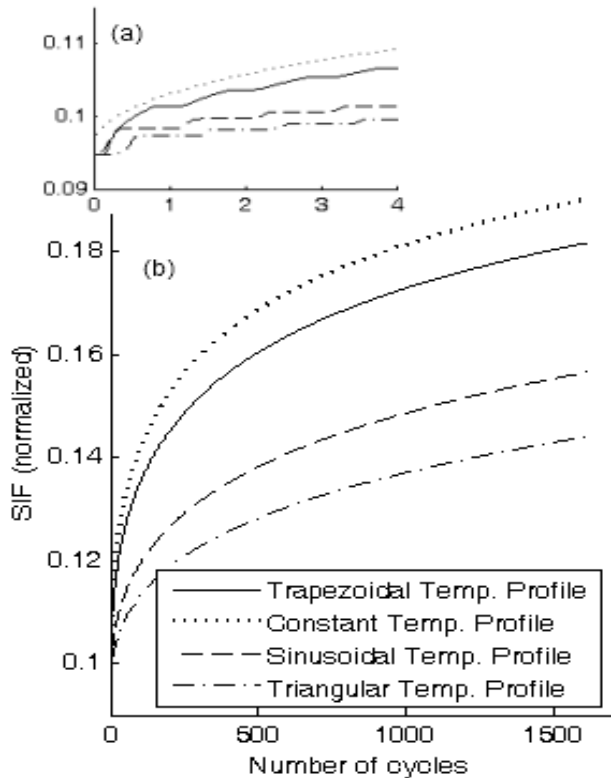


Figure 9: SIF change with number of cycles as TBCs are exposed to different thermal cycles; a) at magnified scale showing actual pattern and b) at reduced scale indicating smooth increase of SIF.

7. CONCLUSIONS

Experimental and simulation studies on the prognostic assessment of thermal barrier coating system were carried out using physics based approaches. Two damage signatures, namely growth of aluminium oxide at the interface between bond coat and top insulating coat and the cracks are responsible for the failure of TBCs. An exponential relationship between the two signatures is established. Temperature being the driving force for diffusion and TGO, four thermal cycle profiles are simulated and fourth-order Runge-Kutta method is used for numerical solution. For TBC system stability and crack tolerance, a modified fracture mechanical model is used assuming that the cracks form and grow in the top TBC layer. The normalized stress intensity factor determines the current health and remaining useful life using the regression

analysis. The TGO and crack tolerance level based on the simulation results vary widely and largely depends on the extent of thermal exposure to TBC.

ACKNOWLEDGEMENT

Authors gratefully acknowledge the generous support and encouragement received from Institute of Aerospace Research, NRC, Ottawa for conducting experiments with thermal barrier coating samples.

REFERENCES

- Chin, H. H., Turbine engine hot section prognostics, <http://www.appliedconceptresearch.com/Turbine%20Engine%20Hot%20Section%20Prognostics.pdf>
- Wood, M. I., (2000), Gas turbine hot section components: the challenge of residual life assessment, *Proceedings of Institution of Mechanical Engrs.*, 214 part A, pp. 193-201.
- Christodoulou, L. & Larsen, J. M., (2005), Material Damage Prognosis: A Revolution in Asset Management, in *Material damage Prognosis* (ed. J M Larsen et.al.), TMS, pp. 3-10.
- Intellistart⁺, Altair avionics/Pratt & Whitney; <http://www.altairavionics.com/>
- SignalProTM, Impact <http://impact.tek.com/Aerospace/Aerospace.html>
- NormNetPHM, Frontier Technology Inc., <http://www.fti.net.com/cm/products/NormNet-prod/NormNet.html>
- Shillington, E.A.G. & Clarke, D.R., (1999), Spalling failure of a thermal barrier coating associated with aluminium in the bondcoat, *Acta Materialia*, vol. 47, 4, pp.1297-1305.
- Evans, A. G., Mumm, D. R., Hutchinson, J. W., Meier, G. H. & Pettit, F. S., (2001), Mechanisms controlling the durability of thermal barrier coatings, *Progress in Materials science*, vol.46, pp. 505-533.
- Kumar, A. N., Nayak, A., Patnaik, A. R., Wu, X. & Patnaik, P.C., (2007), Instability analysis for thermal barrier coatings by fracture mechanical modelings, *Proceedings of GT2007, ASME Turbo Expo 2007: Power for Land, Sea and Air*, GT 2007 – 27489, Montreal, QC.
- Karlsson, A. M, Xu, T. & Evans, A G., (2002), The effect of thermal barrier coating on the displacement instability in thermal barrier system, *Acta Materialia*, vol.50, pp. 1211-1218.
- Chen, W. R., Wu, X., Marple, B. R. & Patnaik, P.C., (2005), Oxidation and crack nucleation / growth in an air-plasma-sprayed thermal barrier coating with NiCrAlY bond-coat, *Surface and Coating and Technology*, vol. 197, pp. 109-115.
- Clarke, D.R., Levi, C. G. & Evans, A. G., (2006), Enhanced zirconia thermal barrier coating systems, *Journal of Power and Energy*, Proc. I. MechE, vol. 220, pp.

85- 92.

- He, M.Y., Hutchinson, J. W. & Evans, A. G.,(2003), Simulation of stresses and delamination in a plasma-sprayed thermal barrier system upon thermal cycling, *Materials science and Engineering*, vol. A345, pp. 172-178.
- LeMieux, D. H., (2003), Online thermal barrier coating monitoring for real time failure protection and life maximization, *Report US Department of Energy*, DE-FC26 -01NT41232.
- Kumar, A., Srivastava, A., Goel, N., & Nayak, A., (2010), Model based approach and algorithm for fault diagnosis and prognosis of coated gas turbine blades, *Proceeding of IEEE/ASME International conference on Advanced Intelligent Mechatronics*, Montreal, QC.
- Kumar, A., Nayak, A., Srivastava, A., & Goel, N. (2009), Experimental validation of statistical algorithm for diagnosis of damage fault, *Proceedings of IEEE Canadian conference on electrical and computer engineering*, St. John's.
- Chen, W.R., Wu, X., Marple, B.R. and Patnaik, P.C.,(2006), The growth and influence of thermally grown oxide in a thermal barrier coating, *Surface and Coating Technology*, vol. 201, pp.1074-1079.
- Carlsson, K.,(2007), *A Study of failure development in thick thermal barrier coating*. Master thesis in Mechanical Engineering, Linkopings University Tekniska Hogskolan, Sweden.
- Anderson, T. L., (1995), *Fracture Mechanics: Fundamentals and Applications*, Boca Raton, CRC Press, USA.
- Mao, W. G., Zhou, Y. C., Yang, L., & Yu, X. H. (2006), Modeling of residual stresses variation with thermal cycling in thermal barrier coatings, *Mechanics of Materials*, vol. 38, pp. 1118-1127.
- Tzimas, E., Mullejans, H., Peteves, S. D. Bressers, J. and Stamm, W.,(2000), Failure of TBC systems under cyclic thermomechanical loading, *Acta Materialia*, 48(18-19),pp. 4699-4707.
- Sidhu, B.S. & Prakash, S. (2005), High temperature oxidation behavior of NiCrAlY bond coats and stellite -6 plasma -sprayed coatings", *Oxidation of Metals*, vol. 63, 314, pp. 241-259.
- Nusier, S. Q., Newas, G. M & Chaudhury, Z. A. (2000), Experimental and analytical evaluation of damage processes in thermal barrier coatings, *International Journal of Solids and Structures*, vol. 37, 18, pp. 2495-2506.

Amar Kumar has more than 25 years of research and consulting experience in the fields of structural materials characterization and development, fracture mechanics, failure analysis and applications. At present Dr. Kumar is working as senior research scientist in the development projects of diagnostics, prognostics and health management

of aeroengine components. Prior to joining the current assignment in 2006, he was employed with Indian Institute of Technology at Delhi, India as a teaching faculty after obtaining his Ph.D degree from the same Institution. Dr. Kumar has published more than 150 research papers in refereed journals, conference proceedings, and technical reports.

Bhavaye Saxena is currently an undergraduate student in the University of Ottawa under the department of Physics. His interests lie in computational physics, and have recently put interest in deriving models related to engineering physics.

Alka Srivastava has 24 years of research, administrative and industrial experience and has a BAsC degree in Electrical engineering from University of Ottawa. At present she is the manager of the R & D Division and leads several teams working in the fields of Fault Tolerance, Prognostic health management etc.

Alok Goel has more than 30 years of manufacturing, research, and industrial experience. He has a M.Sc degree in Advance Manufacturing Systems & Technology from University of Liverpool, UK. At present he is President of OMTEC Inc., (Consulting Engineer's). His research interests are in naval fuel filtration and lube oil separation. He is specialized in manufacturing graded natural fiber as a filler material for plastic reinforcement application.

Physics based Prognostics of Solder Joints in Avionics

Avisekh Banerjee¹, Ashok K Koul¹, Amar Kumar², Nishith Goel³

¹ *Life Prediction Technologies Inc., 23-1010 Polytek Street, Ottawa, Ontario, K1J 9J1, Canada*
banerjeea@lifepredictiontech.com
koula@lifepredictiontech.com

² *Teccis Corporation, 200-210 Colonnade Road, Ottawa, ON, K2E 7L5, Canada*
amar@teccis.ca

³ *Cistel Technology, 40-30 Concourse Road, Nepean, ON, K2E 7V7, Canada*
ngoel@cistel.com

ABSTRACT

Applicability of a physics based prognostics approach for solder joints using microstructural damage models is investigated. A modified deformation mechanism map for the solder alloys is introduced where grain boundary sliding (GBS) plays a dominant role during creep deformation. The high homologous temperature of solder as well as the combined thermal-vibration cycling experienced during typical operating missions necessitates the use of a combined creep-fatigue failure approach. In this work, a PCB consisting of a heat generating chip with Ball-Grid Array (BGA) solder joints is considered for avionics application. A prognostics based Life Cycle Management approach was used to perform the mission analysis, FEA, thermal-mechanical stress analysis and damage accumulation analysis. The remaining useful life (RUL) is predicted for different rupture strains. The uniqueness of this approach lies in the use of microstructure based damage models and consideration of both material and mission variability to predict the RUL under actual usage. The life critical nodes were observed near the junction of the solder joints with the substrate due to high disparities in their coefficients of thermal expansion. In addition, the probabilistic analysis was also performed by randomly varying the grain size and fitting a two-parameter Weibull distribution to the failure data. The model calibration and the results show some practical trends that need to be verified through future experimentation. The simulation results demonstrate the viability of using a physics based approach for the prognosis of solder joint failures in avionics.

Banerjee et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

1. INTRODUCTION

Life prediction analysis of solder joints is a popular, but challenging topic due to high occurrence of failures in the field. The mechanical fault progression leads to electrical failure of solder joints causing about 70% of overall failures in avionics. The failures of the fundamental avionic components like transistors and their interconnections are mostly caused by operating thermal, mechanical and electrical overstresses (Saha, Celaya, Wysocki & Goebel, 2005). Previous studies have considered empirical (Kalgren, Baybutt, Ginart, Minnella, Roemer & Dabney, 2007) or only simplified physics based thermal fatigue (Nasser, Tryon, and Dey, 2005) models for the prognostics of electronic components. However approaches involving top-down multi-component analysis techniques (Kalgren, et al., 2007) combined with empirical models require the availability of significant amount of data along with considerable deviation from the norm to predict the presence of a fault. This makes the early detection or prediction of damage or faults very difficult. Moreover, the relative scaling of electronic components and detectable crack sizes limits the use of traditional empirical damage models from a life prediction or prognostics perspectives. In contrast, only considering thermal fatigue induced transgranular fractures of solder joints (Nasser et al., 2005) may not provide an accurate fault prediction because creep damage may also contribute to the overall damage accumulation process.

The Pb-Sn solder joints in electronic packages function as electrical interconnections, as well as mechanical bonds. The solder joints often consist of materials possessing different thermal expansion coefficients and this imposes cyclic strains under thermal loading fluctuations. Thermal fluctuations can occur due to external temperature variation or internal heat dissipation. These temperature fluctuations can be large in electronic components in avionics. Even small temperature fluctuations can lead to significant cyclic

strain accumulation, depending upon the size of the joint and the difference in the thermal expansion of the joined materials. One of the most important requirements of the new solder materials is the reliability of the solder joints against thermal cycling, flexural bending, and impact loading. An in-depth understanding of the micro-mechanistic processes leading to the solder joint failures under conditions of thermal-mechanical fatigue and creep has been achieved through great deal of research (Dasgupta, Sharma & Upadhyayula, 2001; Joo, Yu & Shin, 2003; Kovacevic, Drogenik & Kolar, 2010; Shi, Wang, Yang & Pang, 2003).

Due to the high homologous operating temperatures, deformation of solder joints is always governed by a combination of creep and TMF processes. Solder joints are exposed to time dependent high temperature deformation mechanisms associated with creep and residual stress relaxation and the joints are also susceptible to low cycle fatigue (LCF) damage accumulation. Creep is the most common and important micromechanical deformation mechanism operative in the solder joints that eventually leads to failure. Microstructural features also influence the material properties and plastic deformation kinetics greatly. For Sn-Pb solders, the phase boundaries are known to be the preferred crack initiation sites. The cracks then propagate preferably along tin-lead or tin-tin grain boundaries (Joo et al., 2003). Continuous TMF loading will also induce creep deformation effects. Since room temperature for eutectic Sn-Pb alloys is around $0.65 T_m$ (T_m is the melting temperature in K), phase changes due to diffusion can also be expected to play a role at higher temperatures leading to accelerated damage accumulation.

Fatigue damage due to vibration loading leads to cyclic plasticity while that due to temperature cycling causes cyclic creep. Plastic deformation of the solder refer to instantaneous time scale and primarily occurs due to slip; while creep due to time dependent and diffusion-assisted mechanisms over long time, namely grain boundary sliding, dislocation glide/climb and mass transport through the grain boundary/matrix. Furthermore, there are interactions between vibration and temperature damage accumulation rates due to factors like material properties changes; microstructural coarsening and interaction between vibration stress and the TMF stress.

The lead-free SAC (Sn-Ag-Copper) is the alternative alloy as Pb is harmful to the environment and human beings. The alloys melt around 250 °C, depending on their composition. Different variations of the SAC alloy, with Ag content from 3.0% to 4.0% are all acceptable compositions. Creep rupture in SAC occurs by the nucleation of cavities and their subsequent growth by continued creep damage accumulation. The 1.5Cu SAC shows the poorest creep ductility because of the brittle cracking of the intermetallic

Cu_6Sn_5 , which provided easy nucleation and crack propagation sites for creep cavities (Joo et al., 2003).

These observations suggest that a number of deformation and failure mechanisms contribute to solder alloy system deformation and fracture depending mainly on applied stress and temperature. Some of these mechanisms include plasticity, dislocation creep and grain boundary deformation accommodated by different processes. To assess the current health and RUL of solder joints, it is important to employ the appropriate constitutive models for deformation and fracture. Combining the constitutive models for various regimes is useful for determining the creep strain rates and the remaining useful life (RUL) for solder joints (Kovacevic et al., 2010; Shi et al., 2003).

Gu and Pecht, (2010) provide several examples of implementing prognostics and health assessment for electronics products in the industry and defense applications. The paper also discusses how the traditional handbook-based reliability prediction methods for electronic products like MIL-HDBK-217 are being replaced by PHM. Approaches like physics-of failure, data-driven and their combination has been discussed in detail.

Hence a reliable physics based prognostics system including both mission as well as microstructural variabilities possess a good potential for facilitating the accurate prediction of the RUL of the PCB. This would enable a user to gauge the health of an existing PCB and optimally plan maintenance schedules as well as help in designing PCBs to withstand the loads for the intended application.

2. PHYSICS-BASED PROGNOSTICS FOR AVIONICS

In this paper, a prognostics-based Life Cycle Management framework is proposed to predict the RUL of avionic components. The combined effect of creep and thermal fatigue loads is considered on the damage evolution leading to intergranular as well as transgranular deformation of solder joints. The major causes of failure of solder joints are TMF cycling arising from the operational changes as well as creep due to the presence of a high operating temperature in terms of the homologous temperature of the eutectic solder. The soldering process also imparts intrinsic residual stresses that arise due to the difference in the thermal properties of the solder/intermetallic/substrate. The stress relaxation caused by the grain/phase boundary sliding leads to creep deformation of the solder joint also leads to crack nucleation during service. At the same time, the variation in the operating loads leads to TMF damage accumulation. The intergranular and transgranular deformation based combined creep-fatigue approach to damage accumulation would thus provide a more accurate simulation of the actual failure of the solder joints and it would also lead to more accurate predictions.

2.1 Deformation Mechanisms

A deformation mechanism map for the Pb-Sn eutectic solder is shown in Figure 1. The deformation mechanism map is a stress-temperature diagram presenting the dependency of normalized stress τ/G (G is the shear modulus) on homologous temperature T/T_m . Elastic region exists only at the very slow strain rate ($<10^{-10}$), while plastic region occurs over yield strength level. The dislocation controlled creep regime consists of three subzones, namely high stress creep regime, low temperature (LT) dislocation glide creep regime, and high temperature (HT) dislocation climb creep regime. Below this regions of creep, the two diffusion controlled regimes exists, namely grain boundary (GB) regime and matrix diffusion regime.

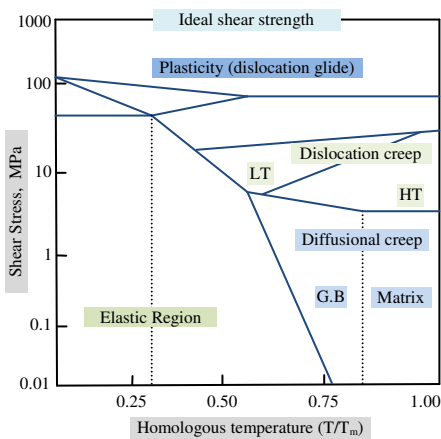


Figure 1: Line diagram for deformation mechanism map of Sn-Pb eutectic solder alloy highlighting the essential features

In parallel with Ashby's deformation mechanism map, Mohammed and Langdon (1974) considered an alternative to this map where grain boundary sliding (GBS) instead of diffusion creep predominates. Other attempts to accommodate GBS field in Ashby type maps have also been presented by Koul, Immariageon and Wallace (1994). In 2002, Wardsworth, Ruano and Sherby (2002) conducted a detailed analysis of all the data on the diffusional creep of engineering alloys and concluded that GBS dominated the deformation which had commonly been confused with diffusional creep. Based on the mechanistic modeling work of Wu and Koul (1993 and 1995), Wu, Yandt and Zhang (2009), presented an alternate map for engineering alloys.

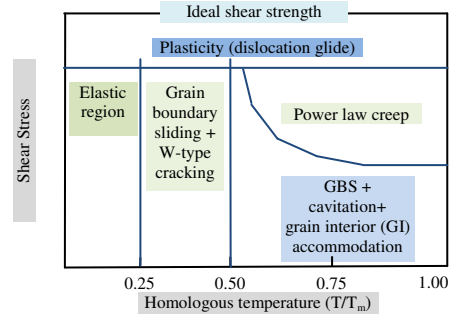


Figure 2: Modified deformation mechanism map with grain boundary sliding

These changes have been incorporated in the form of modifications to the deformation mechanism map presented in Figure 1, while considering the creep behavior of eutectic solder in this study, Figure 2.

PROBLEM FORMULATION

In this work, the problem of prognosis of electronic circuit boards in avionics has been considered and the RUL of solder joints is predicted. A detailed 30 hour long mission suitable for a typical transportation aircraft has been designed and used to determine the operating conditions during the mission. The problem is treated as combined creep and TMF damage accumulation process and analysis based on the microstructural properties of eutectic solder in a Ball-Grid Array (BGA) subject to the mission experienced by the transport aircraft is carried out. The circuit board is assumed to be located in the forward avionics bay of a transport aircraft consisting of a chip (BT substrate) with 8 solder joints mounted on an FR4 board as shown in Figure 3. The dimensions of the different components are also shown in the figure. The substrate has been assumed to have internal heat generation capacity with convection cooling allowing the heat distribution over the entire circuit. The operating conditions like ambient temperature, acceleration change along with the specific mission also need to be considered in physics based prognostics approach. Hence it is proposed that the mission as well as the microstructural variability have to be simultaneously considered for accurate prognosis.

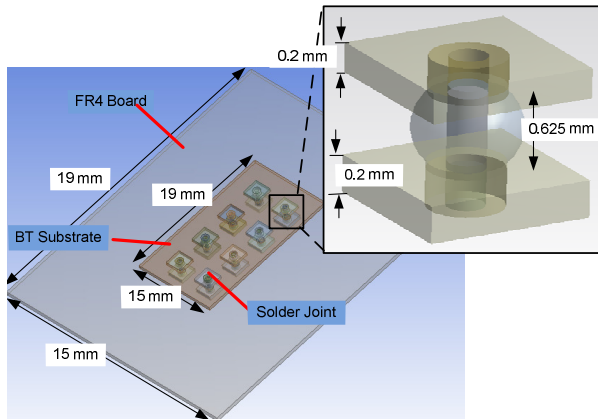


Figure 3: Typical geometry of a PCB for Avionics

4. PROGNOSTICS BASED LCM METHODOLOGY

A bottom-up prognostics based Life Cycle Management (LCM) approach (Koul, Tiku, Bhanot & Junkin, 2007) has been adopted. This involves the systematic consideration of the requisite inputs like material and geometry of the components and the usage. The temperature, stress and strain is calculated to determine the microstructural damage accumulation based nodal life enabling the determination of the RUL as well as the fracture critical location/node. The framework of the prognostic approach is shown in Figure 4 with each module described in details as below:

- **Input Data**

Component geometry: The three dimensional model of the component is created to generate the mesh for subsequent FEA.

In-service operating data: This is required to utilize the actual usage of the component rather than designed usage for more accurate prognosis. Typical operating data collected are RPM, Altitude since it governs the ambient conditions, from where other dependent parameters are calculated to determine the relevant parameters of the mission profile.

Material Data: Microstructural data like grain size, boundary precipitate size, activation energy, etc are requisite for the damage analysis. Simultaneously temperature dependent and independent physical data like elastic modulus, poisson ratio, conductivity, etc are also required for materials used for every component.

- **Pre-Processing**

Mission profile analysis: Once the mission has been obtained from the in-service operating data, a fuzzy logic based mission profile analyzer is used to determine the creep and fatigue loads on the components, their duration or frequency and their sequence.

Thermal and Structural Loading: Based on the in-service operating condition and the mission profile analysis, thermal and mechanical loads are determined

along with the requisite boundary conditions to closely replicate the effect of service exposure.

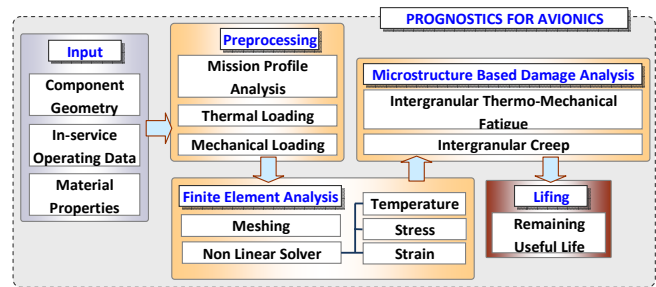


Figure 4: Framework of prognosis based LCM for solder joints in avionics

Finite Element Analysis: Well structured and mapped mesh is generated from the component geometry to conduct the thermal and structural analysis under the pre-determined loading and boundary conditions to calculate the nodal temperature, stress and strain.

- **Microstructure based Damage Analysis**

Microstructure based damage models under intergranular, transgranular and combined creep (Wu and Koul, 1995) and thermo-mechanical fatigue (Neu and Sehitoglu, 1989) has been implemented. These models take into account the microstructure, physical properties and their variation with temperature, operating condition and calibration of empirical coefficients with experimental data.

- **Life Prediction Analysis**

Based on the nodal temperature, stress and strain obtained from FEA, microstructural damage models are applied at each node to determine the accumulated damage as a result of the creep and fatigue loads. Robinson and Miner's damage summation rule is applied to determine the total damage accumulated during each mission and RUL is calculated for each node. This also allows the determination of the primary, secondary and tertiary failure critical locations.

5. SIMULATION SETUP

5.1 Geometry and Meshing

Geometry of the PCB consisting of 8 BGA solder joints, one BT substrate and one FR4 board was created as shown in Figure 3. Structured mapped mesh was generated and symmetry was used as shown in Figure 5 to reduce the computational cost. The two solder joints are numbered 1 and 2 for ease of referencing in the subsequent text. A total of 17663 quadrilateral 3D mesh elements were used for a quarter symmetric model.

5.2 Material Data Collection

Three different types of data were collected, as below:

- *Microstructural data:* Grain size, boundary precipitate size, interlamellar distance, activation energy, diffusion coefficient, etc.
- *Physical properties:* Temperature dependent and independent physical properties like Young’s Modulus, Poisson’s Ratio, Density, CTE, etc.
- *Calibration data:* Creep (strain vs. time) and fatigue test (strain vs. number of cycles) data for solder material (Sn63Pb37).

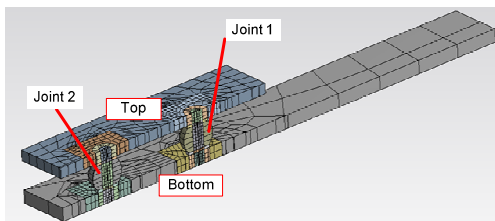
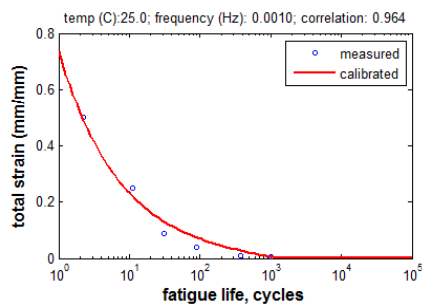


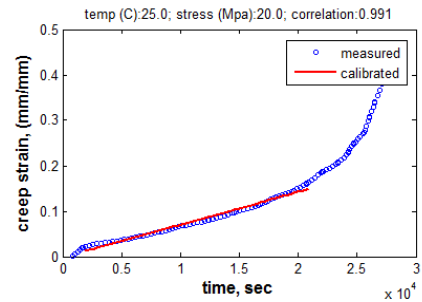
Figure 5: Sectional view of mesh for quarter symmetric circuit board

5.3 TMF and Creep Modeling

The microstructural based damage models for intergranular creep and TMF were calibrated for the eutectic solder alloy. For the creep model, experimentally measured creep life (strain vs. time) test data (Wong, Lau & Fenger, 2004) was utilized to calibrate the measured strain rate due to intergranular deformation caused by grain boundary sliding. For the fatigue model, experimentally obtained fatigue life data (strain vs. number of cycles) was used to calibrate the empirical material constants (Shi, Wang, Yang, & Pang, 2003). The microstructural data for eutectic solder was also obtained from existing literature and applied to both models. The calibrated TMF and creep models with experimental data are shown in Figure 6.



(a) TMF



(b) Creep

Figure 6: Calibration of intergranular damage models

5.4 Mission Profile Analysis

Mission profile closely representing that of a transport aircraft is required. For this purpose a detailed mission profile for typical transport aircraft was generated for a total flight duration of 30 hours with the cruise altitude being around 9,000 meter. Details of the mission were included by incorporating the change in the rpm and altitude at different stages of the mission, as shown in Figure 7. The other dependent parameters like ambient temperature, ambient pressure, acceleration were calculated. An initial temperature of 25°C was assumed at the ground level. Ambient temperature along the mission was calculated from the altitude in the mission which affects all the avionic components. The ambient temperature was added to the temperature profile generated by the chip’s internal heat generation and convective cooling to determine the resultant temperature at every time step of the mission. Moreover the vibratory acceleration amplitude exerted on all the components was also calculated as a function of rpm (Tang & Basaran, 2003; Smith, 2004) along the mission. Based on the temperature and acceleration profile, a fuzzy logic based mission profile analyzer was implemented to determine the creep and fatigue loads on the solder joint. The calibrated damage models were invoked at every time step based on the type of loading.

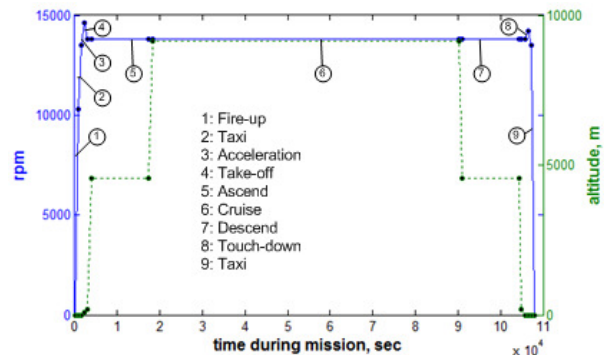


Figure 7: Designed mission profile

5.5 Boundary Conditions

The simplified substrate was assumed to have an internal heat generation of $0.5W/mm^3$ during its operation period which was assumed to be constant throughout the mission. A closed case convection for the front avionics bay region was assumed to have a coefficient of $20W/m^2\text{ }^\circ\text{C}$. The PCB was fixed at the four corners to represent attachment with the aircraft structural frame with screws. The RPM dependent acceleration was applied on all the components which would allow the four corners with least displacement where as the centre of the circuit board would have the maximum deflection.

5.6 Finite Element Analysis

The FEA analysis was performed with ANSYS Workbench with a coupled steady-state thermal and structural analysis. At first the temperature loads were applied on to the components and the thermal results were carried forward for subsequent structural analysis.

6. LIFING ANALYSIS

6.1 Remaining Useful Life

The temperature, stress and strain calculated from FEA based on the mission profile and other operating conditions were applied to the microstructural creep and TMF models. The result of the FEA namely temperature, stress and strain at each node of solder were calculated at each time-step with different fatigue damage models to determine damage accumulated at each node. Robinson and Miner's rule was used to sum the damage (D) for creep and fatigue loads at each load ($i=1$ to n) as below:

$$D = \sum_{i=1}^n \frac{t_i}{t_f} + \sum_{i=1}^n \frac{N_i}{N_f} \quad (1)$$

where t_i is the creep duration and N_i is the number of fatigue cycles for the i -th load, t_f and N_f are the failure creep duration and fatigue cycle. The remaining useful life (RUL) was calculated from the D and the total mission duration time (t_M) as below,

$$RUL = \frac{t_M \times (1 - D)}{D} \quad (2)$$

6.2 Probabilistic Analysis

Once the deterministic life critical nodes were identified based on the nodal temperature, stress and strain over the solder joints, a probabilistic analysis was conducted. In this analysis, the microstructural variability in terms of the grain size variation was considered. The grain size was selected as a major parameter since it plays an important role in the damage accumulation processes arising from combined creep and fatigue mechanisms. For the purpose of studying the variation in the RUL for different grain sizes, a normal

distribution of the grain size was considered. The mean size was the deterministic grain size and the standard deviation was assumed based on variations observed due to different reflow process parameters. Upon randomizing the grain size, probabilistic lifing calculation was carried out under steady-state operating conditions with Monte Carlo Simulation. A two parameter Weibull distribution of the probabilistic remaining useful life was also estimated for the most critical node.

7. RESULTS AND DISCUSSION

7.1 Finite Element Analysis

At first the temperature profile was calculated based on the ambient temperature, heat generated by the chip and convective cooling at each time step of the mission. A typical temperature profile is shown in Figure 8 (a). The temperature was highest over the chip which generates constant heat during the operation. The temperature was lowest at the board furthest away from the heat source, approximately resembling the ambient temperature condition. The thermal loads generated when combined with the mechanical load of vibratory acceleration resulted in maximum deflection at the centre of the board and chip being furthest away from the fixed support as shown in Figure 8 (b). The equivalent stresses and strains were found to be highest near the bottom surface of the solder Joint 1 which is due to the combination of higher temperature variation between the solder and the board as well as lower deformation due to closeness to the fixed support. The typical FEA results in terms of stress and strain distributions are shown in Figure 8 (c) and (d).

7.2 RUL Calculation

The spatial distribution of RUL for different rupture strains is shown in Figure 9. The figure shows that the region close to the bottom interfacing surface of Joint 1 has the lowest life owing to the higher stress and strain concentration and it is most likely to fail at this location. This can be explained on the basis of the presence of higher temperature gradient and lower deformation since its closeness to the fixed support leads to higher stresses.

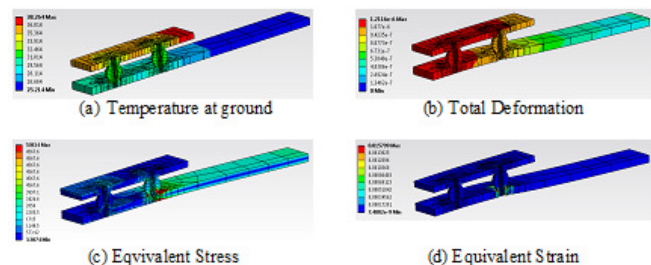


Figure 8: Typical FEA result over the PCB segment

Again more accurate RUL calculation should involve the consideration of the intermetallic layer between the solder and the substrate whose material characterization was beyond the scope of this work. The primary life critical node is at node number 4805 with a RUL of 7,041 hrs. Considering that the intermetallic layer would be highly brittle which makes the selection of a low rupture strain ($\epsilon_{rupture}$) of 0.05% to calculate RUL as the most appropriate engineering solution to this problem. This suggests that the intermetallic layer has to be embrittled to a point where creep failure is dramatically influenced by its volume fraction.

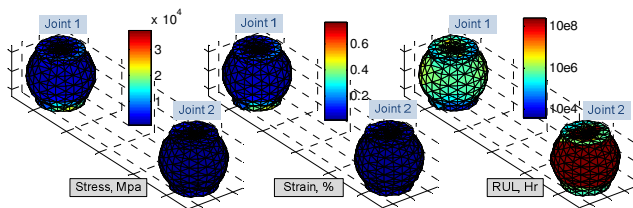


Figure 9: Spatial distribution of stress, strain and RUL over two solders joints

A range of rupture strains were used to recalculate RUL at the life critical node of 4805 along with the damage contribution of TMF and creep and tabulated in Table 1. The table shows that the contribution of the TMF is the largest towards damage accumulation in the solder joints during the normal operation of the aircraft. In-flight cyclic fluctuations will be expected to dominate the contribution to the damage accumulation process.

7.3 Probabilistic Analysis

After determining that the primary fracture critical node is at 4805 node number with 0.05% rupture strain with the RUL

Table 1: RUL and contribution at fracture critical node of 4805

$\epsilon_{rupture}$ (%)	RUL (Hrs)	% Contribution to Damage	
		TMF	Creep
0.001	4,611	61.83	38.17
0.005	6,652	89.03	10.97
0.010	7,041	94.21	5.79
0.050	7,386	98.80	1.20
0.100	7,431	99.41	0.59
0.500	7,468	99.90	0.10
1.000	7,473	99.97	0.03

being 7,386 hrs, a Monte Carlo simulation was conducted with 5,000 normally distributed random samples of microstructural grain size with mean grain size of $2\mu\text{m}$ and standard deviation of $0.40\mu\text{m}$. The Weibull distribution plot of remaining useful life calculated for the randomly distributed grain size at the primary fracture critical node is

shown in Figure 10. Since the $\beta > 1$ it suggests that the usage based failure of the solder with the Mean Time to Failure (MTTF) to be approximately around 8,000 hours service life of the solder joints. Contribution to damage from creep becomes prominent only at very low and may be unrealistic rupture strains. However, for a detailed consideration, creep damage accumulation during ground idle and time between flights should also be included.

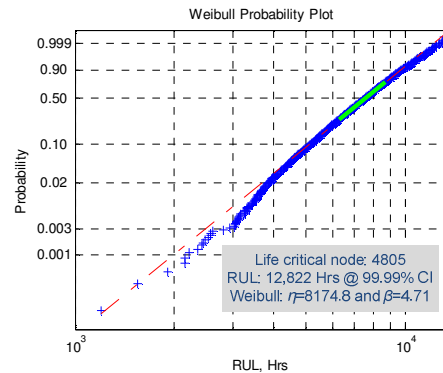


Figure 10: Two parameter Weibull distribution of RUL at life critical node of 4805

8. CONCLUSIONS AND FUTURE WORK

A prognostics based Life Cycle Management approach has been proposed to implement a physics-based prognostic system for eutectic solder joint in avionics. Realistic mission profiles and eutectic solder properties have been incorporated to calculate RUL of the solder joints in a typical PCB under combined thermal and vibratory loading conditions. Microstructure based damage models for creep and fatigue have been calibrated with the properties data for the eutectic solder. Finite Element Analysis and RUL results indicate that the contact surface between the solder and the board accumulated the highest damage thus making it the most likely failure prone zone. It is also observed that the contribution of TMF damage accumulation is dominant during the aircraft operation. The deterministic and probabilistic lifing analysis reiterates the applicability of the prognosis based LCM of solder joints. Further work towards developing more comprehensive prognosis of avionics would include the following:

- Improving the TMF and Creep models by using frequency and cavitation terms
- Extension of the prognosis of other avionics components
- Experimental validation with standardized accelerated life testing in laboratory environment

REFERENCES

- Dasgupta, A., Sharma, P., & Upadhyayula K. (2001). Micro-Mechanics of fatigue damage in Pb-Sn solder due to vibration and thermal cycling, *International Journal*

- of damage mechanics*, vol. 1, pp. 101-132.
- Gu, J., & Pecht M. (2010). Prognostics and Health Assessment Implementation for Electronics Products, *Journal of the IEST*, vol. 53, no. 1, pp. 44-58.
- Joo, D. K., Jin Y., & Shin S. W. (2003), Creep rupture of lead-free Sn-3.5Ag-Cu solders, *Journal of electronic materials*, vol. 32, No. 6, pp 541-547.
- Kalgren, P. W., Baybutt, M., Ginart, A., Minnella, C, Roemer, C. M. J., & Dabney T. (2007). Application of prognostic health management in digital electronic systems, *IEEE Aerospace Conference*, pp. 1-9.
- Koul, A. K., Bhanot, S., Tiku, A., & Junkin B. (2007). Importance of Physics-Based Prognosis for Improving Turbine Reliability: RRA 501KB Gas Turbine Blade Case Study, *ASME 2007 Power Conference*, San Antonio, Texas, USA.
- Koul, A. K., Immarigeon, J. P. & Wallace, W. (1994). Microstructural Control in Ni-Base Superalloys. In *Advances in High Temperature Structural Materials and Protective Coatings* (95-125). National Research Council of Canada, Ottawa.
- Kovacevic, I. F., Drogenik, U., & Kolar J. W. (2010). New physical model for lifetime estimation of power modules, *International Power Electronics Conference*, ETH, Zurich, pp. 2106-2114.
- Mohamed, F. A. & Langdon, T. G. (1974). Deformation Mechanism Maps Based on Grain Size. *Metallurgical Transactions*, 5, 2339-2345.
- Nasser, L., Tryon, R. & Dey A. (2005). Material simulation-based electronic device prognosis, *IEEE Aerospace Conference*, pp. 3579-3584.
- Neu, R., & Sehitoglu, H. (1989), Thermomechanical fatigue, oxidation, and Creep: Part II. Life prediction, *Metallurgical and Materials Transactions A*, vol. 20, pp. 1769-1783.
- Saha, B., Celaya, J. R., Wysocki, P. F., & Goebel K. F. (2009). Towards prognostics for electronics components, *IEEE Aerospace Conference*, Big Sky, MT, pp. 1-7.
- Shi, X. Q., Wang, Z. P., Yang, Q. J., & Pang H. L. J. (2003). Creep behavior and deformation mechanism map of Sn-Pb eutectic solder alloy, *Journal of Engineering Materials and Technology*, ASME, vol. 125, pp. 81 - 87.
- Smith, S. D. (2004). Cockpit seat and pilot helmet vibration during flight operations on aircraft carriers, *Aviation, Space, and Environmental Medicine*, vol. 75, no. 3, pp.247-254.
- Tang, H., & Basaran C. (2003). A damage mechanics-based fatigue life prediction model for solder joints, *Transactions of the ASME, Journal of electronic packaging*, vol. 125, pp.120-125.
- Wong, T. E., Lau, C. Y., & Fenger H. S. (2004). CBGA solder joint thermal fatigue life estimation by a simple method, *Soldering & Surface Mount Technology*, 16/2, 41-45.
- Wadsworth, J., Ruano, O. & Sherby, O. (2002). Denuded zones, diffusional creep, and grain boundary sliding. *Metallurgical and Materials Transactions A, Chemistry and Materials Science*, 33(2), 219-229.
- Wu, X. J., & Koul A. K. (1995). Grain boundary sliding in the presence of grain boundary precipitates during transient creep, *Metallurgical and Materials Transactions A*, vol. 26, pp. 905-914.
- Wu, X. J., Yandt, S. & Zhang, Z. (2009). A Framework of Integrated Creep-Fatigue Modelling. *Proceedings of the ASME Turbo Expo 2009* (GT2009-59087), June 8-12, Orlando, Florida, USA.
- Dr. Avisekh.Banerjee** is Senior Mechanical Engineer at Life Prediction Technologies Inc. (LPTi), Ottawa, ON. He is working on the development of a prognostic based LCM system for avionics. His area of research and interests is diagnostics, prognostics and data trending for failure detection and development of PHM framework.
- Dr. Ashok Koul** is the President of Life Prediction Technologies Inc. (LPTi), Ottawa, ON and also acts as an overall technical advisor. He has 25 years experience in the field of materials engineering and life prediction with extensive experience in managing and supervising research and development activities in gas turbine structures, materials and life cycle management strategies. Over the years he has made key contributions in identifying and applying existing as well as emerging technologies in the field of gas turbine engineering.
- Dr. Amar Kumar** has more than 25 years of research and consulting experience in the fields of structural materials characterization, development and applications. At present Dr. Kumar is working as senior research scientist in the development projects of diagnostics, prognostics and health management. Prior to joining the current assignment in 2006, he was employed with Indian Institute of Technology at Delhi, India as a teaching faculty after obtaining his PhD degree from the same Institution. Dr. Kumar has published more than 200 research papers in refereed journals, conference proceedings, and technical reports.
- Dr. Nishith Goel** is a Founder and President of Cistel Technology Inc. He served as a Technical Advisor and Member of the Scientific Staff of Nortel Networks from 1984 to 1999. Dr. Goel is the co-founder of iPine Networks, Norleaf Networks and CHIL Semiconductor. He has been a Director of Enablence Technologies Inc. and is the Chair of the Queensway-Carleton Hospital Foundation board. Dr. Goel holds an MA in Science, Electrical Engineering and a PhD in Systems Engineering from the University of Waterloo.

Point processes for bearing fault detection under non-stationary operating conditions

Pavle Boškosi¹, and Đani Juričić¹

¹ *Jožef Stefan Institute
Jamova 39 Ljubljana, Slovenia
pavle.boskoski@ijs.si
dani.juricic@ijs.si*

ABSTRACT

Bearing faults represent the most frequent mechanical faults in rotational machines. They are characterized by repetitive impacts between the rolling elements and the damaged surface. The time intervals between two impacts are directly related with the type and location of the surface fault. These time intervals can be elegantly analyzed within the framework of renewal point processes. With such an approach the fault detection and identification can be performed irrespective of the variability of rotational speed. Furthermore, we show that by analyzing the entropy of the underlying counting process by means of wavelet transform, one can perform fault detection and identification without any information about the operating conditions. The effectiveness of the approach is shown on a data-set acquired from a two-stage gearbox with various bearing faults operating under different rotational speeds and loads.

1. INTRODUCTION

According to several surveys (MRWG, 1985a, 1985b, 1985c; Albrecht, Appiarius, & Shrama, 1986) one of the most common mechanical failure are bearing faults. Consequently, a variety of techniques for detection of bearing faults have been developed in the past decades. They rely mainly on analysis of vibrational signals acquired from machines operating under constant and known operating conditions. However, such conditions are rarely met in practice. Therefore, in this paper we address the issue of bearing fault detection under variable and presumably unknown operating conditions within the framework of renewal point processes.

In the currently available approaches, fault detection under variable speed is resolved by acquiring precise information

about the current speed and load. Most common approach in such a case is time-synchronous averaging (TSA), a method which compensates for the speed fluctuations (Zhan, Makis, & Jardine, 2006; Stander & Heyns, 2005). In the same manner Parker et al. (2000) applied higher order spectra analysis for the detection of various bearing faults under different load conditions. Bartelmus and Zimroz (2009) successfully performed fault detection in multi-stage gearboxes by taking into account the information about both variations in speed and load. Although the proposed approaches give satisfactory results they heavily depend on accurate measurements of the current speed and load of the monitored gearbox.

Can bearing faults be reliably detected in spite of unknown variable load and speed conditions? Poulimenos and Fassois (2006) provided a thorough analysis on modeling and analysis of nonstationary vibration signals in time domain. Padovese (2004) gave a hybrid time-frequency approach for analyzing transient signals. Baydar and Ball (2000) performed detection of gear deterioration under different loads using instantaneous power spectrum by employing Wigner-Ville distribution (WVD). They have successfully realized fault detection of gear faults irrespective of the operating conditions.

Another way of overcoming the difficulties induced by variable operating conditions is to analyze the statistical characteristics of the produced vibrational signals. In case of bearing faults, the most informative source can be found in the distribution of the time intervals between two adjacent impacts occurring between the rolling elements and the damaged bearing surface. By doing so we can employ the framework of point processes in modeling the distribution of these times. The framework of point processes was successfully applied in the areas like modeling the neural spikes, earthquake prediction, describing environmental processes etc. However in the field of fault detection, to the best of the authors knowledge, Antoni and Randall (2003) are the only authors that tried to analyze the distribution of these interevent times by treating

Boškosi et.al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

them as an ordinary renewal process. However, their analysis was focused only on cases when bearings are operating under presumably constant and known operating conditions.

In our approach we go one step further by removing the limitation of constant and known operating conditions. Furthermore, we will show that the produced bearing vibrations may be modeled as renewal process with Inverse Gaussian interevent distribution. We will show that with such an approach one can construct an unified model for bearing fault vibrations, capable of modeling both single and multiple bearing faults. The statistical properties of the model additionally allow proper modeling under both constant and variable operating conditions. Finally, we will propose one way of performing fault detection based on the statistical characteristics of the renewal process analyzed through wavelet transform.

2. BASICS OF POINT PROCESSES

The point processes represent a segment of the theory of random processes that are most commonly used for characterizing random collections of point occurrences (Cox & Isham, 1980). In the simplest form, these points usually represent the time moments of their occurrences. This class of point processes is also known as temporal point processes.

Generally it is considered that the observed random points occur at time moments $\dots, t_1, t_2, t_3, \dots$. A point process is *simple* if all the observed points are distinct i.e. $t_i \neq t_j$ for $i \neq j$. Additionally the point process is called *orderly* if the number of points N at any moment t and interval length Δt is:

$$\lim_{\Delta t \rightarrow 0} \Pr\{N[t, t + \Delta t] > 1\} = 0. \quad (1)$$

Besides the occurrence times t and the number of points N another way of defining a point process is by the interevent times, i.e. the time between two adjacent points. Thus, the n th interevent time is defined as $T_n = t_n - t_{n-1}$.

One general goal is to derive the statistical properties of the mechanism that generates the observed random occurrences. The properties of a point process may be specified in several equivalent ways. The most common approach is to specify the non-negative number $N \in \mathbb{Z}^+$ that specifies the number of observed occurrences between time 0 and time T . Another way to specify the statistical characteristics is through the distribution of the interevent times $\{T_1, \dots, T_n\}$ where $T_i = t_i - t_{i-1}$. Finally, the approach for describing the statistical characteristics that will be used throughout this paper is based on the frequency with which the events occur around the time moment t with respect to the history of the process up to that particular moment \mathcal{H}_t . This statistical property is usually called conditional intensity function $\lambda(t, \mathcal{H}_t)$. Each of these specifications is equivalent and the most appropriate one may be used (Daley & Vere-Jones, 2003a).

For the corresponding conditional density function $f(t|\mathcal{H}_t)$ one can also define its corresponding cumulative function

$F(t|\mathcal{H}_t)$. Consequently the conditional intensity function can be defined as:

$$\lambda^*(t) = \frac{f(t|\mathcal{H}_t)}{1 - F(t|\mathcal{H}_t)}. \quad (2)$$

The denominator of (2) is also known as survivor function $s(t)$ (Vreeswijk, 2010):

$$s(t) = \Pr\{\text{event not before } t|\mathcal{H}_t\}. \quad (3)$$

The form of the conditional intensity function completely describes the underlying point process. In general, as shown in Eq. (2), this function depends on both the current time t as well as the complete point process history up to that moment \mathcal{H}_t . However, by allowing specific limitations one can define several specific types of point processes. If we let $\lambda^*(t)$ to become independent of \mathcal{H}_t , it will define a non-stationary Poisson process. A stationary version is defined by fixing the value of $\lambda^*(t) = \text{const.}$ to a specific constant that defines the rate of the underlying Poisson process. With such limitations one can readily show that the interevent times of the Poisson process are independent and distributed with exponential distribution.

A further generalization of this concept is the class of renewal point processes (Lowen & Teich, 2005). Similarly like in the Poisson process, the interevent times of such processes are independent and identically distributed (i.i.d.) but with arbitrary distribution $f(t)$ supported on semi-infinite interval $[0, +\infty)$, i.e. $f(t) = 0$ for $t < 0$. Consequently, the occurrence of a new event becomes dependent only on the time since the previous one.

One can proceed even further by removing the condition of independence of the interevent intervals. If the interevent intervals $\{X_n\}$ form a Markov chain where the length of the X_{n+1} depends only on the length of the previous interval X_n one obtains a so-called Wold process (Daley & Vere-Jones, 2003a). By modeling different transition kernels of the Markov chains one can model various types of point processes (Daley & Vere-Jones, 2003b). The form of the transition directly determines the form of the conditional intensity function (Asmussen, 2003). Therefore, one can define the most suitable transition form of the governing Markov chain that will fit the observed random process. At the same time there is an equivalent opportunity of fitting a specific form of governing chain with respect to an observed history of an arbitrary point process. Such an identification procedure can be implemented by employing well established methods from the area of hidden Markov models.

3. MODELING BEARING FAULTS USING THE POINT PROCESS FRAMEWORK

Generally, the vibrations produced by bearings with localized surface faults have been analyzed in cases of constant and known rotational speed. In such a case the generated vibrational patterns $x(t)$ can be modeled as (Antoni & Randall,

2002; Randall & Antoni, 2011):

$$x(t) = \sum_{i=-\infty}^{+\infty} A_i s(t + \Delta T_i) + n(t), \quad (4)$$

where A_i represent the amplitude of the i th impact, $s(t)$ is the excited impulse response, $n(t)$ is additive background zero-mean noise and $\Delta T_i = T_{i+1} - T_i$ represents the time between two consecutive impacts. The time period ΔT_i contains all the needed diagnostic information.

The intervals ΔT_i can be treated as interevent times of a point process. By imposing a specific distribution of these intervals we can specify a model of the generating point process. Consequently by analyzing the statistical characteristics of such a point process we can infer about the underlying bearing fault.

3.1 Point process model for localized bearing faults

Tandon and Choudhury (1999) specified the characteristic impact frequencies for different bearing surface fault as functions of bearing dimensions and rotational frequency of the rotating ring. Therefore, the interevent times T_i in the model (4) are directly related to the bearing's rotational speed. Thus, in order to model the interevent time distribution we have to specify a suitable condition intensity function. A way to model the rotational speed is by modeling the change in the rotational angle $\theta(t)$ of the rotating ring:

$$\theta(t) = \nu t + \sigma W(t), \quad (5)$$

where $W(t)$ is standard Brownian motion with normally distributed increments with zero mean and some constant variance (Matthews, Ellsworth, & Reasenberg, 2002), ν is directly related to rotational speed and σ accommodate the speed fluctuations. Thus a single evolution occurs when the angle $\theta(t)$ reaches the threshold 2π . A simple realization of such a process is shown in Figure 1. Schrödinger has shown that the distribution of the time needed for a Wiener process (5) to reach a fixed threshold a follows the Inverse Gaussian distribution (Folks & Chhikara, 1978):

$$f(t) = \frac{a}{\sigma\sqrt{2\pi t^3}} \exp\left\{-\frac{(\nu t - a)^2}{2\sigma^2 t}\right\}, \quad (6)$$

usually denoted as $t \sim IG(a/\nu, a^2/\sigma^2)$. Since the parameters ν and σ are constant in time, the resulting point process is stationary with firing rate ν .

3.2 Statistical characteristics of Inverse Gaussian renewal process

Since the Inverse Gaussian renewal process will be the basis of our model we will derive the necessary statistical properties. Besides the conditional intensity function and the interevent times distribution, a point process can be analyzed through its counting process N i.e. the probability distribution $p_N(t)$ of observing N consecutive events within a time

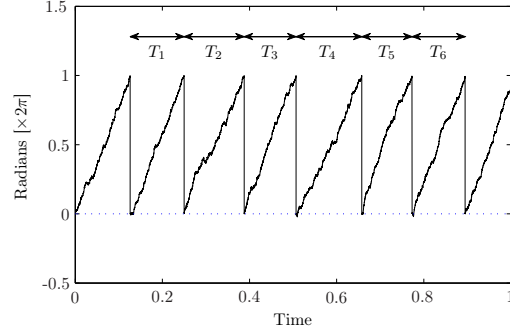


Figure 1. Realization of renewal process with Inverse Gaussian interevent distribution

interval $[t_0, t)$, where usually $t_0 = 0$. In order to derive the distribution $p_N(t)$ one has to calculate the joint probability distribution $p(t_0, t_1, \dots, t_N)$.

Firstly, the probability of a single event occurring up to time t_1 is $p_1(t_1) = p(t_1)$, where $p(t)$ is the probability distribution of a single event. The probability of observing N events up to time t_N is:

$$p_N(t_N) = \int_0^{t_N} p_{N-1}(t_{N-1})p(t_N - t_{N-1})dt_{N-1}, \quad (7)$$

where $p(t_N - t_{N-1})$ is the interevent probability distribution. The Eq. (7) is a convolution of two p.d.f. defined on the non-negative real line, since both $t_n > 0$ and $t_n > t_{n-1}$, and it can be easily calculated using the Laplace transforms of both $p_{N-1}(t)$ and the distribution of interevent times $f(t)$:

$$p_{L,N}(s) = p_{L,N-1}(s)f_L(s) = f_L^N(s), \quad (8)$$

where $p_{L,N-1}(s) = \mathcal{L}\{p_{N-1}(t)\}$, $f_L(s) = \mathcal{L}\{f(t)\}$ and $\mathcal{L}\{\cdot\}$ stands for the Laplace transform.

In case of Inverse Gaussian interevent times the Laplace transform $f_L(s)$ of (6) is:

$$f_L(s) = \exp\left\{\frac{\nu a}{\sigma^2} \left[1 - \sqrt{1 + 2\frac{\sigma^2}{\nu^2} s}\right]\right\} \quad (9)$$

Calculating then the $\mathcal{L}^{-1}\{f_L^N(s)\}$ we obtain (Tweedie, 1957):

$$f_N(t) = \frac{Na}{\sigma\sqrt{2\pi t^3}} \exp\left\{-\frac{(\nu t - Na)^2}{2\sigma^2 t}\right\}. \quad (10)$$

The obtained result has quite intuitive explanation. Namely, in (6) the threshold for the Wiener process was set at a . Therefore the time t needed to observe N consecutive crossings has the same distribution as if one elevated the threshold up to Na .

4. BEARING FAULT DETECTION USING INVERSE GAUSSIAN INTEREVENT DISTRIBUTION

Having in hand the statistical properties of the governing renewal process we can now analyze how the model performs

under different specific operating conditions. The goal of these analysis is to show that the model is valid for both constant and variable operating speed as well as in cases of single and multiple bearing faults.

4.1 Constant rotating speed

In cases when the rotating speed is strictly constant, the value of σ in (5) and (6) will become zero, hence the distribution becomes Dirac impulse i.e. $f(t; \nu, \sigma = 0) = \delta(\nu t - a)$. Consequently, the corresponding point process will be transformed into a truly periodic sequence of impacts.

(Pseudo) Cyclostationarity Small variations in the rotating speed can be accommodated by allowing small values of σ in (6). The autocorrelation function of the stationary renewal process (4) with $\Delta T \sim IG(\nu, \sigma)$ can be derived through its interevent probability distribution. Using (6) as interevent probability distribution it can be readily shown that the autocorrelation function converges to the constant value

$$\lim_{\tau \rightarrow \infty} R_{xx}(\tau) = \frac{2\sigma^2}{a\nu} < \infty. \quad (11)$$

As already analyzed by Antoni and Randall (2002), such a process can be treated as pseudo cyclostationary in cases when σ is sufficiently small, i.e. when the speed fluctuations are just a few percent.

4.2 Variable rotating speed

The modeling of completely arbitrary speed variations can be done by allowing both $\nu_{shaft} = \nu(t)$ and $\sigma_{shaft} = \sigma(t)$ in (6) to become time dependent. The resulting process is called doubly stochastic process which in essence is nonstationary process.

Despite the nonstationary characteristics, for cases where $\nu(t)$ varies slowly, one can employ the so-called modified variability measure C_{V2} . This measure is fairly insensitive to variations in the firing rate of the point process and is defined as (Ponce-Alvarez, Kilavik, & Riehle, 2010):

$$C_{V2} = \frac{2|\tau_{i+1} - \tau_i|}{\tau_{i+1} + \tau_i}, \quad (12)$$

where τ_i represents the interevent time between the events $i - 1$ and i .

4.3 Single bearing fault

A crucial information when analyzing the bearing faults is the underlying shaft speed. The instantaneous shaft speed can be obtained by differentiation of the random process (5) governing the current angle $\theta(t)$

$$\frac{d\theta(t)}{dt} = \omega_{shaft} = \nu_{shaft} + \sigma_{shaft}\eta(t), \quad (13)$$

where $\eta(t)$ is the governing Gaussian process. The rotational speed of each bearing component is directly related

to the speed of the rotating shaft (13) (Tandon & Choudhury, 1999). Consequently, each bearing fault is governed by a random process of form (13) multiplied by a constant C_k . This constant is determined by the geometrical characteristics of the bearing which determine the ratio between the angular speed of the rotating ring and a specific bearing element, i.e. $k \in \{\text{Inner ring, Outer Ring, Bearing Cage, Ball spin}\}$. Consequently, each bearing fault can be represented by a renewal process governed by Inverse Gaussian distribution with $\nu = C_k\nu_{shaft}$ and $\sigma = C_k\sigma_{shaft}$. Consequently, the distribution of the interevent times for the k th component becomes:

$$t_k \sim IG\left(\frac{a}{C_k\nu_{shaft}}, \frac{a^2}{C_k^2\sigma_{shaft}^2}\right) \quad (14)$$

4.4 Multiple faults on different bearing components

As already stated single bearing faults differ in the statistical properties of the governing IG distributions. In cases of multiple bearing faults we can observe the overall produced vibrations as a sum of several random processes each governed by its own IG probability distribution with respect to the underlying fault.

In general case the sum of IG r.v. does not necessarily leads to a result governed by IG distribution. However, the distributions (14) governing the possible bearing faults fulfill the necessary condition that the ratio

$$\frac{Var[t_k]}{E[t_k]} = \frac{a\sigma_{shaft}^2}{C_k\nu_{shaft}^3} \frac{C_k\nu_{shaft}}{a} = \frac{\sigma_{shaft}^2}{\nu_{shaft}^2} \quad (15)$$

remains constant, i.e. independent of C_k . Thus the sum of such renewal processes results into new renewal process with IG interevent distribution:

$$S = \sum_k t_k \sim IG\left(\frac{a}{\nu_{shaft}} \sum_k C_k, \frac{a^2}{\sigma_{shaft}^2} \left(\sum_k C_k\right)^2\right) \quad (16)$$

The Eq. (16) comes in hand for the cases of multiple faults. As shown by Eq. (14), distinctive distribution of interevent times governs each type of bearing fault. In such a case the observed vibrations can be regarded as a sum of several repetitive excitations of possibly different impulse responses, unlike the case of single fault as described by (4). Since such a sum fulfills the conditions (15) the resulting point process can be treated in the same manner as the cases with single fault.

5. DETECTION OF IMPACT TIMES USING WAVELET TRANSFORM

In order to apply the presented framework for bearing fault detection we should be capable of determining the times ΔT_i from (4) as precise as possible. By analyzing the bearing fault

model (4), one can observe that this signal is dominated by sudden excitations of impulse responses positioned at the impact times. The time location of these impacts can be determined sufficiently accurately by analyzing the signal with wavelet transform using a mother wavelet number of vanishing moments v_m higher than the order of the impulse response $s(t)$ in (4) (Unser & Tafti, 2010).

In such a case the selected wavelet will act as a v_m th order differential operator. Consequently, the time moments where the vibration signal $x(t)$ has discontinuities will be marked with wavelet coefficients with higher values. This time moments will coincide with the time moments when the impacts occur.

Therefore by applying Mallat (2008) thresholding process of the calculated wavelet coefficients, one can obtain accurate information about the impact times, i.e. information about the underlying bearing fault. This process is shown in Figure 2.

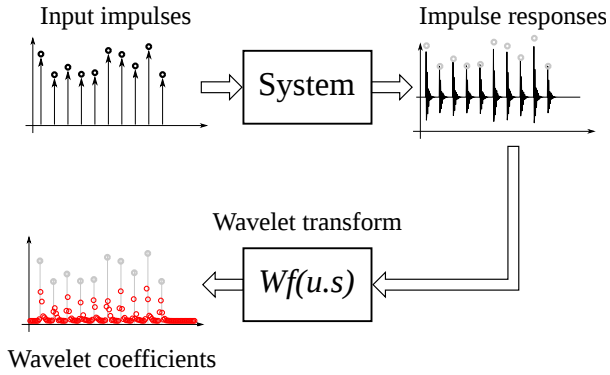


Figure 2. Wavelet as differential operator

Thorough analysis on the influence of the selection of mother wavelet on the accuracy of the decomposition for such signals has been performed by Unser and Tafti (Unser & Tafti, 2010) and Van De Ville, Blu, and Unser (Van De Ville et al., 2005). They have concluded that the crucial parameter is the number of vanishing moments of the mother wavelet rather than the selection of the “optimal” mother wavelet that will closely match the underlying process. By selecting wavelet with sufficiently high number of vanishing moments we can analyze the impulse responses $s(t)$ from (4) regardless of their variable form due to the changes of the transmission path.

5.1 Fault detection procedure

Detecting the impact moments using wavelet transform allows significant simplification in the fault detection process. The calculated wavelet coefficients preserve the statistical characteristics of the probability distribution that is generating the random impulses. Consequently, within a fixed observed window of length T one can use the distribution of the number of impacts N as information that is closely related with the underlying fault.

Due to the orthogonality of the wavelet transform the energy of the observed signal is preserved within the amplitude of the wavelet coefficients. In case when no impacts occur one will observe the wavelet transform just from the noise component $n(t)$ from (4). Therefore, under assumption of Gaussian noise, the energy will be evenly spread throughout the wavelet coefficients. Thus, the entropy of this distribution will be highest. In cases when the impacts are present the bulk of the energy of the signal will be concentrated in a small number of wavelet coefficients coinciding with the impact times, thus the entropy of the wavelet coefficient will decrease. The level of change is directly connected to the number of impulses occurring within the observed time window T . Therefore, by characterizing the distribution $p(N, T)$ of number of impacts N within a time window with length T , one can correlate the changes in the entropy of the wavelet coefficients with a particular bearing fault.

The distribution $p(N, T)$ can be determined by the survivor probability $s_N(t)$ (3). The survivor probability $s_N(t)$ gives a probability of observing the N impact time after a time moment t :

$$s_N(t) = \int_t^{+\infty} f_n(t') dt'. \quad (17)$$

Therefore the probability of observing N impulses within a time window of length T becomes

$$p(N, T) = s_{N+1}(T) - s_N(T). \quad (18)$$

By calculating the Laplace transform of (17) and inserting it in (18) the distribution becomes

$$p(N, s) = \frac{1 - f_L(s)}{s} f_L^n(s), \quad (19)$$

where $f_L(s)$ is the Laplace transform of the IG distribution as defined by Eq. (9). In order to simplify the analysis we will concentrate only on the expected number and the variance of the distribution $p(N, T)$. These values can be approximated by taking into account only a limited number of Taylor expansion terms. Hence for the expected value $E[N, T]$ and the variance $Var[N, T]$ when $f(t) \sim IG(a/\nu, a^2/\sigma^2)$ we obtain

$$\begin{aligned} E[N, T] &= \nu T + \frac{\sigma^2 \nu - 1}{2} \\ Var[N, T] &= \sigma^2 \nu^2 T \end{aligned} \quad (20)$$

As intuitively expected, these two expressions prove that the number of events within a time window depend on the firing rate ν and the variation σ .

However in case of bearing vibrations, as already shown by (16), each bearing fault differ by the factor C_k multiplying the shaft rotational speed and its fluctuation. As a result of this dependence each bearing fault is governed by different interevent distribution $f(t)$, thus the number of expected impulses within a fixed time window of size T will differ among different fault combinations. Consequently, the wavelet energy distribution will be different and the faults will be distinguishable.

Besides the changes caused by different faults, the distribution $p(N, T)$ will change with changes in the rotational speed. As a result of this change the wavelet energy entropy will vary. However, according to (16) the variations in the rotational speed will influence every bearing fault in the same manner, i.e. by adding and additional constant to each coefficient C_k in (16). Consequently, notwithstanding the variations in the speed the entropy values the entropy values for particular bearing fault will be always distinguishable, since the underlying IG distributions will remain different among various bearing faults.

6. EXPERIMENTAL RESULTS

The experimental data was acquired on a laboratory two-stage gearbox (PHM, 2009) (cf. Figure 3). The test runs include 7 different fault combinations and one fault-free reference run. From this set we have used the fault runs that contained bearing faults. Each set-up was tested under 5 different rotational speeds of the input shaft: 30, 35, 40, 45 and 50 Hz. Furthermore, two test runs were performed per each combination of different fault and speed.

The detailed list of the introduced faults is listed in Table 1. It should be noted that bearing faults were introduced only on the bearings 1–3, and all the remaining bearings were kept fault-free during the whole experimental runs. Additionally, the shaft imbalance was introduced on the *Input shaft*, whereas the sheared keyway fault was located on the *Output shaft*.

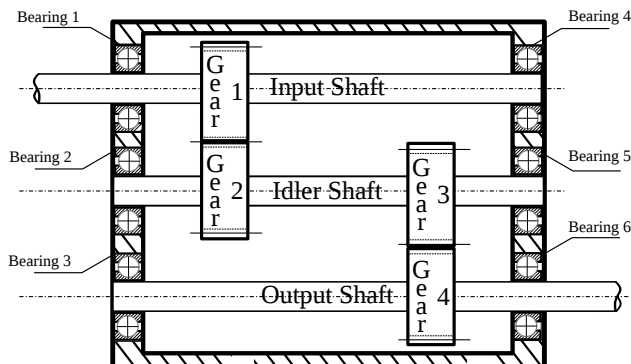


Figure 3. Schematic description of the used two-stage gearbox

6.1 Analysis

Each of the four experimental runs was analyzed using Daubechies8 mother wavelet (Daubechies, 1992). The energy entropies calculated from the corresponding wavelet coefficients are shown in Figure 4. From these results we should note the three key features.

First, the wavelet energy entropy of the fault free run is constant regardless the rotational speed. In absence of fault the observed signal reduces only to background noise $n(t)$ from Eq. (4). Since no information about the machine state is contained in this signal the entropy is constant.

Secondly, the fault 7 shows highest entropy from the other two bearing faults, followed by fault 8 and fault 6 having the lowest entropy. By examining the fault details from Table 1, one can notice that fault 7 contains only a single damaged element, fault 8 two damaged elements and fault 6 with three damaged elements. As already stated in Section 4.4, the occurrence of multiple faults can be treated as sum of several r.v. governed by IG distribution. Thus, according to (16) the resulting random process will have higher firing rate. A higher firing rate in essence contributes to increased number of expected impact occurrences N within a time window T , according to (20). Finally, this effect influences the shape of the wavelet energy distribution in such a manner that the overall entropy decreases.

According to (14), the increase of the rotational speed causes an increase in the firing rate of the IG process, hence decreasing the wavelet energy entropy. This effect has identical influence on all bearing faults. Consequently, as the speed increases the difference among wavelet energy entropies for different bearing faults increases too. Hence, the faults become more distinguishable as the rotational speed increases, as shown in Figure 4.

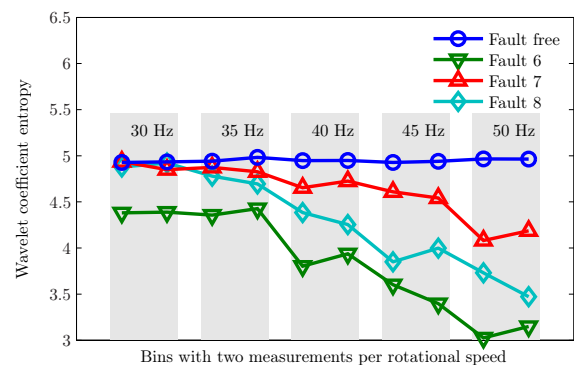


Figure 4. Wavelet coefficients energy entropy for selected bearing faults

6.2 Comments on results and possible improvements

The results support the hypothesis that bearing faults can be detected by employing a statistical model of Inverse Gaussian renewal process and wavelet energy entropy. One of the main assets of the approach is that it requires no information about the operating conditions. This becomes more evident by comparing the fault detection capabilities of this approach with approaches that incorporate information about the operating conditions. A fine example is the study that we have performed on the same experimental data by applying spec-

Run Number	Gear				Bearing ¹			Shaft fault
	1	2	3	4	1	2	3	
#1	Fault Free (FF)							
#6	FF	FF	FF	Broken	Inner	Ball	Outer	Imbalance
#7	FF	FF	FF	FF	Inner	FF	FF	Keyway Sheared
#8	FF	FF	FF	FF	FF	Ball	Outer	Imbalance

¹ Faults were introduced only on Bearings 1–3 (cf. Figure 3). The other three bearings were kept fault-free during all experimental runs. (Boškoski, Juričić, & Stankovski, 2010)

Table 1. Fault details for each experimental run

tral kurtosis (SK) and envelope analysis (Boškoski & Urevc, 2011). Although the bearing fault isolation capabilities of SK are superior, the fault detection results are comparable, i.e. the set of experimental runs containing bearing faults were accurately detected by both approaches.

Additionally this study provides a possible explanation of the results that we have obtained by the analysis of the same experimental set using a set of entropy functions calculated from the wavelet packet coefficients (Boškoski et al., 2010). Those results showed that based solely on the entropy of wavelet packet coefficients one can perform accurate fault detection of gears and bearings regardless of the operating conditions. The relations (16) and (20) provide an explanation how different bearing faults alter the probability distribution of the wavelet coefficients hence modifying its entropy.

An immediate future improvement to this study would be the application of goodness-of-fit tests that will test the hypothesis that the observed point process is governed by IG distribution. The result of such tests can serve as a starting point for deciding whether bearing faults are causing the changes in the observed probability distribution of wavelet coefficients. Furthermore, with such tests we will be able to quantify the effectiveness of the approach by considering the probability of inaccurate detection.

7. CONCLUSION

The bearing fault model based on a renewal process governed by Inverse Gaussian (*IG*) interevent has shown to be capable of modeling the fault vibrational patterns under various operating conditions. This approach provides an unified view on the statistical properties of the produced vibrational signals regardless of the operating conditions. Such a unified concept offers several advantages.

Firstly the rate ν and the variance σ of the *IG* renewal process contain all the necessary information about the present bearing fault. Furthermore, such an approach allows fairly simple modeling of multiple bearing faults, since the resulting process can be treated as a sum of inverse Gaussian random variables. As bearing faults are related to the shaft rotational speed, the necessary condition is fulfilled so the resulting sum

is again governed by Inverse Gaussian distribution.

Secondly, having defined the distribution of the renewal point process we were able to derive the probability of observing N impacts within a time window T . Thus, we have shown that by employing this distribution it is guaranteed that various bearing faults can be distinguished without any knowledge about the geometrical characteristics of the monitored bearings.

Using the distribution of the counting process N we have presented one possible way of using wavelet transform in obtaining an estimate of the number of impacts within a time T by analyzing the wavelet coefficient energy entropy. The results show that various bearing faults can be successfully detected without any knowledge about their geometrical characteristics. Additionally, the behavior of the calculated feature supports the hypothesis that the produced bearing vibrations can be treated as renewal point process with *IG* interevent distribution.

ACKNOWLEDGMENTS

The research of the first author was supported by Ad futura Programme of the Slovene Human Resources and Scholarship Fund. We also like to acknowledge the support of the Slovenian Research Agency through Research Programme P2-0001 and L2-2110.

REFERENCES

- Albrecht, P. F., Appiarius, J. C., & Shrama, D. K. (1986). Assesment of the reliability of motors in utility applications. *IEEE Transactions of Energy Conversion, EC-1*, 39-46.
- Antoni, J., & Randall, R. B. (2002). Differential Diagnosis of Gear and Bearing Faults. *Journal of Vibration and Acoustics, 124*(2), 165-171.
- Antoni, J., & Randall, R. B. (2003). A Stochastic Model for Simulation and Diagnostics of Rolling Element Bearings With Localized Faults. *Journal of Vibration and Acoustics, 125*(3), 282-289.
- Asmussen, S. (2003). *Applied Probability and Queues* (2nd ed.). New York: Springer-Verlag.

- Bartelmus, W., & Zimroz, R. (2009). A new feature for monitoring the condition of gearboxes in non-stationary operating conditions. *Mechanical Systems and Signal Processing*, 23(5), 1528–1534.
- Baydar, N., & Ball, A. (2000). Detection of Gear deterioration under varying load conditions using the Instantaneous Power Spectrum. *Mechanical Systems and Signal Processing*, 14(6), 907–921.
- Bošković, P., Juričić, Đ., & Stankovski, M. (2010). Gear and bearing fault detection under variable operating conditions. In *Annual Conference of the Prognostics and Health Management Society*.
- Bošković, P., & Urevec, A. (2011). Bearing fault detection with application to PHM Data Challenge. *International Journal of Prognostics and Health Management*.
- Cox, D. R., & Isham, V. (1980). *Point Processes*. Cambridge: Chapman and Hall.
- Daley, D., & Vere-Jones, D. (2003a). *An Introduction to the Theory of Point Processes: Elementary Theory and Methods* (Second Edition ed., Vol. 1). New York: Springer-Verlag.
- Daley, D., & Vere-Jones, D. (2003b). *An Introduction to the Theory of Point Processes: General Theory and Structure* (Vol. 2). New York: Springer-Verlag.
- Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia: SIAM.
- Folks, J. L., & Chhikara, R. S. (1978). The Inverse Gaussian Distribution and Its Statistical Application—A Review. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(3), 263–289.
- Lowen, S. B., & Teich, M. C. (2005). *Fractal Based Point Processes*. Wiley-Interscience.
- Mallat, S. (2008). *A wavelet tour of signal processing* (3rd ed.). Burlington, MA: Academic Press.
- Matthews, M. V., Ellsworth, W. L., & Reasenber, P. A. (2002). A Brownian Model for Recurrent Earthquakes. *Bulletin of the Seismological Society of America*, 92(6), 2233–2250.
- MRWG, M. reliability working group. (1985a). Report of Large Motor Reliability Survey of Industrial and Commercial Installations, Part I. *IEEE Transactions of Industry Applications*, IA-21, 853-864.
- MRWG, M. reliability working group. (1985b). Report of Large Motor Reliability Survey of Industrial and Commercial Installations, Part II. *IEEE Transactions of Industry Applications*, IA-21, 865-872.
- MRWG, M. reliability working group. (1985c). Report of Large Motor Reliability Survey of Industrial and Commercial Installations, Part III. *IEEE Transactions of Industry Applications*, IA-23, 153-158.
- Padovese, L. (2004). Hybrid time-frequency methods for non-stationary mechanical signal analysis. *Mechanical Systems and Signal Processing*, 18(4), 1047–1064.
- Parker, B. E., Ware, H. A., Wipf, D. P., Tompkins, W. R., Clark, B. R., & Larson, E. C. (2000). Fault Diagnosis using Statistical change detection in the Bispectral Domains. *Mechanical Systems and Signal Processing*, 14(4), 561–570.
- PHM. (2009). *Prognostics and Health Management Society 2009 Data Challenge*. <http://www.phmsociety.org/competition/09>.
- Ponce-Alvarez, A., Kilavik, B., & Riehle, A. (2010). Comparison of local measures of spike time irregularity and relating variability to firing rate in motor cortical neurons. *Journal of Computational Neuroscience*, 29(1–2), 351-365.
- Poulimenos, A., & Fassois, S. (2006). Parametric time-domain methods for non-stationary random vibration modelling and analysis — A critical survey and comparison. *Mechanical Systems and Signal Processing*, 20(4), 763—816.
- Randall, R., & Antoni, J. (2011). Rolling element bearing diagnostics—A tutorial. *Mechanical Systems and Signal Processing*, 25(2), 485–520.
- Stander, C., & Heyns, P. (2005). Instantaneous angular speed monitoring of gearboxes under non-cyclic stationary load conditions. *Mechanical Systems and Signal Processing*, 19(4), 817—835.
- Tandon, N., & Choudhury, A. (1999). A review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings. *Tribology International*, 32, 469-480.
- Tweedie, M. C. K. (1957). Statistical Properties of Inverse Gaussian Distributions. I. *The Annals of Mathematical Statistics*, 28(2), 362–377.
- Unser, M., & Tafti, P. (2010). Stochastic models for sparse and piecewise-smooth signals. *Signal Processing, IEEE Transactions on*, PP(99), 1–1.
- Van De Ville, D., Blu, T., & Unser, M. (2005). Isotropic polyharmonic B-splines: scaling functions and wavelets. *Image Processing, IEEE Transactions on*, 14(11), 1798–1813.
- Vreeswijk, C. van. (2010). Analysis of Parallel Spike Trains. In S. Grün & S. Rotter (Eds.), (Vol. 7, pp. 3–20). Springer.
- Zhan, Y., Makis, V., & Jardine, A. K. (2006). Adaptive state detection of gearboxes under varying load conditions based on parametric modelling. *Mechanical Systems and Signal Processing*, 20(1), 188–221.

Power Curve Analytic for Wind Turbine Performance Monitoring and Prognostics

Onder Uluyol¹, Girija Parthasarathy², Wendy Foslien³, Kyusung Kim⁴

^{1,2,3,4}*Honeywell International, Inc., Golden Valley, MN, 55422, USA*

onder.uluyol@honeywell.com
girija.parthasarathy@honeywell.com
wendy.foslien@honeywell.com
kyusung.kim@honeywell.com

ABSTRACT

The manufacturer-provided power curve for a wind turbine indicates the expected power output for a given wind speed and air density. This work presents a performance analytic that uses the measured power and the power curve to compute a residual power. Because the power curve is not site-specific, the residual is masked by it and other external factors as well as by degradation in performance of worn or failing components. We delineate operational regimes and develop statistical condition indicators to adaptively trend turbine performance and isolate failing components. The approach is extended to include legacy wind turbines for which we may not have a manufacturer's power curve. In such cases, an empirical approach is used to establish a baseline for the power curve. The approach is demonstrated using supervisory control and data acquisition (SCADA) system data from two wind turbines owned by different operators.

1. INTRODUCTION

High operations and maintenance costs for wind turbines reduce their overall cost effectiveness. One of the biggest drivers of maintenance cost is unscheduled maintenance due to unexpected failures. Using automated failure detection algorithms for continuous performance monitoring of wind turbine health can improve turbine reliability and reduce maintenance costs by detecting failures before they reach a catastrophic stage or cause damage that increases repair costs.

The power curve is a universal measure of wind turbine performance and an indicator of overall wind turbine health. Many failures and performance deterioration mechanisms can manifest in the measured power curve. By exploiting this measure with commonly collected supervisory control

and data acquisition (SCADA) system information, we can provide early indications of failures or severe performance deterioration. This paper presents an approach to wind turbine diagnostics and prognostics that uses nominal power curves and operational data.

While early indication of failure is needed, it is equally important to minimize false warnings; therefore, it is important to determine data variability measures and bounds for normal and anomalous conditions. We use several statistical measures to establish separation between normal or baseline operation and deteriorated conditions.

2. WIND TURBINE PERFORMANCE MONITORING

Performance is described in the context of the underlying process physics of the equipment—in this case, the wind turbine. Wind turbines convert wind kinetic energy into useful electrical energy. As the turbine components deteriorate, the efficiency with which wind energy is converted to electrical energy decreases and the performance of the turbine decreases. Performance degradation can indicate problems such as blade aerodynamic degradation due to leading and trailing edge losses, dirt or ice buildup on blades, drivetrain misalignment, friction caused by bearing or gear faults, generator winding faults, or even pitch control system degradation.

SCADA or operating data of equipment is often used in other industries for accurate and timely detection, diagnostics, and prognostics of failures and performance problems (Bell & Foslien, 2005, Gorinevsky, Dittmar & Mylaraswamy, 2002, Kim & Mylaraswamy, 2006). For example, in turbine engine diagnostics, failures such as turbine degradation, compressor bleed band failure, fuel supply system faults, combustion liner burn-through, and in-range sensor faults can be automatically detected with appropriate diagnostic algorithms. SCADA data provides a rich source of continuous time observations, which can be exploited for overall turbine performance monitoring. With

Onder Uluyol, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

appropriate algorithms, performance monitoring can be matured into individual component fault isolation schemes.

The functional elements of performance monitoring are shown in Figure 1. A performance parameter is computed based on sensor measurements; this parameter can be raw sensor values, sensor values corrected for environmental conditions, residuals with respect to a wind turbine model, component efficiency or aerodynamic parameters. Anomaly detection uses one or more such parameters to test whether the wind turbine is behaving within normal bounds. If the root cause of the anomaly is further classified as a particular component failure, this provides diagnosis. Additional elements involve predictive trending and prognostics, wherein parameters or fault indicators are trended and time to failure is projected.

Use of SCADA data for performance monitoring or fault diagnostics in wind turbines is not as mature as in other industries, such as process and aerospace, where condition-based maintenance (CBM) is more widespread. In some cases, SCADA data, mainly temperature (bearing or generator-winding), have been used along with vibration data for fault detection (Wiggelinkhuizen, et al. 2008, Lekou, et al. 2009). Operating data is also used just to detrend or normalize the vibration or temperature data (Wiggelinkhuizen, et al. 2008). Zaher, McArthur, and Infield (2009) presented a method to use SCADA data for anomaly detection based on neural network models of normal operating modes. The use of power curve based performance monitoring is described in (Zaher & McArthur, 2007). The power curve agent uses a power curve learned from operating data for a healthy turbine. Two pairs of alarm limits are generated: inner and outer. The inner alarm curve is based on the standard deviation for each wind speed bin added to the average in each bin. The outer alarm is chosen by the study of several turbines operating normally.

Caselitz, Giebhardt, Kruger, and Mevenkamp (1996) showed the effectiveness of utilizing spectra of the electrical power output and the vibration measurements to detect the imbalanced rotor, the aerodynamic asymmetry, and the generator bearing faults.

Kusiak presented a method to predict the anomaly, the fault severity, and the fault isolation using data mining tech

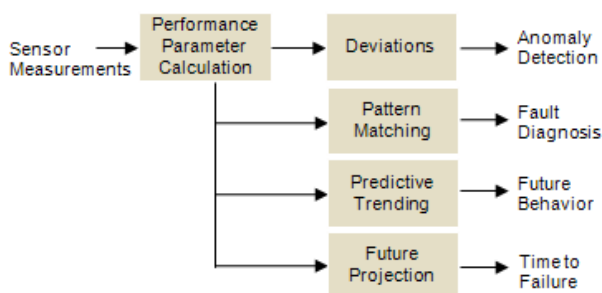


Figure 1. SCADA data-based monitoring

niques and prediction models based on wind speed and power output obtained from SCADA data (Kusiak, 2011).

Anomaly detection can be performed with a series of techniques that range from simple threshold checking to complex statistical analysis. Here, we focus on anomaly and fault detection methods for analyzing sensor data from individual wind turbines. Sensor data used in algorithm development and the approaches are described in the next sections.

3. POWER CURVE ANALYTIC

The power curve is a wind turbine performance specification provided by the manufacturer that indicates performance during operation at different wind speeds. For specific wind turbine operation, power curves are derived from non-dimensional $C_p-\lambda$ (power coefficient versus tip speed ratio) performance curves of the wind turbine design. The nominal power curves are established by the wind turbine manufacturers following published guidelines. One widely-adopted international standard is published in IEC 61400-12-1: Power performance measurements of electricity producing wind turbines (IEC, 2005). The power curve is generally used to estimate the average energy production at a particular location for a given Rayleigh wind profile and to monitor the power production performance of installed wind turbines.

Typical power curves for different air densities for a wind turbine are shown in Figure 2. The operational speed range is between the cut-in speed and the cut-out speed. The cut-in speed is the wind speed at which the turbine begins to generate power. The cut-out speed is chosen to protect the turbine and structure from high loads.

The actual power curve may deviate from the nominal one due to site-specific factors (Tindal, 2008), complex wind regimes (Rareshide, 2009), or changes in component conditions. A complex terrain, as opposed to a benign one (as defined in the standards), and different meteorological

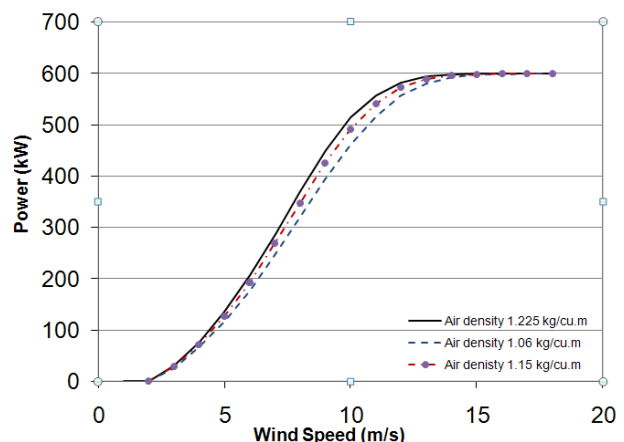


Figure 2. A typical power curve

conditions, such as varying wind direction, wind shear, and turbulence intensity can cause shifts in the power curve from the nominal.

To clearly account for factors affecting the power curve, the magnitude of the deviation from the baseline must first be assessed, and this deviation must then be further processed to generate various indicators that are relevant to different factors and critical wind turbine components.

3.1. Power Curve Generation

We use power curves provided by the manufacturer when available as the base power curve model. In the absence of a manufacturer-provided power curve (e.g., when the wind turbine is a refurbished machine or has undergone several component or control changes), SCADA data can be used to generate one. A number of data fitting approaches have been reported in the literature—from a simple polynomial fitting to a stochastic power curve generation (Milan, 2008) to a more symmetrical sigmoid function or a Gaussian CDF fitting (Yan, 2009). Since wind turbine designs and controllers are optimized for extracting maximum energy through a nonlinear phenomenon and the power coefficient C_p is not constant or symmetrical, we prefer to allow local optima instead of seeking overall symmetry. For this reason, we use polynomial fitting to generate the power curves when a manufacturer provided power curve is not available.

3.2. Power Residual Generation

The difference between the measured actual power and the power expected based on the power curve is called the power residual. Since generated power depends on the air mass as well, a family of power curves may be specified for different air densities. Hence, before we can calculate the power residual, we need to obtain the air density, which can be calculated using either of the following equations.

$$\rho = p / RT \tag{1}$$

or

$$\rho = (p_0 / RT) \exp(gz/RT) \tag{2}$$

where ρ is the air density at location in kg/m^3 , p_0 is the standard sea level atmospheric pressure, p is the air pressure in $\text{Newtons}/\text{m}^2$, T is the ambient air temperature in Kelvin, z is the location altitude in meters, and R is the specific gas constant ($287 \text{ J kg}^{-1} \text{ Kelvin}^{-1}$).

When air density, wind speed, and, in turn, the expected power are available, the power residual can be readily calculated:

$$\text{Power}_{\text{residual}} = \text{Power}_{\text{actual}} - \text{Power}_{\text{expected}} \tag{3}$$

3.3. Operational Metrics

Although the wind turbine is designed to operate between the cut-in and cut-out wind speeds, its power response to various factors discussed above is not identical across the wind speed range. Figure 3 visualizes the variation in the power residual with respect to wind speed, denoted by the blue dots. This plot illustrates the residual or power deviation of the baseline data from the power curve. Even in the case of baseline data (data used for power curve generation), there is variation in the distribution of residuals across wind speeds. The analysis presented in the following sections are based on characterizing these residual statistical metrics for the baseline and other cases—the difference in which can be visualized in plots, but need quantitative measures for automated analytics.

Notice that the variation starts small at low wind speeds, then expands in both positive and negative directions as the wind speed increases and tapers off once the rated power is reached, forming a bird-like shape which we call the Hummingbird model. To delineate the nominal and anomalous residuals with respect to the Hummingbird model, wind speed bins are defined and the standard deviation of the power residual for each bin is calculated. Three-sigma from the mean residual for each wind speed bin is used to set the upper and lower bounds on the residuals. The residual points that are outside these bounds for a particular wind speed bin are marked and used for anomaly detection as explained in the next section. Recall that the power curve shown in Figure 2 had first a concave segment followed by a convex segment. These two segments respond to increasing turbulence intensity in opposite manner—the power increases in the concave region while it decreases in the convex region as the turbulence intensity increases. Such factors determine the variability characteristics of the residuals at different wind speeds and provide a way to characterize the operational envelope.

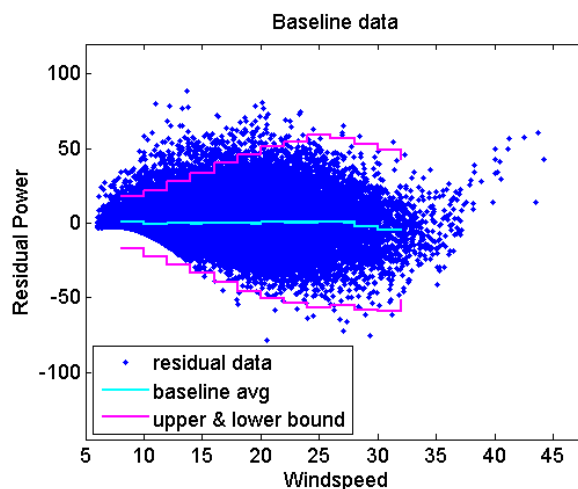


Figure 3. Power residual scatter plot of the baseline data

To model the operational envelope and be able to identify any data point that lies outside of it, Osadciw, Yan, Ye, Benson, and White used the Kaiser window fitting approach (Osadciw, et al. 2010). We prefer an industrial process control approach to define the operational parameters. This approach is naturally adaptive and easily accounts for performance changes due to normal component wear and other factors.

By adjusting the baseline period and the window size, changes in different time scales can be detected. For example, if the baseline is established using data collected from a newly installed wind turbine, any long-term changes in the turbine performance such as the deterioration of the aerodynamic performance of the rotor blades can be detected. However, using data only from a recent period to establish the baseline would mask any long-term performance degradation while exposing symptoms of an impending component failure.

In line with the standard practice of wind speed binning, we determine the power residuals for each bin and compute the corresponding bin statistics such as the mean and variance. For analysis, we also set a nominal operational boundary for each bin at some multiple of the standard deviation for that bin in the baseline data (3-sigma in this case). In Figure 3, the operational boundary is indicated by the staircase magenta lines surrounding the nominal variation (and defining the Hummingbird). Note that this operational boundary is not a ‘threshold’ in the anomaly detection sense. The n-sigma boundary provides insight into the variability of the residuals inside each bin and gives us an opportunity to characterize the shape of the residual distribution curve. This curve forms the basis for developing condition indicators that could separate nominal operation from faulty or deteriorated operation. Notice that although the Hummingbird in Figure 3 has a curvy shape, the nearly

straight horizontal line in the middle indicates that the mean power residual for the baseline remains close to zero. Also note that at this early stage of development of an algorithm, we do not characterize the power curve model as accurate or not accurate with respect to the baseline data. We characterize only the baseline residual metrics and compare these metrics with subsequent time periods, including those with failure on the horizon.

3.4. Operational Regime Based Condition Indicators

Having defined the operational boundary, we can now generate various statistics and other parametric variables that we call condition indicators (CI). The CIs can be as simple as the mean of the power residual for a wind speed bin. We can also calculate higher statistics such as skewness to measure distribution symmetry and kurtosis to see how peaked or flat a distribution we obtain for each wind speed bin. These indicators can be computed using an appropriate set of data for the baseline to detect short- and long- term changes.

4. TEST CASES

We have tested the power curve analytic approach with the SCADA data from two different wind turbines belonging to two different operators.

4.1. Data Set I

We obtained Data Set I from a mid-power wind turbine that supplies power to a university campus and sells excess power to the grid. It recently came out of 5-year warranty with the turbine manufacturer. The SCADA data is available in 10 minute and hourly intervals for 2006-2010.

Figures 4 and 5 show the power residuals plotted using the winter and summer 2008 data.

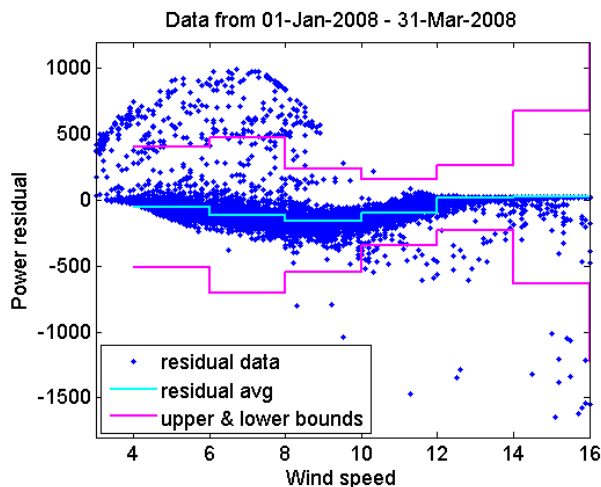


Figure 4. Power residuals in winter, 2008

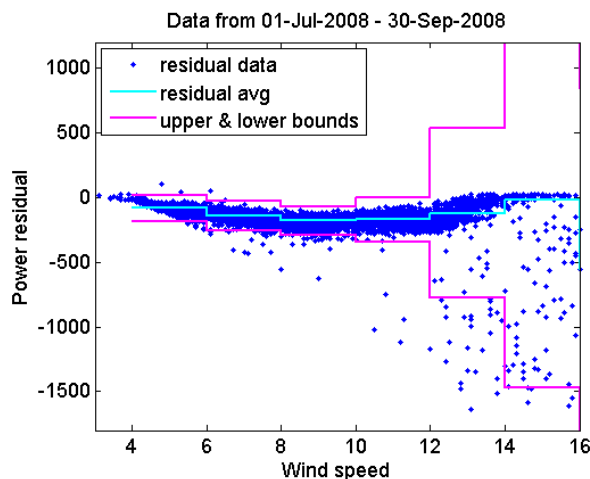


Figure 5. Power residuals in summer, 2008

Figure 6 shows a mean power residual condition indicator (CI_MPR) computed for each season in 2008 and 2009. Notice that CI_MPR is indistinguishable at low and high speeds, but it clearly shows a shift from 2008 to the next year at mid speeds. The shift indicates a noticeable improvement in the turbine performance in 2009. Unfortunately, the maintenance logs are not available from this wind turbine for us to verify the results or track the cause of the improvement to a particular maintenance action.

4.2. Data Set II

We collected Data Set II from a small, reconditioned wind turbine that provides power to the operator’s office building, and the excess power is sold to the grid. The data is available at 1-min sampling rate.

This operator encountered an issue with the gearbox during routine, semi-annual maintenance in October, 2009. The low-speed gear was moving axially on the input shaft of the gearbox. To proactively repair this condition, the gearbox had to be removed from the turbine and taken to the rebuilding facility. The gearbox was disassembled and the low-speed shaft sizing was corrected to prevent the axial movement. The gearbox was then reassembled and reinstalled in the turbine.

This maintenance event provides a good test case for the power curve analytic approach. As a first step of our analysis, the data was split up by quarter for each year. The first quarter data from 2009 was used to establish the baseline. The power residuals were generated for the remaining quarters. Notice that the CI_MPR in Figure 7, plotted as a broken yellow line, drops further away from the baseline as the wind speed increases. Although this provides an indication of anomaly, it is not yet clear whether the drop

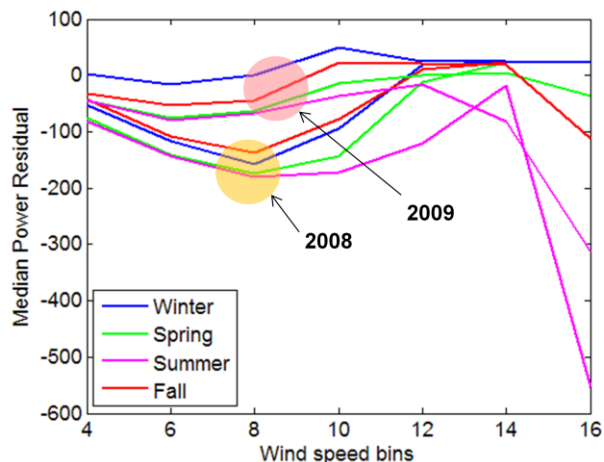


Figure 6. Improvement in WT performance at mid-range wind speeds.

in CI_MPR is the result of seasonal variations. Since we do not have many years’ worth of data, this is hard to ascertain.

Building on this first indication of an anomaly, we compute two other condition indicators: Skewness (CI_SKEW) and kurtosis (CI_KURT). Figure 8 clearly shows that the power residual symmetry as measured by the CI_SKEW for the Q3_09 is much more skewed than the other quarters. Figure 9 provides more CI_KURT evidence for the anomaly. It is clear that seasonal variations are not a consideration for either of these indicators, and any small variations between datasets are completely dominated by the indicator curve for the quarter with the failure.

The preceding analysis is based on lumped data for certain quarters. Diagnostics and prognostics depend on the underlying measurements; very exclusive sensor measurements for particular failure modes provide more accurate and earlier warnings of that failure. Since power generated is a very broad measure, how early can any such deviations from normal be detected? We performed the same analysis for moving 30-day windows with 1-day progression intervals. Figures 10 and 11 show the variation of skewness and kurtosis of residual distribution in each wind speed bin. The moving window plots started deviating from the normal around Sept 30 to Oct 3.

Notice that in Figures 8-11, the biggest difference between the suspect data sets and the baselines occur at around 24 mph. By focusing on this wind speed bin, we can take a closer look at the data to see any early indication of the impending failure.

Figures 12 and 13 show the CI_SKEW and CI_KURT for the wind speed bin at 24 mph, computed daily, with the 30-day moving windows from the days preceding the failure. The last day that the data was collected before dismantling the turbine was October 22, 2009. In the figures, several days from the earlier periods are also included for comparison.

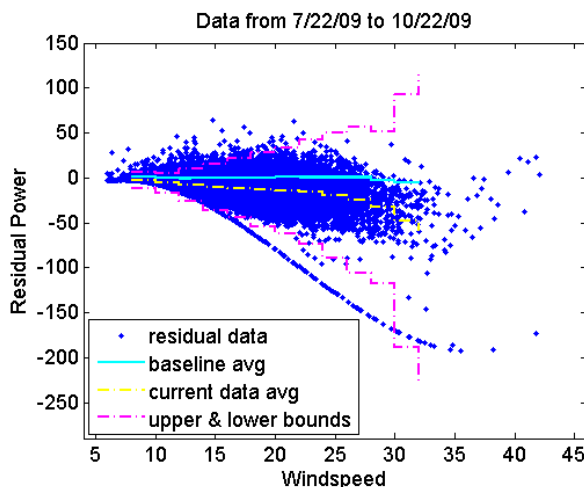


Figure 7. Power residuals in fall, 09

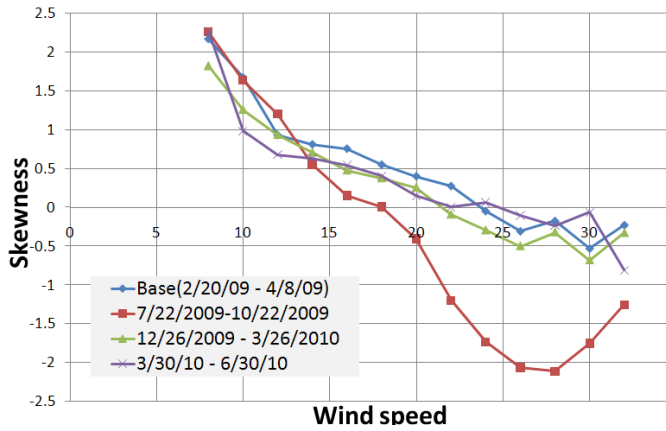


Figure 8. Skewness per quarter for each wind speed

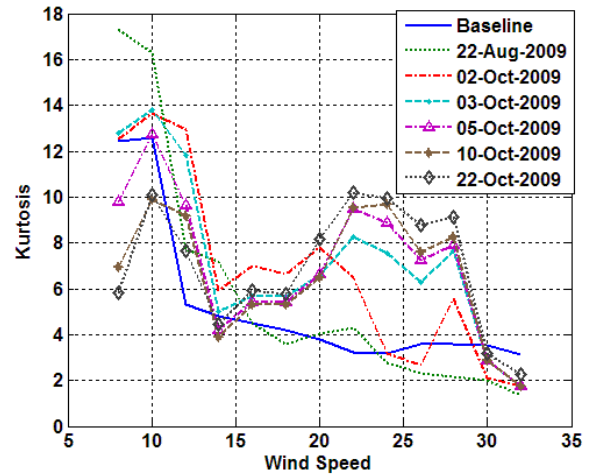


Figure 11. Kurtosis of power residual distribution in a 30-day moving window

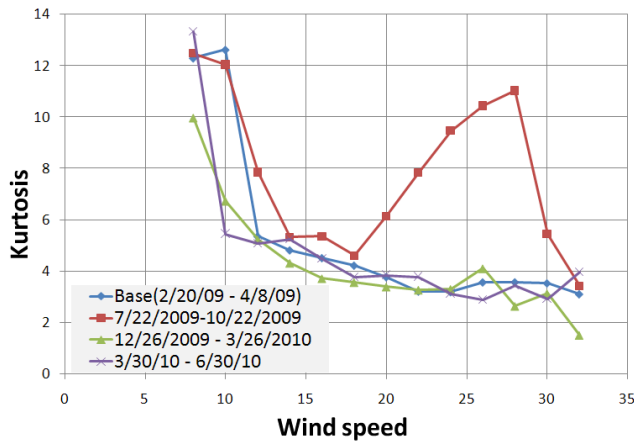


Figure 9. Kurtosis per quarter for each wind speed

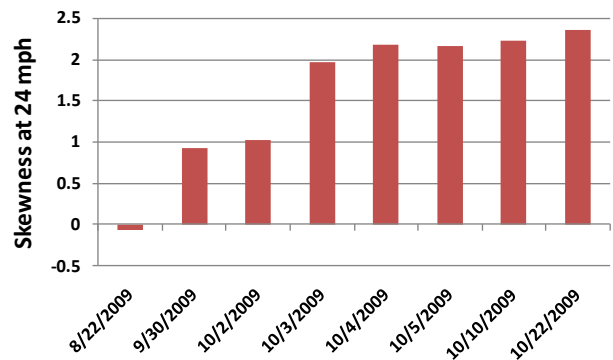


Figure 12. Skewness in days preceding the gearbox failure

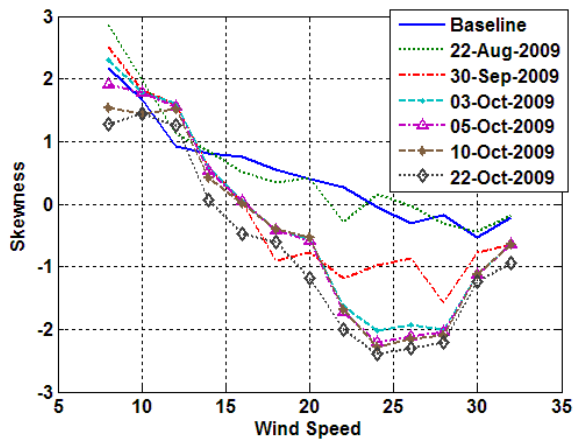


Figure 10. Skewness of power residual distribution in a 30-day moving window

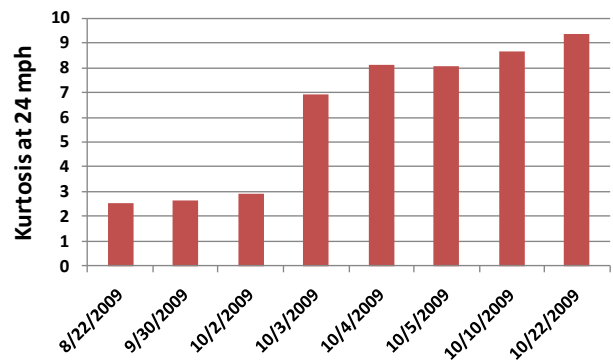


Figure 13. Kurtosis in days preceding the gearbox failure

Notice that on October 3, 2009 there is a significant rise in both CI_SKEW and CI_KURT, and the CIs remain at these new elevated levels until the failure. This shows that the first indication of the impending failure occurred about 20 days before the failure and that both indicators seem to be robust as demonstrated by the consistency in the elevated levels until the failure.

Note that these condition indicators can only quantify the wind turbine's difference in operation compared to the baseline or other periods of data. Our work analyzed the data with statistical measures to see whether the CIs capture approaching failures. At this stage, we cannot associate the anomaly to a particular failure—especially using a broad measure such as power. However, since the gearbox failure was noted and repaired and since no other major repairs or adjustments were performed during that timeframe, it is likely that the gearbox failure was manifested in the CIs. With additional data and experience, it may be possible to associate changes in CIs in particular bins to particular failure modes or operational changes.

In this gearbox failure case, the scheduled maintenance coincided with the developing failure. The operator was able to correct the problem in time and, in their own words, “it allowed us to salvage all gearing and shafts. Had the problem progressed, it would have damaged the components beyond repair and greatly increased the cost of the repair.”

5. CONCLUSION

We showed that the wind turbine power curve analytic is useful for assessing wind turbine performance and generating robust indicators for component diagnostics and prognostics. The analytic takes advantage of a universal measure of wind turbine performance with commonly collected SCADA information and provides easy configuration based on process control approaches for condition-based monitoring. Condition-based rather than hours-based maintenance enables high reliability and low maintenance costs by eliminating unnecessary scheduled maintenance.

As demonstrated in the gearbox failure case in Data Set II, early detection of an impending failure can save an operator costly repairs and long downtimes.

The wind turbine performance analytic power curve analysis method clearly separates out pre-failure data from other normal operating data. Instead of simply assigning uniform thresholds for power curve deviation, our approach uses operational regime based condition indicators. Operational regime-based CIs prevent false alarms (recognizing unique regime variabilities) and increases the possibility of fault isolation (different faults may manifest at different regimes). It emphasizes detecting slow performance degradation caused by component wear as well as degradation due to an impending failure. Condition indicators that not only take into account the variability of the power residual, but also

the distribution shape and symmetry, provide additional means of detecting and isolating failure cause.

ACKNOWLEDGMENT

We are grateful for cooperation from Great River Energy, Minnesota, Broadwind Energy Services, South Dakota, and University of Minnesota, Morris. We thank the US Department of Energy for the support for this work under Award Number DE-EE0001368.

REFERENCES

- Bell, M. B. & Foslien, W. K. (2005). Early event detection—results from a prototype implementation. In *17th Annual Ethylene Producers' Conference*, Session TA006-Ethylene Plant Process Control. Spring National Meeting (pp. 727-741), Apr. 10-14, Atlanta, GA.
- Caselitz, P., Giebardt, J., Kruger, T. & Mevenkamp, M. (1996). Development of a fault detection system for wind energy converters. *EUWEC'96* (pp1004–1007), May 2–24, Goteborg, Sweden.
- Gorinevsky, D., Dittmar, K., Mylaraswamy, D. & Nwadiogbu, E. (2002). Model-based diagnostics for an aircraft auxiliary power unit. *IEEE Conference on Control Applications*. (pp 215-220), Sept 18–20, Glasgow, Scotland.
- International Electro-technical Commission (IEC) (2005): Wind turbines—Part 12-1: Power performance measurements of electricity producing wind turbines. *IEC 61400-12-1*, first edition.
- Kim, K. & Mylaraswamy, D. (2006). Fault Diagnosis and Prognosis of Gas Turbine Engines Based on Qualitative Modeling. In *ASME TurboExpo*. (881–889), May 8–11, Barcelona, Spain.
- Lekou, D.J., Mouzakis, F., Anastasopoulou, A. A. & Kourosis, D. (2009). Fused Acoustic Emission and Vibration Techniques for Health Monitoring of Wind Turbine Gearboxes and Bearings. In *EWEC2009*.
- Milan, P. (2008). *The stochastic power curve analysis of wind turbines*. MS Thesis, University of Oldenburg, Germany.
- Kusiak, A., & Li, W. (2011). The prediction and diagnosis of wind turbine faults. *Renewable Energy* 36, pp 16-23.
- Osadciw, L. A., Yan, Y., Ye, X., Benson, G. & White, E., (2010) Wind Turbine Diagnostics based on Power Curve Using Particle Swarm Optimization. Book chapter in *Wind Power Systems: Applications of Computational Intelligence*, Wang, Lingfeng; Singh, Chanan; Kusiak, Andrew (Eds.) Springer.
- Rareshide, E., Tindal, A., Johnson, C., Graves, A., Simpson, E., Blegg, J., Harris, T., & Schoborg., D. (2009). Effects of complex wind regimes on turbine performance. Podium Presentation at the *AWEA WINDPOWER Conference*, May 4–7. Chicago, IL.
- Tindal, A., Johnson, C., LeBlanc, M., Harman, K., Rareshide, E., & Graves, A. (2008). Site-specific

- adjustments to wind turbine power curves. Poster presentation at the *AWEA WINDPOWER Conference*, June 1–4, Houston, TX.
- Wiggelinkhuizen, E., Verbruggen, T., Braam, H., Rademakers, L., Xiang, J. & Watson, S. (2008), Assessment of condition monitoring techniques for offshore wind farms, *J. Sol. Energy Eng.* 130 031004-1-9, DOI: 10.1115/1.2931512.
- Y. Yan, L. A. Osadciw, G. Benson, & E. White, (2009). Inverse data transformation for change detection in wind turbine diagnostics. *Proceedings of 22nd IEEE Canadian Conference on Electrical and Computer Engineering* (pp. 944–949), May 3–6, Delta St. John's, Newfoundland and Labrador, Canada.
- Zaher, A.S. & McArthur, S.D.J. (2007) A Multi-agent fault detection system for wind turbine defect recognition and diagnosis. In *Proceedings of Power Tech 2007*, (pp. 22–27), July 1–5, Lausanne, Switz.
- Zaher, A., McArthur, S.D.J. & Infield, D.G., (2009), Online wind turbine fault detection through automated SCADA data analysis, wind energy. Published online in Wiley Interscience DOI: 10.1002/we.319 (www.interscience.wiley.com).

Prognostics of Power MOSFETs under Thermal Stress Accelerated Aging using Data-Driven and Model-Based Methodologies

José R. Celaya¹, Abhinav Saxena², Sankalita Saha³ and Kai F. Goebel⁴

^{1,2}SGT Inc., NASA Ames Research Center, Prognostics Center of Excellence, Moffett Field, CA, 94035, USA

Jose.R.Celaya@nasa.gov
Abhinav.Saxena@nasa.gov

³MCT, NASA Ames Research Center, Prognostics Center of Excellence, Moffett Field, CA, 94035, USA

Sankalita.Saha@nasa.gov

⁴NASA Ames Research Center, Prognostics Center of Excellence, Moffett Field, CA, 94035, USA

Kai.Goebel@nasa.gov

ABSTRACT

An approach for predicting remaining useful life of power MOSFETs (metal oxide field effect transistor) devices has been developed. Power MOSFETs are semiconductor switching devices that are instrumental in electronics equipment such as those used in operation and control of modern aircraft and spacecraft. The MOSFETs examined here were aged under thermal overstress in a controlled experiment and continuous performance degradation data were collected from the accelerated aging experiment. Die-attach degradation was determined to be the primary failure mode. The collected run-to-failure data were analyzed and it was revealed that ON-state resistance increased as die-attach degraded under high thermal stresses. Results from finite element simulation analysis support the observations from the experimental data. Data-driven and model based prognostics algorithms were investigated where ON-state resistance was used as the primary precursor of failure feature. A *Gaussian process regression* algorithm was explored as an example for a data-driven technique and an *extended Kalman filter* and a *particle filter* were used as examples for model-based techniques. Both methods were able to provide valid results. Prognostic performance metrics were employed to evaluate and compare the algorithms.

1. INTRODUCTION

Power semiconductor devices such as MOSFETs (Metal Oxide Field Effect Transistors) and IGBTs (Insulated Gate Bipolar Transistors) are essential components of electronic

and electrical subsystems in on-board autonomous functions for vehicle controls, communications, navigation, and radar systems. Until very recently it was common wisdom that electronic devices fail instantly without any prior indication of failure. Therefore, current maintenance schedules are usually based on reliability data available from the manufacturer. This approach works well in aggregate on a large number of components, but, owing to the statistics, failures on individual components are not necessarily averted. For mission critical systems it is extremely important to avoid such failures. This calls for condition based prognostic health management methods. The science of prognostics is based on the analysis of failure modes, detection of early signs of wear and aging, and fault conditions. Predictions are made *in-situ* on individual in-service components. The signs of early wear are then correlated with a damage propagation model and suitable prediction algorithms to arrive at a remaining useful life (RUL) estimate.

To carry out prognostics on electronic components it is essential to understand the failure modes, their effects, and the physics of fault propagation. This requires analysis of run-to-failure data. Since more often than not current systems are not adequately instrumented to provide necessary information from electronic components to build health management algorithms, dedicated experiments are needed to fill that gap. In particular, accelerated aging allows collecting run-to-failure data in a manageable timeframe. The prognostic technique for a power MOSFET presented in this paper is based on accelerated aging of MOSFET IRF520Npbf (which comes in a TO-220 package). The aging methodology utilizes thermal and power cycling and was validated with tests using 100V power MOSFET devices. The major failure mechanism for

Celaya et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

the stress conditions is die-attachment degradation, typical for discrete devices with lead-free solder die attachment. It has been determined in these experiments that die-attach degradation results in an increase in ON-state resistance due to its dependence on junction temperature. Increasing resistance, thus, can be used as a precursor of failure for the die-attach failure mechanism under thermal stress. Data collected from these experiments were augmented by a finite element analysis simulation based on a two-transistor model. The features based on normalized ON-resistance were computed from *in-situ* measurements of the electro-thermal response. A Gaussian process regression (GPR) framework to predict time to failure was used as a data-driven prognostics technique. The extended Kalman filter (EKF) and the particle filter (PF) were used as model-based prognostics techniques based on the Bayesian tracking framework.

2. RELATED WORK

In (Saha, Celaya, Wysocki, & Goebel, 2009a) a model-based prognostics approach for discrete IGBTs was presented. RUL prediction was accomplished by using a particle filter algorithm where the collector-emitter current leakage has been used as the primary precursor of failure. A prognostics approach for power MOSFETs was presented in (Saha, Celaya, Vashchenko, Mahiuddin, & Goebel, 2011). There, the threshold voltage was used as a precursor of failure; a particle filter was used in conjunction with an empirical degradation model. The latter was based on accelerated life test data.

Identification of parameters that indicate precursors to failure for discrete power MOSFETs and IGBTs has received considerable attention in the recent years. Several studies have focused on precursor of failure parameters for discrete IGBTs under thermal degradation due to power cycling overstress. In (Patil, Celaya, Das, Goebel, & Pecht, 2009), collector-emitter voltage was identified as a health indicator; in (Sonnenfeld, Goebel, & Celaya, 2008), the maximum peak of the collector-emitter ringing at the turn of the transient was identified as the degradation variable; in (Brown, Abbas, Ginart, Ali, Kalgren, & Vachtsevanos, 2010) the switching turn-off time was recognized as failure precursor; and switching ringing was used in (Ginart, Roemer, Kalgren, & Goebel, 2008) to characterize degradation. For discrete power MOSFETs, on-resistance was identified as a precursor of failure for the die-solder degradation failure mechanism (Celaya, Saxena, Wysocki, Saha, & Goebel, 2010a; Celaya, Patil, Saha, Wysocki, & Goebel, 2009). A shift in threshold voltage was named as failure precursor due to gate structure degradation fault mode (Celaya, Wysocki, Vashchenko, Saha, & Goebel, 2010b; Saha, et al., 2011).

There have been some efforts in the development of degradation models that are a function of the usage/aging

time based on accelerated life test. For example, empirical degradation models for model-based prognostics were presented in (Saha, et al., 2009a) and (Saha, et al., 2011) for discrete IGBTs and power MOSFET respectively. Gate structure degradation modeling discrete power MOSFETs under ion impurities was presented in (Ginart, Ali, Celaya, Kalgren, Poll, & Roemer, 2010).

3. BACKGROUND

3.1. Accelerated Aging Experiments

Accelerated aging approaches provide a number of opportunities for the development of physics-based prognostics models for electronic components and systems. In particular, they allow for the assessment of reliability in a considerably shorter amount of time than running long-term reliability tests. The development of prognostics algorithms face some of the same constraints as reliability engineering in that both need information about failure events of critical electronics systems. However, these data are rarely ever available. In addition, prognostics requires information about the degradation process leading to an irreversible failure; therefore, it is necessary to record *in-situ* measurements of key output variables and observable parameters in the accelerated aging process in order to develop and learn failure progression models.

Thermal cycling overstress leads to thermo-mechanical stresses in electronics due to mismatch of the coefficient of thermal expansion between different elements in the component's packaged structure. The accelerated aging applied to the devices presented in this work consisted of thermal overstress. Latch-up, thermal run-away, or failure to turn ON due to loss of gate control were considered as the failure condition. Thermal cycles were induced by power cycling the devices without the use of an external heat sink. The device case temperature was measured and directly used as control variable for the thermal cycling application. For power cycling, the applied gate voltage was a square wave signal with an amplitude of $\sim 15V$, a frequency of 1KHz and a duty cycle of 40%. The drain-source was biased at 4Vdc and a resistive load of 0.2Ω was used on the collector side output of the device. The aging system used for these experiments is described in detail in (Sonnenfeld, et al., 2008). The accelerated aging methodology used for these experiments was presented in detail in (Celaya, et al., 2010a).

Figure 1 shows an X-ray image and a scanning acoustic image of the device after degradation. It can be observed that the die-attach solder has migrated resulting in voids. This supports the observation that the thermal resistance from junction to case has increased during the stress time resulting in increase of the junction temperature and ON-resistance ($R_{DS(ON)}$). Figure 2 presents a plot of the measured $R_{DS(ON)}$ as a function of case temperature for several

consecutive aging tests on the same device. For each test run, the temperature of the device was increased from room temperature to a high temperature setting thus providing the opportunity to characterize $R_{DS(ON)}$ as a function of time at different degradation stages. It can be observed how this curve shifts as a function of aging time, which is indicative of increased junction temperature due to poor heat dissipation and hence degraded die-attach.

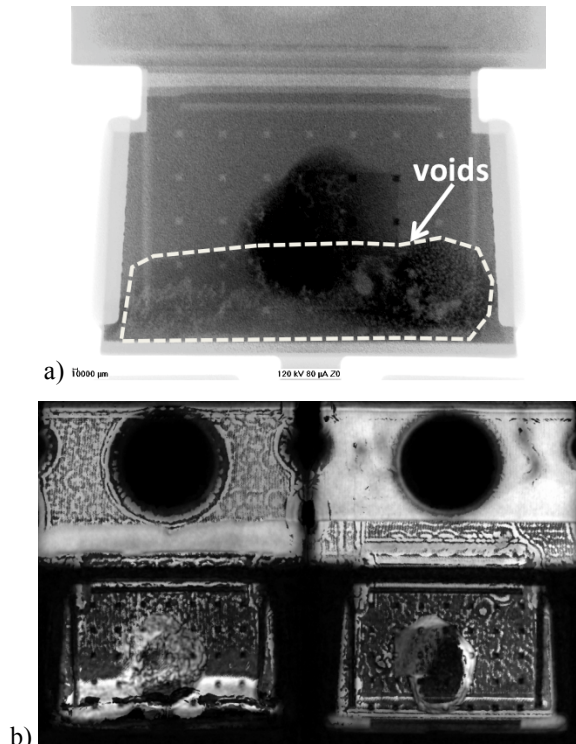


Figure 1. Failure analysis of a device after thermal overstress aging: a) X-ray microscopy of the degraded device and b) scanning acoustic microscopy of the degraded device.

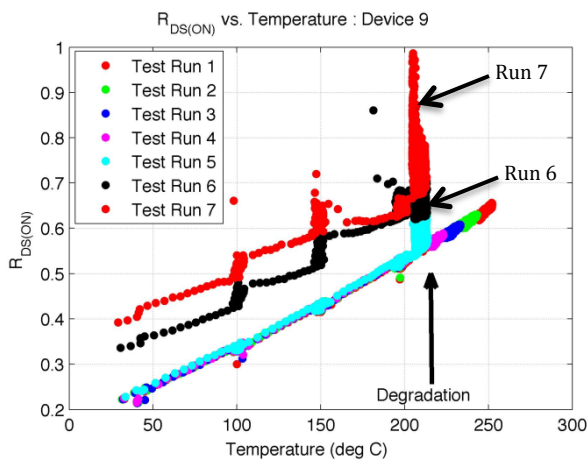


Figure 2. $R_{DS(ON)}$ degradation process due to die-attach damage.

Seven aging runs were performed in order to provide evidence of the underlying hypothesis that damage accumulates as a function of aging time and that damage rate is higher for aging under higher stress conditions like higher operating temperature. Please refer to (Celaya, et al., 2010a) for further details on the experiments.

3.2. Device Physics Modeling

In earlier work, a finite element model (FEM) was developed for a power MOSFET similar to the IRF520Npbf in order to simulate the physical phenomenon under thermal stresses. This work was originally presented in (Celaya, Saxena, Vashchenko, Saha, & Goebel, 2011b). I-V characteristics at different gate bias voltage (V_{gs}) were obtained while keeping the generic simulation parameters reasonably close to the tested MOSFETs. From the mixed-mode simulation of a single transistor model it was observed that the safe operation area (SOA) becomes limited at higher temperatures by critical voltages and currents that can be identified by the instability points in the simulation results. Please refer to (Celaya, et al., 2011b) for further details on the simulation setup and results.

The two-transistor model in figure 3 was developed to represent a device with partial die-attach degradation. The objective was to represent a degraded device of total area W_t , with two independent power MOSFET transistors with area W_1 and W_2 respectively and $W_t = W_1 + W_2$. The first transistor in the model represents the area of the device without die-attach damage and nominal thermal resistance from junction to case. The second transistor represents the area of the device with degraded die-attach and increased thermal resistance from junction to case. The second transistors runs by principle at higher temperature representative of hot spot formation on the device.

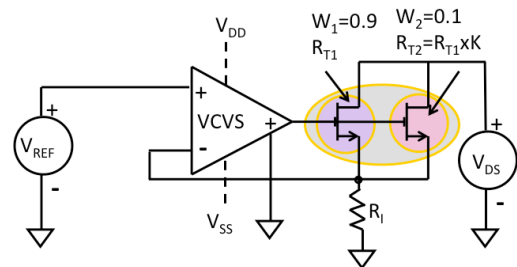


Figure 3. Two-transistor model circuit for mixed-mode simulation. Finite element models were used for each transistor.

The first transistor has original default parameters including the thermal resistance R_{T1} and area factor 90% while the second transistor depicts degradation due to electro-thermal stress represented by 10% of area with deviation of the thermal resistance coefficient scaled by the parameter K . As can be seen from the simulation results in figure 4, even a small deviation in the thermal resistance of the second transistor ($R_{T2}=KxR_{T1}$) results in significant reduction of the

critical voltage in auto bias conditions. Please refer to (Celaya, Saxena, Vashchenko, Saha, & Goebel 2011) for further details on the simulation setup and results.

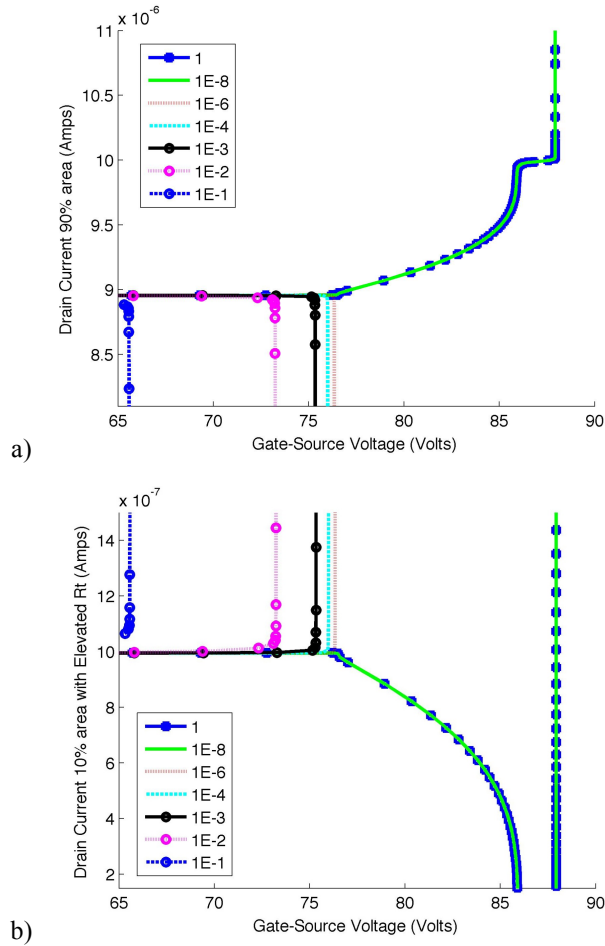


Figure 4. Results of numerical analysis for different thermal resistance parameters K of the $W_2=10\%$ second transistor model region at 450K heat sink and $R_{T2}=K \times R_{T1}$; a) nominal transistor with area W_1 , b) degraded transistor with area W_2 .

This model appears to be a good candidate for use in a physics-based degradation model. The model parameters K , W_1 and W_2 could be varied as the device degrades as a function of usage time, loading and environmental conditions. Parameter W_1 defines the area of the healthy transistors. The lower this area is, the larger is the degradation in the two-transistor model. Parameter K serves as a scaling factor for the thermal resistance of the degraded transistors. The larger this factor is, the larger is the degradation in the model. Similar to the empirical model used in this work and presented in later sections, the parameters of the two-transistor model should be estimated based on the actual fault progression dynamics.

3.3. Drain to source ON state resistance as a health state assessment parameter

In-situ measurements of the drain current (I_D) and the drain to source voltage (V_{DS}) are recorded as the device is in the aging regime and the power cycling is at 1 kHz square waveform. The ON-state resistance in this application was computed as the ratio of V_{DS} and I_D during the ON-state of the square waveform. As indicated in section 3.1, this parameter allows the observation of the die-attached degradation process and it is used in this study as a feature that reflects the state of health of the device. It is broadly understood that $R_{DS(ON)}$ increases as the junction temperature of the devices increases. In our accelerated aging setting, it is not possible to measure junction temperature directly, as a result, the increase in junction temperature is observed by monitoring the increase in $R_{DS(ON)}$ (Figure 2). Furthermore, junction temperature is also a function of the case temperature, which is also measured and recorded *in-situ*. Therefore, the measured $R_{DS(ON)}$ was normalized to eliminate the case temperature effects and reflect only changes due to degradation.

Due to manufacturing variability, the pristine condition $R_{DS(ON)}$ varies from device to device. In order to take this into account, the normalized $R_{DS(ON)}$ time series is shifted by applying a bias factor representing the pristine condition value. The resulting trajectory ($\Delta R_{DS(ON)}$) from pristine condition to failure represents the degradation process due to die-attach failure and represents the increase in $R_{DS(ON)}$ through the aging process.

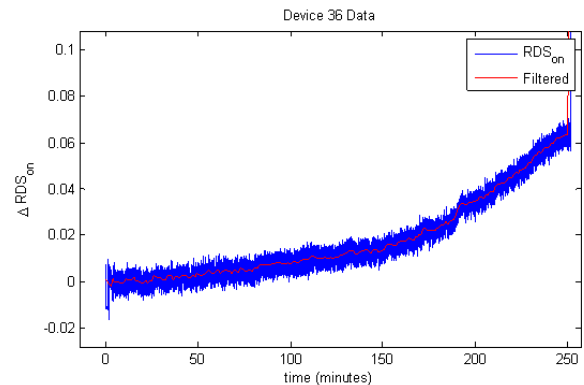


Figure 5. Normalized ON-state resistance ($\Delta R_{DS(ON)}$) and filtered trajectory for device #36.

As described earlier, these measurements are taken during the power cycle regime on the ON-state portion of the square switching signal. These measurements do not have a fixed sampling rate due to the nature of the implementation of the data acquisition system. On average, there is a transient response measurement every 400 nS. This consists of a snapshot of the transient response which includes one full square waveform cycle. The algorithms under consideration benefit from a uniform sampling in terms of ease of implementation and reduced complexity. Since the

complexity of GPR is $O(n^3)$, computational effort increases with number of data points and hence it is important to keep the number of training points low. A similar issue is also present on the EKF and PF. Therefore a resampling of the curve was carried out to have uniform sampling and a reduced sampling frequency on the failure precursor trajectory. In order to cope with these restrictions, the signals were filtered by computing the mean of every one minute long window (see Figure 5).

4. PROGNOSTICS ALGORITHMS

A prognostics algorithm in this application predicts the remaining useful life of a particular power MOSFET device at different points in time through the accelerated life of the device. Three algorithms are considered in this article, a data-driven algorithm based on the Gaussian process regression framework, and two model-based algorithms, the extended Kalman filter and the particle filter, which are based on the Bayesian estimation framework.

As indicated earlier, $\Delta R_{DS(ON)}$ is used in this study as a health indicator feature and as a precursor of failure. The prognostics problem is posed in the following way:

- A single feature is used to assess the health state of the device ($\Delta R_{DS(ON)}$).
- It is assumed that the die-attached failure mechanism is the only active degradation during the accelerated aging experiment.
- Furthermore, $\Delta R_{DS(ON)}$ accounts for the degradation progression from nominal condition through failure.
- Periodic measurements with fixed sampling rate are available for $\Delta R_{DS(ON)}$.
- A crisp failure threshold of 0.05 increase in $\Delta R_{DS(ON)}$ is used.
- The prognostics algorithm will make a prediction of the remaining useful life at time t_p , using all the measurements up to this point either to estimate the health state at time t_p in a regression framework or in a Bayesian state tracking framework.
- It is also assumed that the future load conditions do not vary significantly from past load conditions.

Six accelerated aging tests for power MOSFETs under thermal overstress were available. Figure 6 presents the $\Delta R_{DS(ON)}$ trajectories for the six cases. Cases #08, #09, #11, #12 and #14 are used for algorithm development purposes. They are used either as training data for regression models, as empirical data for degradation models or as data to quantify prior distributions' parameters of model and measurement noise and initial conditions. Case #36 is used to test the algorithms. The algorithms are developed and tested on the accelerated aging test timescale. In a real world operation, the timescale of the degradation process and therefore the RUL predictions will be considerably larger. It

is hypothesized that even though the timescale will be larger, it remains constant through the degradation process and the developed algorithms and models would still apply under the slower degradation process. On the other hand, the algorithms under consideration have been used on several other prognostics applications. Here, by using accelerated aging data with actual device measurements and real sensors (no simulated behavior), we attempted to assess how such algorithms behave under these more realistic conditions.

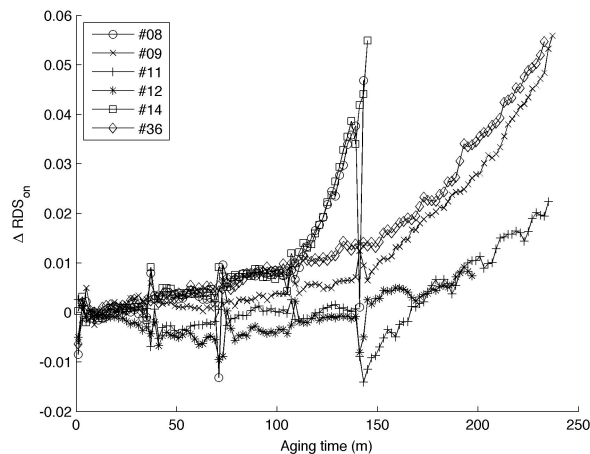


Figure 6: $\Delta R_{DS(ON)}$ trajectories for all MOSFETs, #36 is used to test algorithms and the rest are used for degradation model development and algorithm training (if required).

4.1. Degradation modeling

An empirical degradation model is suggested based on the degradation process observed on $\Delta R_{DS(ON)}$ for the five training devices. It can be seen that this process grows exponentially as a function of time and that the exponential behavior starts at different points in time for different devices. An empirical degradation model can be used to model the degradation process when a physics-based degradation model is not available. This methodology has been used for prognostics of electrolytic capacitors using a Kalman filter (Celaya, Kulkarni, Biswas, & Goebel, 2011a). There, the exponential degradation model was posed as a linear first order dynamic system in the form of a state-space model representing the dynamics of the degradation process. The proposed degradation model for the power MOSFET application is defined as

$$\Delta R_{DS(ON)} = \alpha(e^{\beta t} - 1), \quad (1)$$

where t is time and α and β are model parameters that could be static or estimated on-line as part of the Bayesian tracking framework. This model structure is capable of representing the exponential behavior of the degradation process for the different devices (see Figure 6). It is clearly observed that the parameters of the model will be different for different devices. Therefore, the parameters α and β need to be estimated online in order to ensure accuracy. This

empirical degradation model is posed as a dynamic system as follows. Let $R = \Delta R_{DS(ON)}$, then

$$\frac{dR}{dt} = R\beta + \alpha\beta. \quad (2)$$

In this model, α and β are also state variables that change through time. Therefore, the model is a non-linear dynamic system and Bayesian tracking algorithms like the extended Kalman and particle filter are needed for on-line state estimation.

4.2. Gaussian process regression

Gaussian Process Regression (GPR) is a data-driven technique that can be used to estimate future fault degradation based on training data collected from measurement data. First, a prior distribution is assumed for the underlying process function that may be derived from domain knowledge (Goebel, Saha, & Saxena, 2008). Then this prior is tuned to fit available measurements which is used with the probabilistic function for regression over the training points (Rasmussen & Williams, 2006). The output is a mean function to describe the behavior and a covariance function to describe the uncertainty. These functions can then be used to predict a mean value and corresponding variance for a given future point of interest. The behavior of a dynamic process is captured in the covariance function chosen for the Gaussian process. The covariance structure also incorporates prior beliefs of the underlying system noise. A covariance function consists of various hyper-parameters that define its properties. Proper tuning of these hyper-parameters is key in the performance. While a user typically needs to specify the type of covariance function, the corresponding hyper-parameters can be learned from training data using a gradient based optimization (or other optimization) such as maximizing the marginal likelihood of the observed data with respect to hyper-parameters (Rasmussen & Williams, 2006).

4.3. Extended Kalman filter

Extended Kalman filter allows for the implementation of the Kalman filter algorithm for on-line estimation on non-linear dynamic systems (Meinhold & Singpurwalla, 1983; Welch & Bishop, 2006). This algorithm has been used in other applications for health state estimation and prognostics (Saha, Goebel, & Christophersen, 2009b). The extended Kalman filter general form is as follows.

$$\begin{aligned} x_k &= f(x_{k-1}, u_{k-1}) + w_{k-1} \\ y_k &= h(x_k) + v_k, \end{aligned} \quad (3)$$

where f and h are non-linear equations, w_{k-1} is the model noise and v_k is the measurement noise. Noise is considered to be normally distributed with zero mean and known variance. For the prognostics implementation using the degradation model in equation (1) the state variable is

defined as $x = \{R, \alpha, \beta\}$, therefore f is a vector valued function. Equation (2) gives the state transition equation for variable R ; α and β are considered constant but they need to be estimated, therefore $\frac{d\alpha}{dt} = \frac{d\beta}{dt} = 0$ as part of the state transition function f . Measurements of R are available periodically but not for α and β . Therefore y_k will be a scalar representing the measured R at step k and h will be a scalar function defined as $h(x_k) = R$.

4.4. Particle filter

Particle filters (PFs) are based on Bayesian learning networks and are often used to track progression of system state in order to make estimations of remaining useful life (RUL). Bayesian techniques also provide a general rigorous framework for such dynamic state estimation problems. The core idea is to construct a probability density function (pdf) of the state based on all available information. In the Particle Filter (PF) approach (Arulampalam, Maskell, Gordon, & Clapp, 2002; Gordon, Salmond, & Smith, 1993) the pdf is approximated by a set of particles (points) representing sampled values from the unknown state space, and a set of associated weights denoting discrete probability masses. The particles are generated and recursively updated from a nonlinear process model that describes the evolution in time of the system under analysis, a measurement model, a set of available measurements and an a priori estimate of the state pdf. In other words, PF is a technique for implementing a recursive Bayesian filter using Monte Carlo (MC) simulations, and as such is known as a sequential MC (SMC) method.

Particle filter methods assume that the state equations can be modeled as a first order Markov process with the outputs being conditionally independent which can be written as:

$$\begin{aligned} x_k &= f(x_{k-1}) + \omega_k \\ y_k &= h(x_k) + v_k \end{aligned}$$

where, k is the time index, x denotes the state, y is the output or measurements, and both ω and v are samples from noise distributions. For this application, the PF framework was used to first track the degradation of $R_{DS(ON)}$ and then predict the remaining useful life of the power MOSFET based on whether the damage threshold has been reached by $R_{DS(ON)}$. The degradation model is presented in (1) and α and β are coefficients that are estimated initially by simple curve fitting for a few initial iterations. The PF uses the parameterized exponential growth model for $\Delta R_{DS(ON)}$, described above, for the propagation of the particles in time where the state vector is $\Delta R_{DS(ON)}$. The measurement vector comprises of the $\Delta R_{DS(ON)}$ parameters inferred from measured data. The values of α and β are learnt from regression on few initial inputs and are used as initial estimates for the filter.

5. REMAINING USEFUL LIFE PREDICTION RESULTS

This section presents the results of the three algorithms implemented. Device #36 was used to test the RUL predictions provided by the different algorithms. RUL predictions for device #36 are made at t_p : 140, 150, 160, 170, 180, 190, 195, 200, 205 and 210 minutes into aging. Subtracting the time when the prediction was made from the time when the predicted increase in resistance crosses the failure threshold gives the estimated remaining component life. As more data become available, the predictions are expected to become more accurate and more precise.

Figure 7 presents the state estimation results for $\Delta R_{DS(ON)}$ and the forecasting of $\Delta R_{DS(ON)}$ after measurements are no longer available. In this figure, measurements are available up to time t_p . They are used by all three algorithms to adjust the state estimation. The prediction step starts after t_p and time of failure $t_{EOL}=228$ hrs. A detail plot focusing around t_{EOL} is presented in Figure 8.

Analysis of the subplots from top to bottom shows how the prediction progresses as more data become available and the device gets closer to end of life. It also illustrates how prognostics is a series of periodic RUL predictions throughout the life of the device. The results as presented in Figure 7 and Figure 8 do not allow for a direct comparison among the three algorithms under consideration. Rather, it is to visually assess the estimation and prediction process. A quantitative assessment of the performance is required for direct comparison.

Figure 9 presents the α - λ performance metric for the three algorithms. This metric quantifies and visualizes the RUL prediction performance through time (Saxena, Celaya, Balaban, Goebel, Saha, Saha, & Schwabacher, 2008). The y-axis represents the estimated RUL at the time indicated in the x-axis. Ground truth RUL (RUL^*) information is used in this metric in order to assess the quality of the estimated RUL trajectories and it is identified by the 45° line in the plot. From this metric it was observed that the GPR approach is able to make predictions only at a considerably later time compared to the model-based approaches. This behavior is expected since the GPR method is data-driven and does not have the benefit of a model of the degradation process. Instead, the degradation process needs to start to get close to the elbow point of the exponential behavior in order for the prediction of RUL to become reasonably accurate. In general, the three approaches are all able to handle the RUL prediction process and predictions enter the α bound early in the life of the device. The RUL prediction results along with the prediction error are tabulated in Table 1.

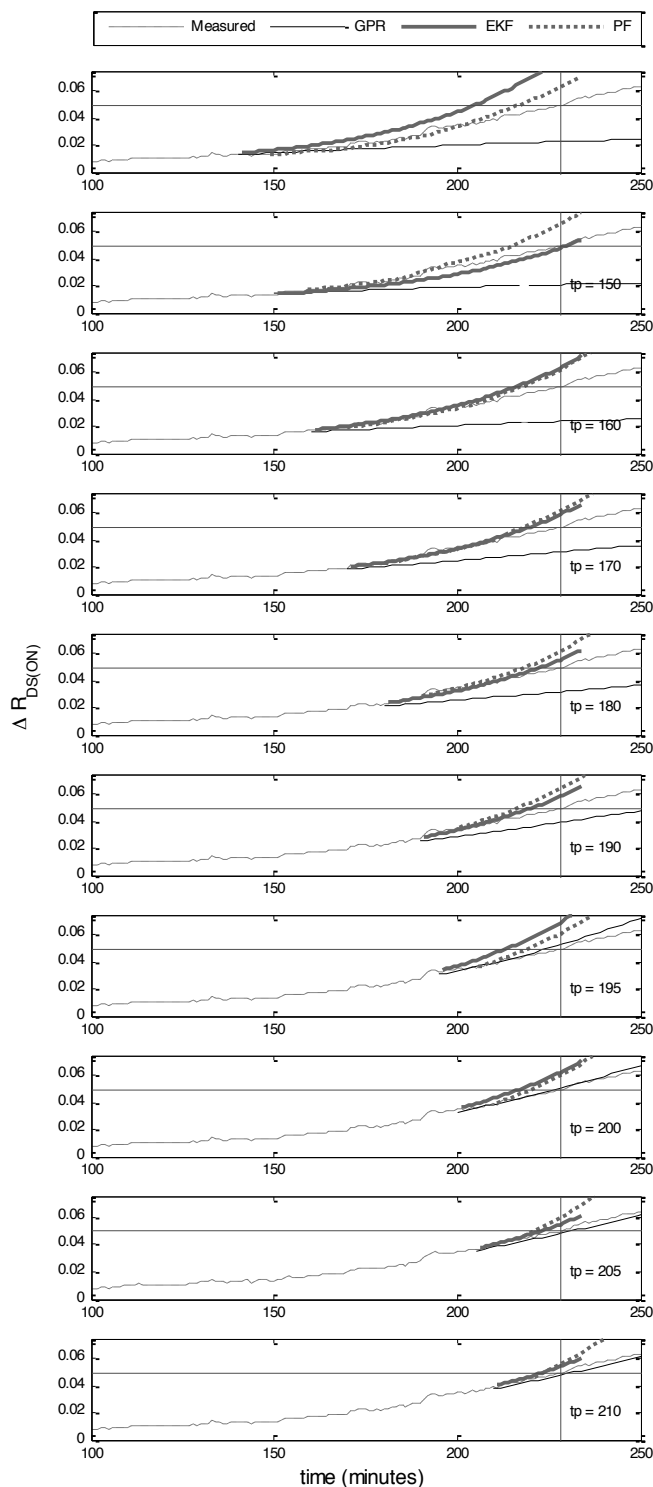


Figure 7: Health state ($\Delta R_{DS(ON)}$) tracking and forecasting for GPR, EKF and PF. Forecasting at t_p : 140, 150, 160, 170, 180, 190, 195, 200, 205 and 210 (min).

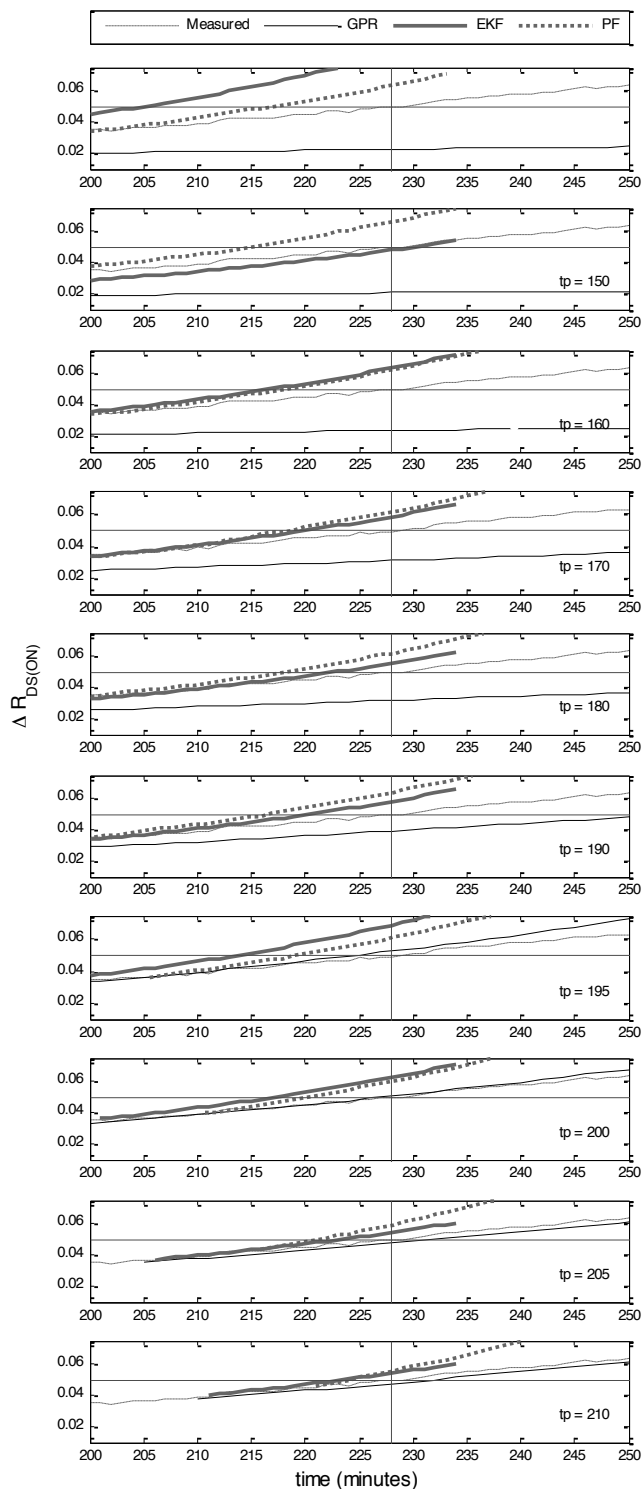


Figure 8: Detail of the health state ($\Delta R_{DS(ON)}$) tracking and forecasting for GPR, EKF and PF. Forecasting at t_p : 140, 150, 160, 170, 180, 190, 195, 200, 205 and 210 (min).

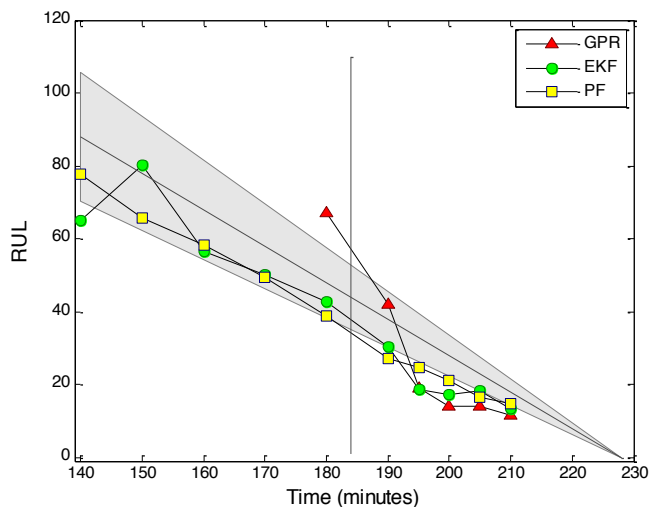


Figure 9: RUL prediction performance assessment for GPR, EKF and PF using the α - λ prognostics metric.

Table 1: RUL prediction results for GPR, EKF and PF at different t_p and $t_{EOL}=228$ hrs. RUL prediction error is between parentheses.

t_p	RUL*	GPR	EKF	PF
140	88	N/A	64.98 (23.02)	77.65 (10.35)
150	78	N/A	80.22 (-2.22)	65.85 (12.15)
160	68	N/A	56.64 (11.36)	58.33 (9.67)
170	58	N/A	50.15 (7.85)	49.47 (8.53)
180	48	73.2 (-25.2)	42.75 (5.25)	38.68 (9.32)
190	38	33.4 (4.6)	30.35 (7.65)	27.14 (10.86)
195	33	17.6 (15.4)	18.57 (14.43)	24.76 (8.24)
200	28	14.6 (13.4)	17.24 (10.76)	21.09 (6.91)
205	23	13.8 (9.2)	18.28 (4.72)	16.66 (6.34)
210	18	11.8 (6.2)	13.46 (4.54)	14.68 (3.32)

6. CONCLUSION

The paper reports on a case study of employing data-driven and model-based techniques for the prediction of remaining life of power MOSFETs. Several strong assumptions were made that need to be challenged in order to make the proposed process practical for field use. For instance, the future operational conditions and loading of the device are

considered constant at the same magnitudes as the loads and conditions used during accelerated aging. In addition, the algorithm development is conducted using accelerated life test data. In real world implementation, the degradation process of the device would occur in a considerably larger time scale. Determining the relationship between signatures from accelerated aging and signatures from “natural” aging is a topic of future work.

The algorithms considered in this study have been used as prognostics algorithms in different applications and are regarded as suitable candidates for component level prognostics. This work attempts to further the validation of such algorithms by presenting them with real degradation data including measurements from real sensors, which include all the complications (noise, bias, etc.) that are regularly not captured on simulated degradation data.

The *in-situ* data available for empirical degradation model development could be used to assess the two-transistor model parameters on an on-line tracking framework. The two-transistor model has the added advantage of being suitable to be included along the dynamics of the subsystem or system level. For instance, if the device is part of a power supply, the two-transistor model could be used as part of the whole power supply transfer function, therefore generating a system-level physics-based model with degradation parameters linked to the die-attach degradation process.

ACKNOWLEDGEMENT

This work was funded by the NASA Aviation Safety Program, projects IVHM and SSAT.

NOMENCLATURE

RUL	remaining useful life
$R_{DS(ON)}$	ON-state drain to source resistance
SOA	safe operation area of the power MOSFET
K	scaling factor for thermal resistance on the two-transistor model
W_1	area of nominal transistor in the two-transistor model
W_2	area of degraded transistor in the two-transistor model
R_{T1}	junction to case thermal resistance of the nominal transistor in the two-transistor model
R_{T2}	junction to case thermal resistance of degraded transistor in the two-transistor model
$\Delta R_{DS(ON)}$	normalized deviation in ON-resistance from drain to source
t_p	time of RUL prediction
t_{EOL}	time of end of life (time of failure)
I_D	drain current
V_{DS}	drain to source voltage
RUL*	ground truth for RUL

REFERENCES

- Arulampalam, S., Maskell, S., Gordon, N. J., & Clapp, T. (2002). A tutorial on particle filters for on-line non-linear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2), 174-188.
- Brown, D., Abbas, M., Ginart, A., Ali, I., Kalgren, P., & Vachtsevanos, G. (2010). *Turn-off Time as a Precursor for Gate Bipolar Transistor Latch-up Faults in Electric Motor Drives*. Paper presented at the Annual Conference of the Prognostics and Health Management Society 2010.
- Celaya, J., Kulkarni, C., Biswas, G., & Goebel, K. (2011a). *Towards Prognostics of Electrolytic Capacitors*. Paper presented at the AIAA Infotech@Aerospace, St. Louis, MO.
- Celaya, J., Saxena, A., Wysocki, P., Saha, S., & Goebel, K. (2010a). *Towards Prognostics of Power MOSFETs: Accelerated Aging and Precursors of Failure*. Paper presented at the Annual Conference of the Prognostics and Health Management Society 2010.
- Celaya, J. R., Patil, N., Saha, S., Wysocki, P., & Goebel, K. (2009). *Towards Accelerated Aging Methodologies and Health Management of Power MOSFETs (Technical Brief)*. Paper presented at the Annual Conference of the Prognostics and Health Management Society 2009.
- Celaya, J. R., Saxena, A., Vashchenko, V., Saha, S., & Goebel, K. (2011b). *Prognostics of Power MOSFET*. Paper presented at the 23rd International Symposium on Power Semiconductor Devices & IC's (ISPSD), San Diego, CA.
- Celaya, J. R., Wysocki, P., Vashchenko, V., Saha, S., & Goebel, K. (2010b). *Accelerated aging system for prognostics of power semiconductor devices*. Paper presented at the 2010 IEEE AUTOTESTCON.
- Ginart, A., Roemer, M., Kalgren, P., & Goebel, K. (2008). *Modeling Aging Effects of IGBTs in Power Drives by Ringing Characterization*. Paper presented at the IEEE International Conference on Prognostics and Health Management.
- Ginart, A. E., Ali, I. N., Celaya, J. R., Kalgren, P. W., Poll, S. D., & Roemer, M. J. (2010). *Modeling SiO₂ Ion Impurities Aging in Insulated Gate Power Devices Under Temperature and Voltage Stress*. Paper presented at the Annual Conference of the Prognostics and Health Management Society 2010.
- Goebel, K., Saha, B., & Saxena, A. (2008). *A Comparison of Three Data-Driven Techniques for Prognostics*. Paper presented at the Proceedings of the 62nd Meeting of the Society For Machinery Failure Prevention Technology (MFPT).
- Gordon, N. J., Salmond, D. J., & Smith, A. F. M. (1993). Novel Approach to Nonlinear/Non-Gaussian

Bayesian State Estimation. *IEE Proceedings Radar and Signal Processing*, 140(2), 107-113.

- Meinhold, R. J., & Singpurwalla, N. D. (1983). Understanding the Kalman Filter. *The American Statistician*, 37(2), 123-127.
- Patil, N., Celaya, J., Das, D., Goebel, K., & Pecht, M. (2009). Precursor Parameter Identification for Insulated Gate Bipolar Transistor (IGBT) Prognostics. *IEEE Transactions on Reliability*, 58(2), 276.
- Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*.
- Saha, B., Celaya, J. R., Wossocki, P. F., & Goebel, K. F. (2009a). *Towards prognostics for electronics components*. Paper presented at the Aerospace conference, 2009 IEEE.
- Saha, B., Goebel, K., & Christophersen, J. (2009b). Comparison of prognostic algorithms for estimating remaining useful life of batteries. *Transactions of the Institute of Measurement and Control*, 31(3-4), 293-308. doi: 10.1177/0142331208092030
- Saha, S., Celaya, J. R., Vashchenko, V., Mahiuddin, S., & Goebel, K. F. (2011). *Accelerated Aging with Electrical Overstress and Prognostics for Power MOSFETs*. Paper presented at the IEEE EnergyTech 2011.
- Saxena, A., Celaya, J., Balaban, E., Goebel, K., Saha, B., Saha, S., & Schwabacher, M. (2008, 6-9 Oct. 2008). *Metrics for evaluating performance of prognostic techniques*. Paper presented at the Prognostics and Health Management, 2008. PHM 2008. International Conference on.
- Sonnenfeld, G., Goebel, K., & Celaya, J. R. (2008). *An agile accelerated aging, characterization and scenario simulation system for gate controlled power transistors*. Paper presented at the IEEE AUTOTESTCON 2008.
- Welch, G., & Bishop, G. (2006). An Introduction to the Kalman Filter (TR 95-041): Department of Computer Science, University of North Carolina at Chapel Hill.

BIOGRAPHIES

José R. Celaya is a research scientist with SGT Inc. at the Prognostics Center of Excellence, NASA Ames Research Center. He received a Ph.D. degree in Decision Sciences and Engineering Systems in 2008, a M. E. degree in Operations Research and Statistics in 2008, a M. S. degree in Electrical Engineering in 2003, all from Rensselaer Polytechnic Institute, Troy New York; and a B. S. in Cybernetics Engineering in 2001 from CETYS University, México.

Abhinav Saxena is a Research Scientist with SGT Inc. at the Prognostics Center of Excellence NASA Ames Research Center, Moffett Field CA. His research focus lies in developing and evaluating prognostic algorithms for engineering systems using soft computing techniques. He is a PhD in Electrical and Computer Engineering from Georgia Institute of Technology, Atlanta. He earned his B.Tech in 2001 from Indian Institute of Technology (IIT) Delhi, and Masters Degree in 2003 from Georgia Tech. Abhinav has been a GM manufacturing scholar and is also a member of IEEE, AAAI and ASME.

Sankalita Saha is a research scientist with Mission Critical Technologies at the Prognostics Center of Excellence, NASA Ames Research Center. She received the M.S. and Ph.D. degrees in Electrical Engineering from University of Maryland, College Park in 2007. Prior to that she obtained her B.Tech (Bachelor of Technology) degree in Electronics and Electrical Communications Engineering from the Indian Institute of Technology, Kharagpur in 2002.

Kai Goebel received the degree of Diplom-Ingenieur from the Technische Universitt Mnchen, Germany in 1990. He received the M.S. and Ph.D. from the University of California at Berkeley in 1993 and 1996, respectively. Dr. Goebel is a senior scientist at NASA Ames Research Center where he leads the Diagnostics and Prognostics groups in the Intelligent Systems division. In addition, he directs the Prognostics Center of Excellence and he is the Associate Principal Investigator for Prognostics of NASA's Integrated Vehicle Health Management Program. He worked at General Electric's Corporate Research Center in Niskayuna, NY from 1997 to 2006 as a senior research scientist. He has carried out applied research in the areas of artificial intelligence, soft computing, and information fusion. His research interest lies in advancing these techniques for real time monitoring, diagnostics, and prognostics. He holds 15 patents and has published more than 200 papers in the area of systems health management.

Structural Integrity Assessment Using In-Situ Acoustic Emission Monitoring

Masoud Rabiei¹, Mohammad Modarres², and Paul Hoffman³

¹*Impact Technologies, Rochester, NY, 14623, USA*
masoud.rabiei@impact-tek.com

²*University of Maryland, College Park, MD, 20742, USA*
modarres@umd.edu

³*NAVAIR 4.3.3 Structures Division, Patuxent River, MD 20670, USA*
paul.hoffman@navy.mil

ABSTRACT

The work presented in this paper is focused on monitoring fatigue crack growth in metallic structures using acoustic emission (AE) technology. Three different methods are proposed to utilize the information obtained from in-situ monitoring for structural health management.

Fatigue crack growth tests with real-time acoustic emissions monitoring are conducted on CT specimens made of 7075 aluminum. Proper filtration of the resulting AE signals reveals a log-linear relationship between fracture parameters (da/dN and ΔK) and select AE features; a flexible statistical model is developed to describe the relationship between these parameters.

Bayesian inference is used to estimate the model parameters from experimental data. The model is then used to calculate two important quantities that can be used for structural health management: (a) an AE-based instantaneous damage severity index, and (b) an AE-based estimate of the crack size distribution at a given point in time, assuming a known initial crack size distribution.

Finally, recursive Bayesian estimation is used for online integration of the structural health assessment information obtained from AE monitoring with crack size estimates obtained from empirical crack growth model. The evidence used in Bayesian updating includes observed crack sizes and/or crack growth rate observations.

¹This research was conducted while the author was a graduate research assistant at the Center for Risk and Reliability at University of Maryland, College Park.

Rabiei et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

1. INTRODUCTION

Acoustic emissions are elastic stress waves generated by a rapid release of energy from localized sources within a material under stress (Mix 2005). Acoustic emissions often originate from defect-related sources such as permanent microscopic deformation within the material and fatigue crack extension.

Despite significant improvements in AE technology in recent years, quantitative interpretation of the AE signals and establishing a correlation between them and the source events remains a challenge and a topic for active research. In recent years, AE research has focused on two main areas; the first area has to do with characterizing the wave propagation through complex geometries which has proved to be an extremely difficult problem. The second area of research is concerned with processing the AE waveforms in an intelligent way (depending on the application) in order to extract useful information that can be traced back to the source event (Holford et al. 2009). The approach presented in this paper is in line with the second area.

In the first part of this paper, the problem of monitoring fatigue crack growth using AE technique is investigated. A statistical model is developed that correlates important crack growth parameters, i.e., crack growth rate, da/dN , and stress intensity factor range, ΔK , with select AE features. Next, this model will be used to calculate two important quantities that can be used for structural health management: (a) an AE-based instantaneous damage severity index, and (b) an AE-based estimate of the crack size distribution at a given point in time, assuming a known initial crack size distribution. Finally, the outcome of the statistical model described above will be used as direct “evidence” in a recursive Bayesian estimation framework to

update the model parameters as well as the estimated crack size distribution.

2. CRACK GROWTH MONITORING USING ACOUSTIC EMISSION

Fatigue crack growth is a well-known source of acoustic emission inside materials. Several researchers have studied the connection between fatigue crack growth behavior and the resulting acoustic emissions (Hamel et al. 1981; Bassim et al. 1994). Certain features of acoustic emission signals are found to be stochastically correlated with key fatigue parameters, such as stress intensity factor range, ΔK , and crack growth rate, da/dN . Two of the most commonly used AE parameters in fatigue are the AE count c and its derivative, count rate dc/dN . For a given AE signal, c is defined as the number of times that the signal amplitude exceeds a predefined threshold value. Accordingly, dc/dN is defined as the derivative of c with respect to time (measured as elapsed fatigue cycles).

The following form has been proposed by (Bassim et al. 1994) for the relationship between dc/dN and ΔK :

$$\frac{dc}{dN} = A_1(\Delta K)^{A_2} \quad (1)$$

where A_1 and A_2 are the model parameters. Our goal is to use the AE parameter as the predictor to estimate the fatigue parameter; therefore, Eq. (1) is solved for ΔK and linearized as follows (Rabiei et al. 2009):

$$\log \Delta K = \alpha_1 \log \left(\frac{dc}{dN} \right) + \alpha_2 \quad (2)$$

where $\alpha_1 = A_1^{-1/A_2}$ and $\alpha_2 = 1/A_2$ are the new model constants to be estimated from data.

The significance of Eq. (2) is that once the model parameters are determined experimentally, this equation can be used to estimate ΔK by monitoring the acoustic emissions and extracting the dc/dN parameter from the observed signals—thus obviating the need for complex modeling and calculations used in fracture mechanics to calculate ΔK .

The second parameter that will be estimated via AE monitoring is the crack growth rate, da/dN . Based on the Paris equation (Paris & Erdogan 1963), da/dN is expected to have a log-linear relationship with ΔK while the crack growth is in the stable region. According to Eq. (2), ΔK itself has a log-linear relationship with dc/dN , which results in the following equation:

$$\log \left(\frac{da}{dN} \right) = \beta_1 \log \left(\frac{dc}{dN} \right) + \beta_2 \quad (3)$$

where β_1 and β_2 are the model parameters that describe the log-linear relationship between da/dN and dc/dN . From a structural monitoring perspective, this relationship means

that on average, the rate of crack growth can be estimated solely based on features extracted from AE signals. This is a significant outcome because by knowing the rate of the crack growth and the initial crack size, the size of the crack can be estimated at any given time without knowing the specific load history or complex ΔK calculations. This fact will be used to develop an AE-based crack growth model that can predict the crack size as a function of observed AE signals.

2.1. Experimental test setup and procedure

A series of experiments were designed to validate the proposed relationship in Eqs. (2) and (3) and to generate the experimental data required for fitting the statistical model that will be introduced in the next section.

The experiments consisted of two separate parts that ran in parallel: the first part is a standard fatigue crack growth test in which a notched aluminum specimen undergoes cyclic loading, which causes a crack to initiate from the notch and grow until fracture; the second part is real-time AE monitoring—on the same specimen and while the crack is growing—to capture the AE signals resulting from the propagation of the crack inside the material.

Fatigue tests were carried out on standard compact tension (CT) specimens (ASTM E647-08 2008) made of 7075 aluminum alloy. The test setup is shown in Figure 1. The goal of the experiment was to record the AE signals generated by fatigue crack growth. To do so, we used a PCI-2 AE monitoring system supplied by Physical Acoustic Corporations¹ to monitor the CT specimen during the crack growth test. The most crucial step in AE monitoring is to distinguish the AE signals originating from the source event of interest (e.g. crack tip) from extraneous noises.

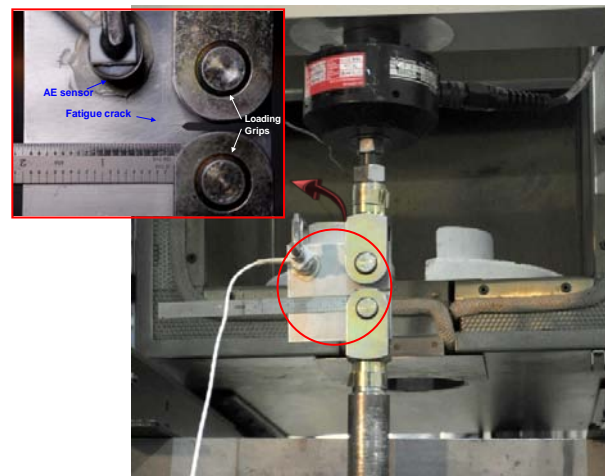


Figure 1: CT specimen instrumented with AE sensor and mounted on MTS machine

¹ <http://www.pacndt.com>

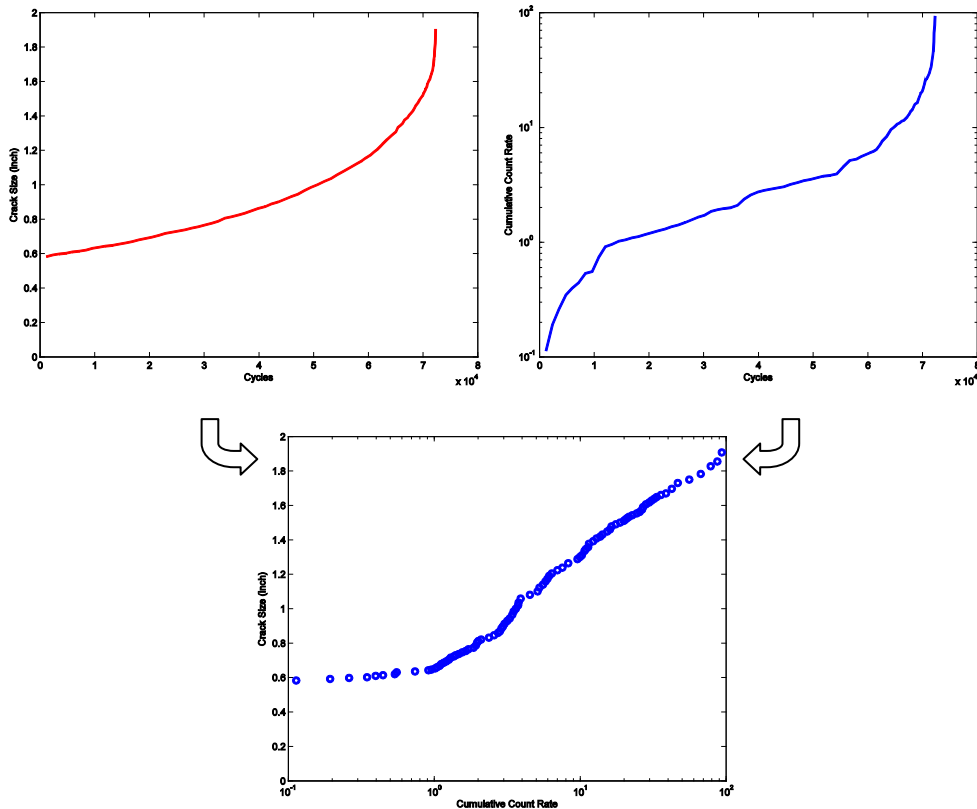


Figure 2: Cumulative AE count rate versus crack size

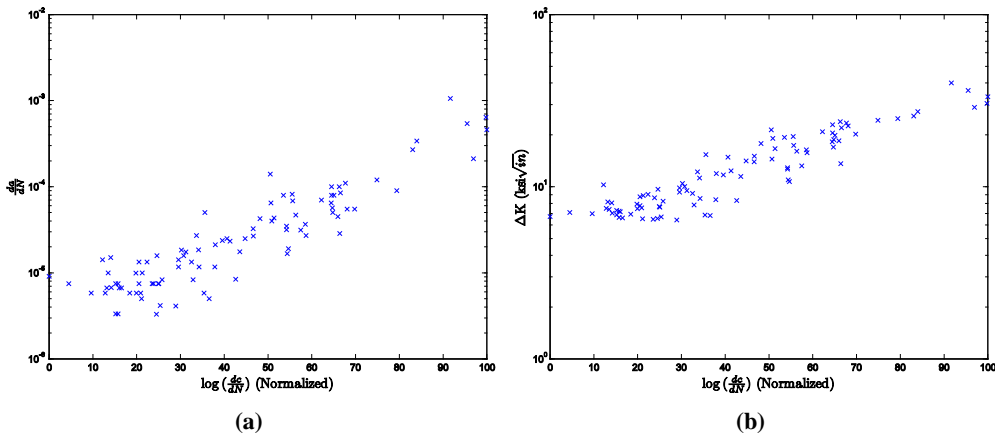


Figure 3: The linear correlation observed between dc/dN and da/dN (a) and ΔK (b) in a crack growth test

The source of the noise can be both internal (e.g., surface rubbing at loading pins, internal rubbing of crack surfaces) and external (e.g., noise from the hydraulic loading actuators). Various de-noising techniques were used to distinguish AE signals from the background noise. See (Rabiei 2011) for detailed information about the fatigue test setup, crack measurement technique and proper AE filtrations in crack growth monitoring.

Once proper filtration has been applied to the signals, the correlation between AE and crack growth parameters can be seen. Figure 2 shows that the increasing trend in crack size has a linear relationship with the cumulative AE count rate (on a log scale) for cracks larger than 0.6 inches. This suggests that in theory, the crack size can be measured by monitoring the cumulative AE count rate, if the relationship between the two is fully characterized and modeled.

Figure 3 shows the correlation between the AE parameter, dc/dN , and the fatigue parameters da/dN (a) and ΔK (b) on a log-log scale. These are the same data shown in Figure 2 but presented here in terms of derivatives. The linear correlation between fatigue and AE parameters is evident in this figure.

The dataset collected using the experimental procedure described here will be used to build a statistical model that can be used for AE-based structural health management.

2.2. Statistical model development

It was shown that on average, a log-linear relationship can be assumed between fracture parameters (da/dN or ΔK) and AE parameter (dc/dN). A statistical model is developed to describe the relationship between these parameters.

Let X denote dc/dN as the independent variable in the regression analysis, and Y denote either da/dN or ΔK as the dependent variable that we are interested in estimating. Regression analysis estimates the conditional expectation of the dependent variable given the independent variable — that is, the average value of the dependent variable when the independent variable is fixed. Another way of looking at this problem is to partition the dependent variable Y into a deterministic component given by function $\phi(\cdot)$ of the independent variable X , plus a zero-mean random component, ϵ , that follows a particular probability distribution. That is,

$$Y = \phi(X; \Theta) + \epsilon \quad (4)$$

The addition of the random term makes the above relationship a statistical model, meaning that the functional relationship between the response variable Y and the predictor variable X holds only in an average sense, not for every data point. Based on the experimental results in previous section, it seems reasonable to assume a linear form for the regression function $\phi(\cdot)$ where $\Theta = (\alpha_1, \alpha_2)$ when Y represents ΔK and $\Theta = (\beta_1, \beta_2)$ when Y represents da/dN .

To complete the model, the error term ϵ must be fully specified as well. Here we adopt the classic regression assumption that the errors are independent and identically-distributed (i.i.d.) random variables and follow a normal probability distribution:

$$\epsilon \sim N(0, \sigma) \quad (5)$$

The mean of the error distribution is zero, and its standard deviation is the unknown parameter σ . Another classic assumption in regression analysis is that the error has a constant variance for all observations regardless of the value of independent variable X . In this application, however, it is reasonable to assume that a small crack is harder to measure, and as the crack becomes larger, the measurement

of its length becomes more accurate. Accordingly, the da/dN and ΔK values associated with data points coming from smaller cracks could be less accurate than those from larger cracks.

One way to account for this effect is to release the constant variance assumption and allow σ to change as a function of the independent variable X . This will result in a flexible model that can capture any change in the error distribution based on the available data. Here, we choose a flexible two-parameter exponential relationship to capture the potential trend in σ ,

$$\sigma = \gamma_1 \exp(\gamma_2 X) \quad (6)$$

This function can capture both increasing and decreasing trends of σ for positive and negative values of γ_2 , respectively. It also reduces to the standard constant variance case if γ_2 is equal to zero. It is important to note that it is not necessary to have any prior knowledge about the trend of σ ; γ_1 and γ_2 are in fact treated as additional unknown parameters and will be estimated using the observed data.

2.3. Bayesian parameter estimation

Numerous procedures have been developed for parameter estimation and inference in regression analysis. Here we adopt a Bayesian approach to parameter estimation often referred to as *Bayesian regression*.

In Bayesian inference, the initial belief about the distribution of the parameters (*a priori* distribution) is systematically updated according to Bayes' theorem (Eq. (7)), based on some kind of evidence or available observations (Figure 4).

$$p(\Theta|D) = \frac{p(D|\Theta)p(\Theta)}{p(D)} \quad (7)$$

where Θ is the vector of model parameters to be estimated and D denotes the set observations to be used in the updating process. $p(\Theta)$ is the *a priori* distribution of model parameters while $p(\Theta|D)$ is the *a posteriori* probability of the model parameters once updated by the observations.

The model that was developed in the previous section can be summarized in the following form:

$$Y = \alpha_1 X + \alpha_2 + \epsilon$$

where

$$\epsilon \sim N(0, \sigma), \quad (8)$$

$$\sigma = \gamma_1 \exp(\gamma_2 X)$$

The likelihood can be defined based on the distribution of the error term, ϵ . To do so, the error $\epsilon_i = y_i - (\alpha_1 x_i + \alpha_2)$ for every data point (x_i, y_i) is calculated.

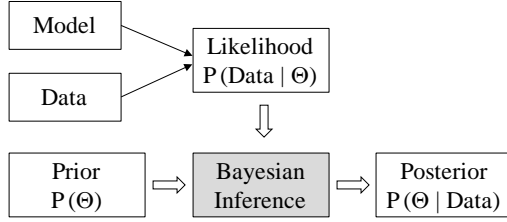


Figure 4: Bayesian Inference Framework

Next, the likelihood of each data point can be defined according to $\epsilon_i \sim N(0, \gamma_1 \exp(\gamma_2 x_i))$. This can be written explicitly as,

$$p(D|\alpha_1, \alpha_2, \gamma_1, \gamma_2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2} \left(\frac{y_i - (\alpha_1 x_i + \alpha_2)}{\gamma_1 \exp(\gamma_2 x_i)}\right)^2\right) \quad (9)$$

The likelihood in Eq. (9) is based on the assumption that the data points are independent and therefore the likelihood for dataset D is simply the multiplication of the likelihood function for every data point.

This study began with no past experience, and therefore non-informative (uniform) prior distributions for all parameters $\alpha_1, \alpha_2, \gamma_1$ and γ_2 were chosen.

The denominator in Bayes' theorem acts as a normalization factor and can be written as,

$$p(D) = \int p(D|\theta)p(\theta)d\theta \quad (10)$$

In practice, numerical approaches such as Monte Carlo-based methods are used to calculate the multidimensional integral in Eq. (10). Here we used WinBUGS (Cowles 2004) to obtain the posterior distributions; WinBUGS is a software package for Bayesian analysis of complex statistical models using Markov chain Monte Carlo (MCMC) methods. Interested readers can refer to (Ntzoufras 2009) for a good reference on Bayesian modeling using WinBUGS. For further reading on MCMC methods in general, see (Gelman et al. 2003; Gamerman & Lopes 2006).

Once the posterior distribution $p(\theta|D)$ is calculated, the inference process is complete. The next step is to use the developed model for prediction using unobserved data. In other words, the model (with posterior parameters) will be used to calculate the distribution of dependent variable Y for a given input X .

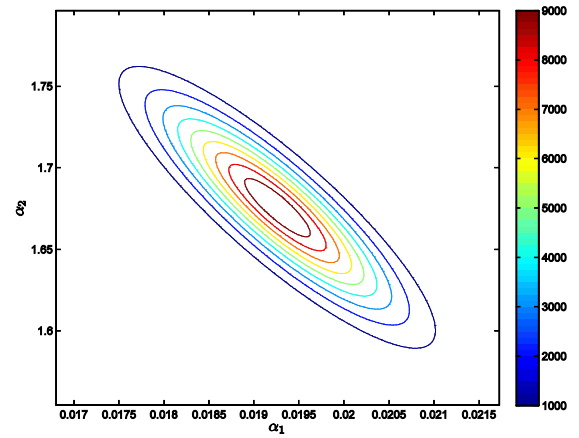
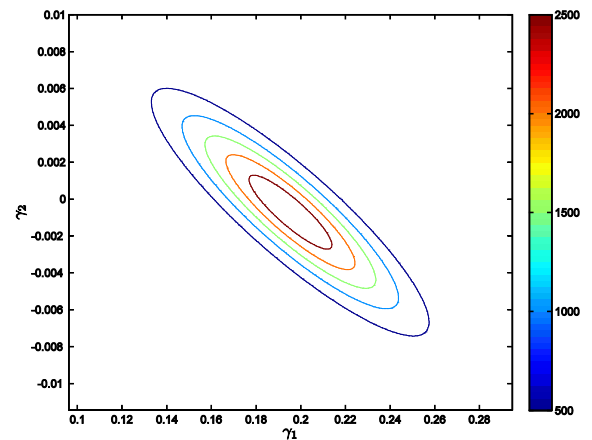
The posterior *predictive distribution* is the distribution of unobserved observations (prediction) conditional on the observed data. Let D be the observed data, θ be the vector of parameters, and D_{pred} be the unobserved data; the posterior predictive distribution is defined as follows,

$$p(D_{pred}|D) = \int p(D_{pred}|\theta)p(\theta|D)d\theta \quad (11)$$

Here again, we are dealing with a multi-dimensional integral that should be calculated numerically. The same MCMC procedure described above can be used to generate samples from the posterior predictive distribution based on draws from the posterior distribution of θ .

2.3.1. Parameter estimation results

Figure 5 shows the contour plot of the posterior joint distribution of parameters α_1 and α_2 . The figure shows that these two parameters are highly correlated (Correlation coefficient $\rho = -0.88$). Similar results are presented in Figure 6 for the parameters γ_1 and γ_2 . These variables are also highly correlated ($\rho = -0.89$), which highlights the importance of considering their joint distribution (rather than marginal distributions) when using the model for prediction.


 Figure 5: Contour plot of the posterior joint distribution of parameters α_1 and α_2 .

 Figure 6: Contour plot of the posterior joint distribution of parameters γ_1 and γ_2 .

It was previously described that the flexible model in Eq. (6) was used to define the standard deviation of the dependent variable Y . For any given input X , one can calculate the corresponding distribution of σ by knowing the joint distribution of γ_1 and γ_2 which was one of the outcomes of the parameter estimation process. This result is shown in Figure 7. Note that for this particular dataset, the median value of σ is relatively constant (it has a slight decreasing trend) over the range of values of $\log dc/dN$. This is consistent with the fact that the estimated value of γ_2 is close to zero (see Figure 6), which means that the relationship in Eq. (6) reduces to a constant variance case where $\sigma_{\Delta K} \approx \gamma_1$. Notice the change in the calculated bounds of σ over the range of $\log dc/dN$. The tighter bounds in the middle of the range are due to a higher density of data points in this region, which results in a more confident estimate in this range.

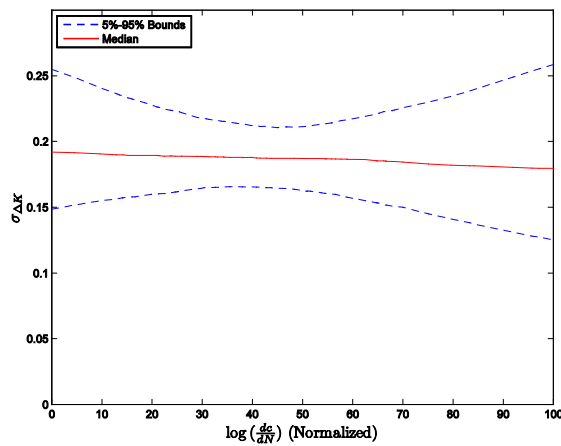


Figure 7: Distribution of $\sigma_{\Delta K}$ as a function of the independent variable $\log dc/dN$

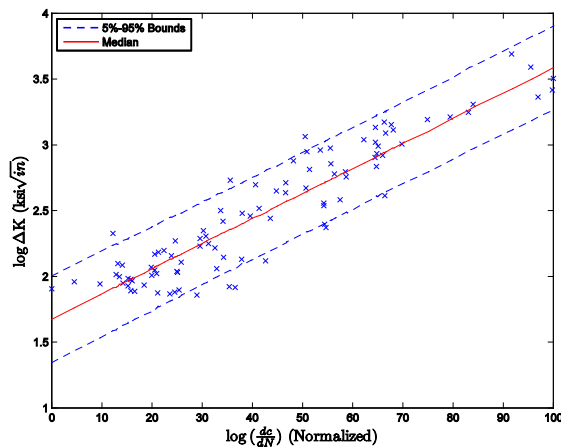


Figure 8: Posterior predictive distribution of $\log \Delta K$ as a function of $\log dc/dN$

Once all the model parameters are estimated, Eq. (11) can be used to calculate the posterior predictive distribution for the dependent variable $\log \Delta K$ as a function of the independent variable $\log dc/dN$, given past observations, D . The result is presented in Figure 8 where the posterior distribution is shown by its median and the 5% and 95% prediction bounds.

The procedure described above can be repeated to fit the model in Eq. (8) to the $\log da/dN$ versus $\log dc/dN$ dataset as well. The models developed in this section provide a quantitative means for relating the crack growth parameters to the AE parameters. In the remainder of this paper, this concept will be used to develop a complete SHM solution based on AE monitoring.

3. STRUCTURAL HEALTH MANAGEMENT USING AE

Three novel approaches are proposed for structural health management using AE monitoring. In all of these approaches, the statistical model developed in the previous section will be utilized to calculate system health parameters solely based on AE monitoring data.

3.1. AE-based damage severity assessment

In this section, we will calculate the probability of structural failure (as defined here) due to crack growth using AE monitoring data.

As a crack grows in a structure, the value of the stress intensity factor ΔK associated with it increases as well. For a standard CT specimen, this relationship is defined as follows (ASTM E647-08 2008):

$$\begin{aligned} \Delta K &= f(a) \\ &= \frac{\Delta P}{B\sqrt{W}} \frac{2\alpha}{(1-\alpha)^{3/2}} (0.886 + 4.64\alpha \\ &\quad - 13.32\alpha^2 + 14.72\alpha^3 - 5.6\alpha^4) \end{aligned} \quad (12)$$

where ΔP is the range of the applied force cycles, W and B are the width and thickness of the CT specimen, respectively, and α is the dimensionless crack size defined as a/W . Equation (12) shows that ΔK , in general, depends on the geometry of the structure, amplitude of the applied load cycles and the instantaneous size of the crack. For a given structure, assuming that the geometry is fixed, a large ΔK represents either a large crack size and/or high load amplitude applied to the structure. ΔK can therefore be considered a criticality parameter that describes the potential of the crack for further growth at any given point in time.

On the other hand, the resistance of a material to stable crack propagation under cyclic loading is characterized by its fracture toughness, K_{Ic} (Anderson 1994). At any point during the crack growth, if the stress intensity exceeds the fracture toughness of the material, the crack growth transitions from stable to non-stable/rapid growth regime where failure is imminent (Figure 9).

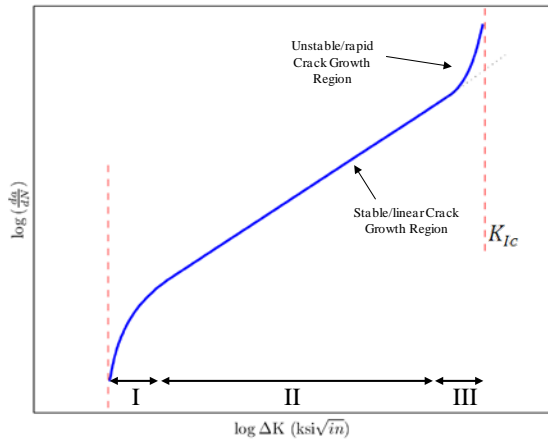


Figure 9: Crack growth sigmoid curve showing both stable and unstable crack growth regions.

In other words, the crack growth is stable as long as K_{max} is less than the fracture toughness of the material, K_{IC} . This fact is used to define an AE-based measure of risk, R_{AE} , as follows,

$$R_{AE} = p(K_{max} > K_{IC}) \quad (13)$$

where K_{max} is defined according to Eq. (12) for $\Delta P = P_{max}$.

Our objective is to assess the health of the structure based only on AE monitoring. To do so, the statistical model developed previously is used in the following way:

Step 1: Estimate the model parameters (θ) using experimental data for a given structure,

Step 2: Monitor the structure using the AE technique and extract the dc/dN parameter from the observed signals,

Step 3: At any given time, use Eq. (11) to calculate the posterior predictive distribution of ΔK as a function of instantaneous AE parameter, dc/dN .

Step 4: Use Eq. (13) to calculate R_{AE} (noting that $K_{max} = \Delta K / (1 - R)$ for constant amplitude loading with loading ratio R).

Figure 10 shows the outcome of the above procedure for steps 1-3. The structure is monitored using the AE technique, and the dc/dN feature is extracted from the signals at different values of elapsed cycles, N . At any given cycle N , the posterior predictive distribution as a function of the instantaneous AE feature, dc/dN , can be calculated. As the number of cycles increases, the crack continues to grow, and therefore, the distribution of K_{max} gradually shifts towards larger values.

Following step 4 in the procedure described above, R_{AE} can be calculated for any given cycle N according to Eq. (13). The result is shown in Figure 11. As shown in this figure, R_{AE} increases (non-monotonically) throughout the experiment.

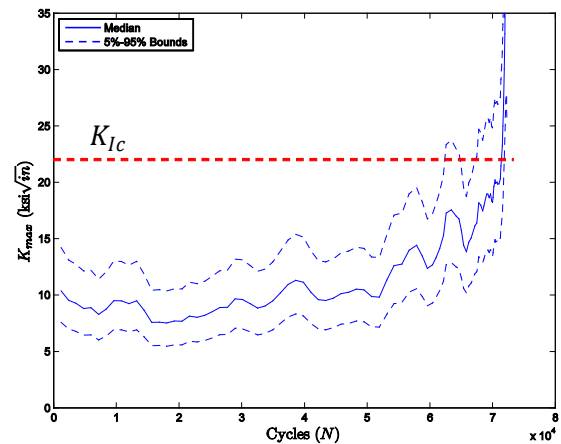


Figure 10: Probability distribution of K_{max} as a function of applied fatigue cycles, N

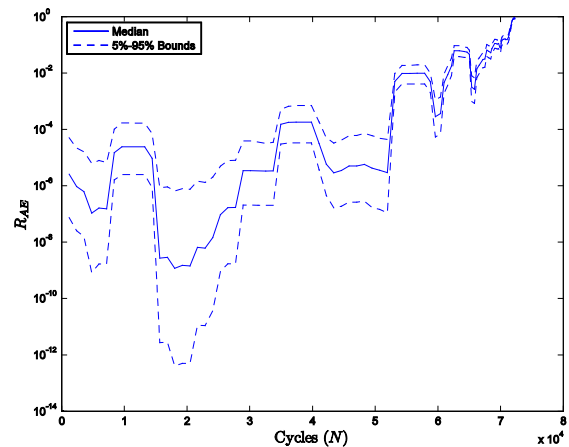


Figure 11: AE-based risk factor, R_{AE} , calculated as a function of applied fatigue cycles, N

The fluctuations in this figure are in fact a direct result of the fluctuations in the input AE feature, dc/dN , which also matches the trend in Figure 10. The AE-based risk factor defined here is an *instantaneous* exceedance probability calculated based on the average value of dc/dN for any given interval. The AE feature has an overall increasing trend that may fluctuate due to instantaneous dynamics of the crack growth. So the best way to interpret the result in Figure 11 is to treat it as a red/green warning mechanism to alert the decision-maker in real-time about the increased risk factor at a given cycle based on the current AE readings.

3.2. AE-based crack growth model

For a given initial crack size, if the rate of crack growth can be estimated, then the crack size itself can be easily calculated by a summation over crack size increments starting from the known initial size. This is the logic behind

most crack growth models. In these models, however, the rate of crack growth is usually calculated based on its empirical relationship with the ΔK parameter, which itself has a complex derivation even for simple geometries.

In the approach presented here, the rate of crack growth is estimated directly from AE monitoring using the statistical model that was developed earlier. The process of estimating crack size using this AE-based crack growth model is summarized in Figure 12.

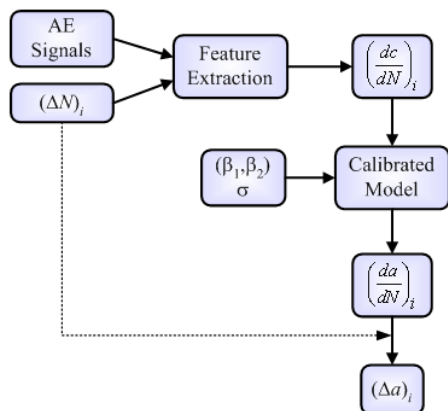


Figure 12: Flowchart of the AE-based crack growth model (Rabiei et al. 2010)

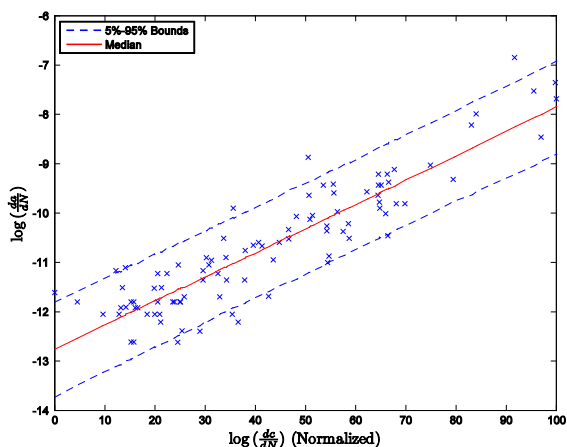


Figure 13: Posterior predictive distribution of $\log da/dN$ as a function of $\log dc/dN$

The process starts by finding the parameters of the model in Eq. (8), where $Y = \log da/dN$ and $X = \log dc/dN$, based on relevant experimental data. The resulting posterior predictive distribution will be used to estimate the distribution of da/dN for any given input dc/dN .

Consider a crack growth experiment where crack growth-related AE signals are recorded throughout the test. For any given interval of elapsed cycles, ΔN_i , the corresponding average AE feature $(\Delta c/\Delta N)_i$ can be calculated. Figure 14 shows the feature extracted from such data during crack

growth in a CT specimen. The probability distribution of the crack extension Δa_i corresponding to the interval ΔN_i can be calculated using Eq. (11). This is shown in Figure 15 using the input AE data shown in Figure 14 and the calibrated model shown in Figure 13.

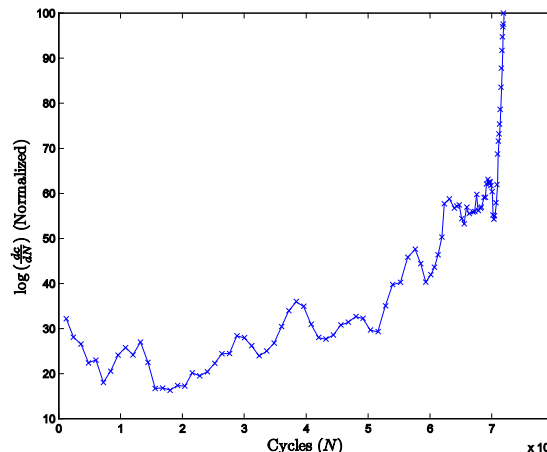


Figure 14: The AE count rate feature extracted from signals obtained during crack growth in a CT specimen

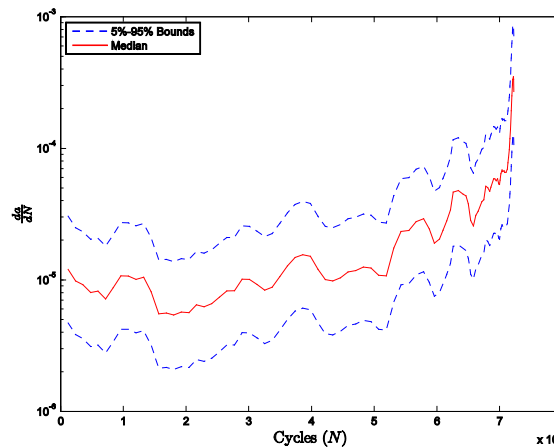


Figure 15: Crack growth rate as a function of applied fatigue cycles predicted via AE monitoring

If the crack size is known at the beginning of the interval, a probability distribution for the crack size at the end of the interval can be easily obtained. By repeating this process for consecutive intervals, multiple crack growth trajectories can be generated, as shown in Figure 16.

The main feature of the AE-based crack growth model presented here is that the rate of crack growth is determined experimentally, and therefore, there is no need to have any information about the amplitude of the applied loading cycles to the structure. This approach, however, relies heavily on a calibrated statistical model that should describe the relationship between an NDI feature of interest

($\log dc/dN$ in this case) and the crack growth rate. Developing a robust model that can capture this relationship with minimum uncertainty is a difficult task that is still a topic of continued research.

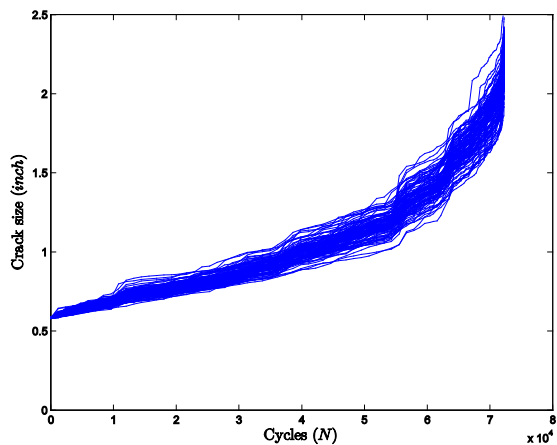


Figure 16: Crack growth trajectories obtained via AE-based crack growth model

recursively update the empirical model prediction as well as the model parameters using crack growth rate and crack size observations.

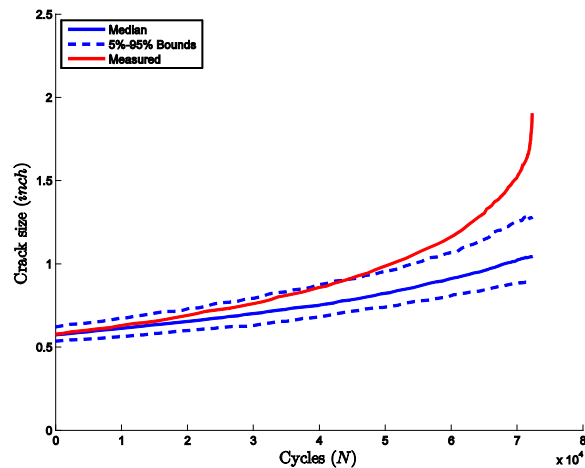


Figure 17: Probabilistic crack growth simulation result using empirical model

3.3. Bayesian knowledge fusion

So far two approaches have been proposed to use AE for quantitative structural health management. A third approach will be discussed here which seeks to use AE findings as an independent source of information to update the outcome of empirical crack growth models.

Several models of varying complexity, e.g. (Forman et al. 1997) and (Walker 1970), have been proposed to describe the crack growth phenomenon. The outcomes of these models suffer from uncertainty from various sources including material properties, model parameters and the model structure. Despite all efforts to capture various sources of uncertainty, the final outcome of the empirical models could still be far from true crack size.

Consider the crack growth test described earlier in this paper. Figure 17 shows the true crack growth trajectory along with empirical model prediction for the CT specimen being considered here. The model in this case consistently underestimates the true crack size. This shows that the actual crack growth rate in the experiment was higher than what was predicted by the model. Several factors (including uncertainty in model structure, uncertainty in model parameters or presence of rogue flaw) could contribute to the poor performance of the empirical model. It is therefore highly desirable to update the model estimates using an independent source of information.

Using the statistical model presented earlier, the AE signals can be translated into crack growth rate information and be used to update the empirical model prediction. (Rabiei 2011) proposed an efficient Bayesian framework to

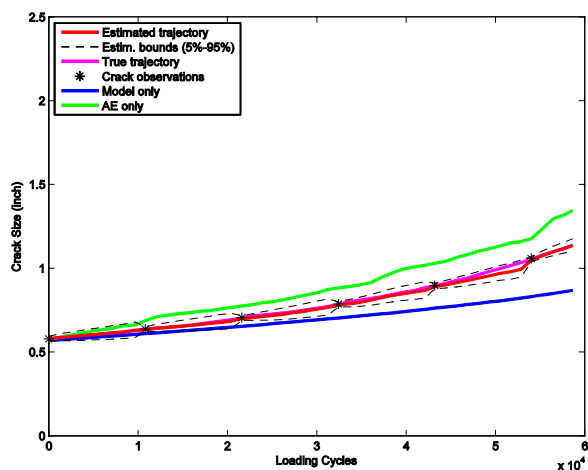


Figure 18: Recursive Bayesian estimation of crack size using crack size and AE-based crack growth rate observations

In Figure 18 the updated crack size estimate for the specimen described above is presented. This result is obtained by: a) recursively updating the crack growth rate based on the AE data, and b) updating crack size at fixed intervals using crack size observations (e.g. periodic inspections). In this figure, the line marked as *model only* is the outcome of the empirical crack growth model. The *AE only* line, on the other hand, shows the crack growth trajectory as predicted solely by the AE-based crack growth model. The *estimated trajectory* is the fusion result obtained via recursive Bayesian estimation. In this particular case, since the empirical model consistently underestimates while

the AE-based approach consistently overestimates the crack size, the fusion results in an enhanced crack size prediction. It is important to note that this observation is based on results from limited experimentation and cannot be generalized. The fusion outcome is dependent on the performance of the individual techniques fused together. Obviously, if both the model and the AE observations overestimate the crack size in one application, the fused result will also be an overestimation of the true crack trajectory.

4. CONCLUSION

Three new approaches were proposed for quantitative structural health management using in-situ AE monitoring: in the first approach, an AE-based risk measure, R_{AE} , was defined as the probability that the crack growth will transition from the stable to non-stable/rapid growth regime. The transition probability was calculated as the probability that K_{max} exceeds the fracture toughness of the material, K_{Ic} . In the proposed approach, K_{max} is calculated as a function of real-time AE monitoring data using the calibrated statistical model developed in this paper.

In the second approach, AE monitoring data was used to calculate the instantaneous distribution of crack growth rate, da/dN . For a given initial crack size and with crack growth rates obtained from AE monitoring, the crack size distribution was estimated as a function of elapsed fatigue cycles.

Recursive Bayesian estimation technique was used to fuse the outcome of the empirical crack growth model with crack size observations as well as the online crack growth rate observations obtained from AE monitoring.

5. REFERENCES

- Anderson, T.L., 1994. *Fracture Mechanics: Fundamentals and Applications, Second Edition* 2nd ed., CRC.
- ASTM E647-08, 2008. *Standard Test Method for Measurement of Fatigue Crack Growth Rates*, ASTM International.
- Bassim, M.N., St Lawrence, S. & Liu, C.D., 1994. Detection of the onset of fatigue crack growth in rail steels using acoustic emission. *Engineering Fracture Mechanics*, 47(2), pp.207-214.
- Cowles, M.K., 2004. Review of WinBUGS 1.4. *The American Statistician*, 58(4), p.330-336.
- Forman, R.G., Kearney, V.E. & Engle, R.M., 1997. Numerical analysis of crack propagation in cyclic-loaded structures.
- Gamerman, D. & Lopes, H.F., 2006. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*, Chapman & Hall/CRC.
- Gelman, A. et al., 2003. *Bayesian Data Analysis, Second Edition* 2nd ed., Chapman & Hall.
- Hamel, F., Bailon, J.P. & Bassim, M.N., 1981. Acoustic emission mechanisms during high-cycle fatigue. *Engineering Fracture Mechanics*, 14(4), pp.853-860.
- Holford, K.M. et al., 2009. Acoustic emission for monitoring aircraft structures. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 223(5), pp.525-532.
- Mix, P.E., 2005. *Introduction to nondestructive testing: a training guide*, Wiley-Interscience.
- Ntzoufras, I., 2009. *Bayesian Modeling Using WinBUGS*, Wiley.
- Paris, P. & Erdogan, F., 1963. A critical analysis of crack propagation laws. *Journal of Basic Engineering*, 85(4), p.528-534.
- Rabiei, M., 2011. *A Bayesian framework for structural health management using acoustic emission monitoring and periodic inspections*. College Park: University of Maryland.
- Rabiei, M., Modarres, M. & Hoffman, P., 2009. Probabilistic Structural Health Monitoring Using Acoustic Emission. In Annual Conference of the Prognostics and Health Management Society 2009. San Diego, CA.
- Rabiei, M., Modarres, M. & Hoffman, P., 2010. Towards Real-Time Quantification of Fatigue Damage in Metallic Structures. In Aircraft Airworthiness & Sustainment (AA&S 2010). Austin, TX.
- Rahman, S. & Rao, B.N., 2002. Probabilistic fracture mechanics by Galerkin meshless methods—part II: reliability analysis. *Computational mechanics*, 28(5), p.365-374.
- Walker, K., 1970. The effect of stress ratio during crack propagation and fatigue for 2024-T3 and 7075-T6 aluminum. *Effects of environment and complex load history on fatigue life*, p.1-14.

Study on MEMS board-level package reliability under high-G impact

Jiuzheng Cui, Bo Sun, Qiang Feng, ShengKui Zeng

School of Reliability and Systems Engineering, Beihang University, 37 Xueyuan Road, Beijing, China
cuijiuzheng@ste.buaa.edu.cn
sunbo@buaa.edu.cn

ABSTRACT

Under high-G (10^4 g or above) impact load conditions, the reliability of micro electro mechanical systems (MEMS) board-level package and interconnection are critical concerned that influence the mission success of total projectile. This paper conducts a research on this problem to analyze package reliability using finite element modelling (FEM) and simulation method. Theoretical analysis and mathematical model for failure mechanism of MEMS package under high-G impact are conducted and established. A FEM dynamic analysis is conducted on a typical MEMS board-level leadless chip carrier (LCC) package. Results show that the solder joints are one of the key weakness points that influence the reliability of MEMS package. The maximum effective stress in the structure occurs at the outer corner in the outermost solder point, and the alloy cover and printed circuit board (PCB) have a greater deformation.*

1. INTRODUCTION

In gun-shooting and projectile process, the projectile and its inner components (such as MEMS gyroscope, accelerator, and other electrical components) are suffering large inner pressure and high acceleration load. This type of load features as an extremely peak acceleration (10^4 g level or above) and duration of extremely short time (such as 10ms) (Lou *et.al.*, 2005), (Vinod *et.al.*, 2008), (Jiang *et.al.*, 2004). Under this load conditions requirement, it's difficult to design and manufacture a reliable fine component used in projectile. The failure of component is frequently found in this type of usage environment that will influence the total projectile reliability.

MEMS and electronics component board level package and interconnections (solder joints) are key weakness points that influence reliability. While LCC (Leadless Chip Carrier) is the general package type that adopted in MEMS and other electronics with its remarkable advantage of small scale and

lower cost (Wei *et.al.*, 2004), (Thomas *et.al.*, 2008). LCC package technique belongs to SMT (Surface Mount Technology), and its reliability problem has got more and more attention. Surface mounted technology (SMT) is widely used in MEMS package, as the solder joints of SMT are significantly small (typically a few millimeters), which turns the solder joints between the printed circuit board (PCB) and MEMS chip into the most weak component during the impact(Tee *et.al.*, 2004).

In present literature and report, the research activities have been focusing on the reliability of board level interconnections in drop impact (with 10^2 ~ 10^3 g level acceleration) for portable electronics (JEDEC Standard, 2003), (Younis *et.al.*, 2007), covering experimental work together with analytical and numerical modeling studies (Sirkar and Senturia, 2002), (Suhir and Burke, 2000), (Yu *et.al.*, 2003), (Tee *et.al.*, 2004). But the fundamental understanding of the reliability of board level interconnections to high-g impact (with 10^5 g level acceleration) remains limited and is a subject which is still need further studied. Also in these drop impact studies, the analysis is especially focus on all kinds of BGA (Ball Grid Array) packages used in portable electronics (Yu *et.al.*, 2003), (Tao *et.al.*, 2006). While the study on the LCC package reliability generally used in MEMS is not found yet.

This paper provides a research on the reliability of solder joints / interconnections in high-g impact and deal with the dynamics of board level impact and package reliability assessment using finite element modeling and simulation method. Theoretical analysis of failure mechanism of MEMS package is conducted to provide a physical explanation and a FEM dynamic analysis is conducted on a typical LCC MEMS package. Efforts will provide reference for development and practical utilization of MEMS components.

2. THEORETICAL ANALYSIS OF FAILURE MECHANISM

Numerical simulation and experimental validation study have confirmed differential flexing between the PCB and the component packages as the primary driver for the failure of board level interconnections during high-g impact (Wong,

* This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

2006). The interconnections (solder joints) of MEMS device and electronics component are key weakness that influences the reliability of MEMS, for the differential deflection between the PCB and MEMS device together with mechanical resonance (Tee et.al., 2003).

Generally, the devices such as MEMS components can be taken as rigid body compared to PCB, bending moment will be introduced in PCB during high impact, which will make the PCB produce flexure deformation. The flexure deformation will introduce repeated pressure and compressive stress to the interconnections (solder joints) between MEMS device and PCB, which will result in connection failure of solder joints. The smaller the deformation is, the smaller the stress and strain are, and the time to recover to static equilibrium (the oscillation of impact) is shorter, which can increase the reliability of solder joints.

The PCB are fixed to the base by bolt, considering the transfer process of stress wave, the dynamic response of MEMS package during high-g impact can be simplified to a board supported by each side and the load can be simplified to uniform load applied to underside of PCB, the two-dimensional simplified universal model is shown in Fig.1.

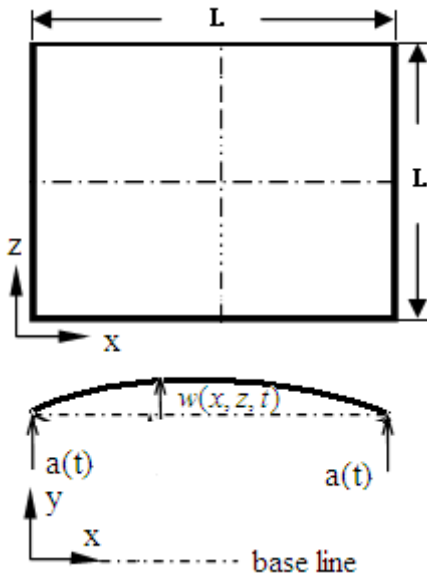


Fig.1. Simplified PCB model

Based on the simplified model, two failure mechanisms can be obtained: overstress fracture under high impact and fatigue fracture for multi-impact. For onboard devices, the launch environment take place only once, thus overstress fracture is concerned in this study.

The equations of motion under high-g impact are as follows (Suhir, 2002).

$$D \frac{\partial^4 w(x, z, t)}{\partial x^2} + \rho h \frac{\partial^2 w(x, z, t)}{\partial t^2} = 0 \quad (1)$$

Where $D = \frac{Eh^3}{12(1-\nu^2)}$ is the bending stiffness of PCB, ρ , E ,

h and ν are the density, elastic module, thickness and Poisson ratio of the substrate

The flexure deformation of PCB $w(x, z, t)$ can be defined as the linear superposition of model units, and the superposition equation is as follows:

$$w(x, z, t) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} X_i(x) X_j(z) u_{ij}(t) \quad (2)$$

Where $u_{ij}(t)$ represents the displacement of micro-deformation unit corresponds to ij order mode shape function. The mode function $X_i(x)$ and $X_j(z)$ of PCB, the natural frequency ω_{ij} , and the displacement u of the board can be calculated respectively according to the following equation

$$X_i(x) = \sin(i\pi/a)x \quad (3)$$

$$X_j(z) = \sin(jz/b)z \quad (4)$$

$$\omega_{ij} = \pi^2 \gamma (i^2 + j^2) / L^2 \quad (5)$$

$$\frac{u_{ij}}{a(t)} \omega_{ij}^2 = \frac{16}{ij\pi^2} \psi_{ij} \quad (6)$$

For most of the differential equations of motion, there is a large number of nonlinear terms, the general analytical solution cannot be obtained, and numerical solution is needed by computer analysis, FEM is commonly used (including explicit finite elements, implicit global model and transient dynamic response simulation such as LS-DYNA) and so on.

High impact test are normally conducted to understand the actual response and verify the analysis results. The actual acceleration response curves on different sites (impact table, base plane, and test vehicle) are acquired for further numerical analysis. (See next section)

3. FINITE ELEMENT MODEL AND DYNAMIC RESPONSE ANALYSIS

3.1 The structure of MEMS

In military applications, MEMS gyroscopes is mainly used for navigation guidance, attitude determination and stability, etc., the projectile usually stand with tremendous

acceleration when launching, which would cause the solder joints fracture.

In this paper, a typical LCC package of MEMS gyroscopes in high impact is analyzed. The comb is installed inside the ceramic which is surface mounted on the PCB, brazing technology are used to cap sealing in vacuum condition. The package shown in Fig.2 is formed of the cover, ceramic, PCB and solder joints, the comb is ignored during the modeling for the mass and volume is much smaller than the package and the PCB.

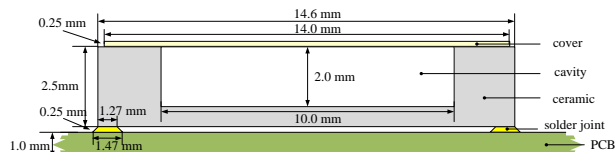


Fig.2. MEMS package structure and geometry

3.2 FE modelling and analysis

FE modeling is proven to be a very efficient tool for design analysis and optimization of IC and MEMS packaging, because of advantages of economic, saving time and being able to provide comprehensive information.

For the characters of high-g impact, the transient dynamic analysis is conducted using direct time-integration method. This can be further divided into implicit or explicit algorithm. While the contact duration may be extremely short and intense or when the impacting bodies of interest are excited into very high frequencies response, such as the

cases of drop impact on rigid surface. For all of these reasons, the explicit algorithm is adopted in this analysis.

Finite element model was established according to the MEMS gyroscope. The structure is mainly combined of the cover, ceramic, PCB and solder joints. Solid element (SOLID 164) was used to model all components including the PCB in this model. Due to symmetry of the package, a quarter of the model was established in order to simplify modeling and save compute time, finite element model of LCC package is shown in Fig.3.

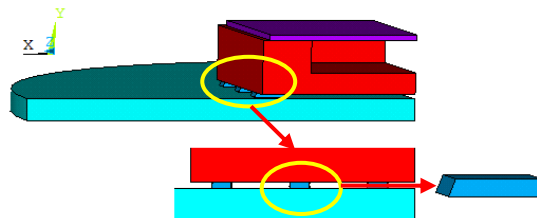


Fig.3. Finite element model of MEMS package

The material properties of MEMS packaging used in the FEM are shown in Table 1. The bilinear kinematic hardening model is chosen for the solder joints to be closer to the engineering practice, and linear material model is used for the rest component of the package. Generally, the damping of the system is often set between 0.01-0.25, where a fixed value of 0.03 is recommended here (Zhao et.al., 2004).

Table 1 Material properties used in FEM

Structure	Materials	Density (kg/m^3)	Young's modulus (GPa)	Poisson's	Yield strength (MPa)	Tangent modulus (GPa)
PCB	FR4	1900	22	0.3	450	—
Cover	alloy	8460	138	0.31	380	—
Body	ceramal	3920	344	0.22	580	—
Solder joint	Sn3.5Ag	7400	52.7	0.4	22.5	3.09

The hexahedral element is used to mesh the FEM in this paper for the low stiffness and high accuracy compared to the tetrahedron element. The mesh of the solder joints are refined to get a precise result of the stress in solder joints. The model is consisting of 10,772 nodes and 7,371 units.

The input-G method developed by Tee (Tee *et.al.*, 2004) is used in this dynamic analysis. This method is more accurate and much faster, and bypasses many technical difficulties in conventional dynamic model such as adjusting the parameters of contact surfaces, defining contact type, etc. In this way, only the package itself needs to be modeled. The impact acceleration pulsed which is measured and built from the actual test from the missile (see Fig.4.), are imposed on the supports of the board-level test vehicle as

load condition. A very detailed finite element model of the board and package was constructed and simulated by LS-DYNA (ANAYS).

Fig.4. shows the impact pulse according to the actual measurement which features as an impact acceleration pulse with a duration of 0.01s, the initial value of $1.44 \times 10^4 m/s^2$ and peak acceleration of 12600g at 0.0026s, at 0.01 seconds, the acceleration reduced to $4.71 \times 10^4 m/s^2$, after 0.01s, the load was removed.

In ideal conditions, MEMS packaging only produced a straight up displacement (y direction) and the displacement (x, z) of PCB is confined in the whole process during the impact. As the result of a quarter models, symmetry

boundary constraints should be applied to xy and yz surface on the inner side of the model.

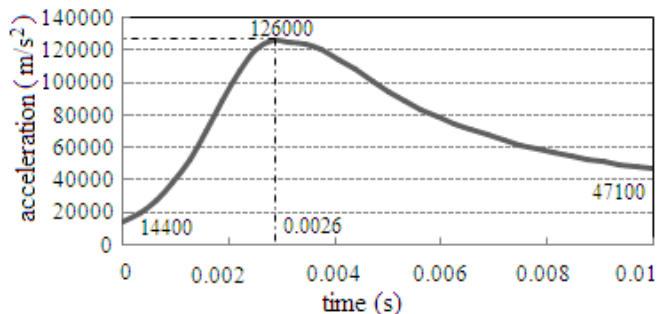


Fig.4. Acceleration curve

Explicit nonlinear algorithm, a more suitable method, was used for dynamic response analysis. Contact surface cannot penetrate each other by the outer surface, so automatic single surface contact was set for global contact. Hourglass is a model in theory, not exist in the actual process, it is a mathematically stable but physically impossible state. Solutions will be useless because of hourglass, and if the total hourglass of the model can be greater than 10%, the result is generally a fault result. The hourglass was set 10% of the total internal energy in this paper. The results file output interval was 100, and time-history output Interval was 100, the solution termination time was 0.03s.

3.3 Results and Discussion

Fig.5 shows the typical simulation analysis results for the contours of von-mises stress of MEMS. The response of MEMS package under high-g impact is a dynamic process, the maximum effective stress occurs in the outermost corner and reached to 27.17MPa, more than the yield strength of solder joint (22.5MPa), which make the solder joint produce plastic deformation and maybe fracture during the impact; The maximum plastic strain has reached 2.7×10^{-3} . Fig.6 shows Time-History curve of solder joint at the maximum stress point. The solder joints experience tensile stress first, and then stress changes to attenuating. The effective stress reached its peak at 0.0027s, and after about 0.0105s, the stress tends to stabilize.

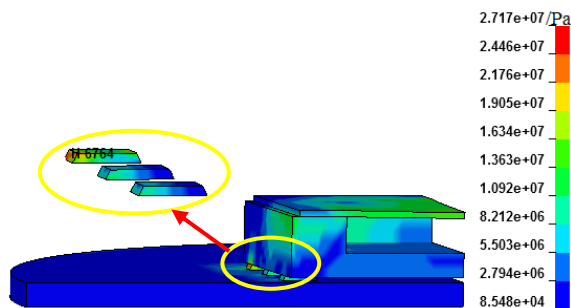


Fig.5. The stress distributing of the finite element simulation

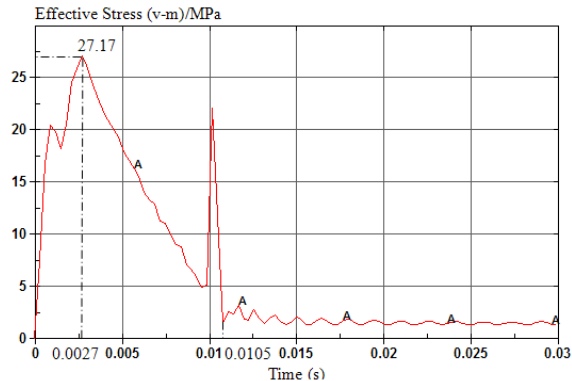


Fig.6. Effective stress Time-History of maximum stress point

By comparing the deflection-time curve of the PCB and the stress-time curve of the solder joints, it finds that the maximum stress in solder joint reaches its peak at the moment of the deflection of the PCB reaches its maximum, and the effective stress has a slight fluctuation along with the acceleration curve during the impact. It suggests that peeling stress in solder joints is mainly caused by the deflection of the PCB board during impact.

In addition, the deformation of the alloy cover is large during the impact; Fig.7 shows the cover had bending phenomenon to the cavity. At the same time, difference between the maximum and minimum displacement of the unit is the maximum deformation, the maximum deformation of the cover is 0.08mm, which is 32% of the cover thickness (0.25mm), and it is 4% of the cavity depth MEMS devices (2mm), if the deformation of alloy cover is too large, it will squeeze the internal structure in LCC package, then failure occurs. Therefore, the proposed design of the LCC package should reserve larger space for the upper part to avoid excessive deformation of the cover. In addition, by using potting processing, the squeeze effect would be more serious.

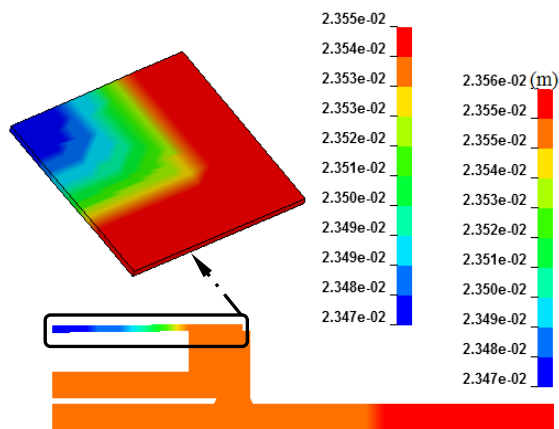


Fig.7. The deformation distributing of the cover Meanwhile, the edge of PCB occurred to bend up, as shown in Fig.8. And at the moment of maximum stress occurred,

the deformation of PCB reached a maximum of 0.01mm, which is 1% of its own thickness (1mm), and it is 5% of solder joint height (0.2mm), thus, the squeeze will extrusion the solder joint which is the main reason to produce internal stress of the solder joint.

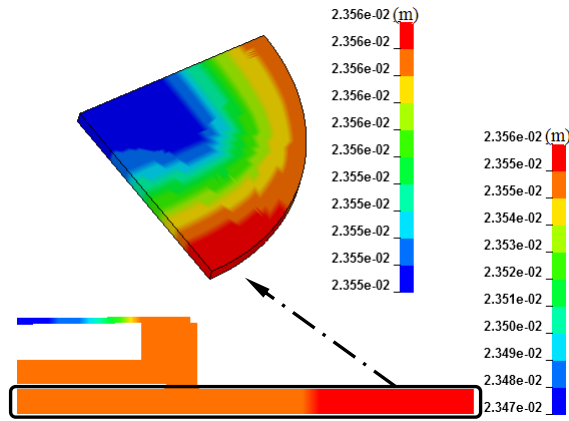


Fig.8. The deformation distributing of the PCB

4. CONCLUSIONS

This paper conducts a research on the dynamic response of MEMS gyroscope board-level package reliability under high-G impact. Theoretical analysis and mathematical model for failure mechanism of MEMS package (Leadless Chip Carrier, LCC) under high-G impact are established and analyzed. Analytical solutions that provide physical insights to the dynamics of PCB and the interconnection stresses have been presented.

Under high-g impact, solder joint is the key weakness that influence MEMS package, which will be fractured and failure easily. The response of MEMS package is a dynamic process. Moreover, the maximum effective stress in the structure occurs at the outer corner in the outermost solder point, and the alloy cover and PCB have a greater deformation.

The room between the cover and the device inside MEMS should be carefully designed, and the distant must be longer than 10% of the cavity depth.

Considering the failure mechanism of solder joint that the flexure deformation of the PCB produces stress inner solder joint, we recommend minimizing the length and width of the PCB or increasing the thickness of the PCB or increasing the stiffness of the PCB so as to decrease the stresses in the solder joint.

ACKNOWLEDGEMENT

This study was conducted under the contact of NanJing University of Science and Technology. The author also would like to thank the member of City PHM Team and the reviewers of this paper for their valuable suggestions.

REFERENCES

- Lou Xia, Shi Jinjie, Zhang Wei, et al. Study on the Packaging Technology for a High-G MEMS Accelerometer [C]. *Proceedings of Electronic Packaging Technology Conference, Singapore: IEEE*, 2005, 103-106
- Vinod Chakka, Mohamed B Trabia. Modeling and Reduction of Shocks on Electronic Components within a Projectile [J]. *International Journal of Impact Engineering*, 2008,35(11) 1326-1338
- Jiang Qiyu, MaoHu Du, Le Luo, et al. Simulation of the Potting Effect on the High-G MEMS Accelerometer [J]. *Journal of Electronic Materials*,2004,33(8):893-898
- Wei Lingyun, Mei Zhao, Qiang Guo et al. SMT Solder Joint's Shape and Location Optimization using Modified Genetic Algorithm in the Dynamic Loadings [C]. *Proceedings of the International Conference on the Business of Electronic Product Reliability and Liability*, 2004, 169-173
- Thomas F. Marinis, Joseph W. Soucy and James G. Lawrence. Vacuum Sealed MEMS Package with an Optical Window [C]. *Proceedings of Electronic Components and Technology Conference*, 2008, 804-810
- Tee T Y, Luan Jing-en. Advanced Experimental and Simulation Techniques for Analysis of Dynamic Responses during Drop Impact [C]. *Proceedings of the 5th Electric Component and Technology Conference, Las Vegas: IEEE*, 2004,1088-1094
- JEDEC Standard JESD22-B111. Board Level Drop Test Method of Components for Handheld Electronic Products[S], 2003
- Younis, Daniel Jordy, et al. Computationally Efficient Approaches to Characterize the Dynamic Response of Microstructures under Mechanical Shock [J]. *Journal of Microelec-tromechanical systems*, 2007:16(3), 628-638
- Sirkar V T, Senturia S D. The Reliability of Microelectromechanical System (MEMS) in Shock Environments [J]. *Journal of Microelectromechanical systems*, 2002, 11(3):206-214
- Suhir E, Burke R. Analysis and Optimization of the Dynamic Response of a Rectangular Plate to a Shock Load Acting on its Support Contour with Application to Portable Electronic Products [J]. *Advanced Packaging*, 2000, 122(1):3-5
- Yu Q, Kikuchi H, Ikeda S Y, et al. Dynamic behavior of electronics package and impact reliability of BGA solder joints [J]. *Structure and Materials*, 2003, 12:55-64
- Tee T Y, PEK Eric, et al. Novel Numerical and Experimental Analysis of Dynamic Response under Board Level Drop Test [C]. *Proceedings of 5th International Conference on Thermal and Mechanical Simulation and Experiments in Microelectronics and Microsystems, Brussels: IEEE*, 2004, 1131-1124

Tao Jianlei, Qu Xin, Wang Jiaji. Simulation of Average Strain Energy Density (SED) in the BGA Soldering during the Drop Test [C]. *Proceedings of 7th International Conference on Electronics Packaging Technology*, Shanghai: 2006,1-6

Wong E H. Drop impact test - mechanics & physics of failure

Tee T Y, Ng H S, Zhong Z P, et al. Design for Enhance Solder Joint Reliability of Integrated Passives Device under Board Level Drop Test and Thermal Cycling Test [C]. *Proceedings of 5th EPTC Conference*. Singapore: IEEE, 2003,210-216

E Suhir. Could Shock Tests Adequately Mimic Drop Test Conditions? [C] *Proceedings of Electronic Components and Technology Conference*. Singapore: IEEE, 2002,563-573

Zhao Mei, Zhou Haitang, et al. Mechanical vibration and noise [M]. *Beijing: Science and Technology Press*, 2004(In Chinese)

Jiuzheng Cui is currently a student of School of Reliability and Systems Engineering at Beihang University in Beijing, China. He received his B.E. degree in detective guidance and control from Northwest Polytechnic University. His current research interests include reliability of electronics, physics of failure, prognostics and health management, failure analysis of electronics, reliability engineering, and integrated design of product reliability and performance.

Dr. Bo Sun is a reliability engineer and a member of the faculty of School of Reliability and Systems Engineering at Beihang University in Beijing, China. He received his Ph.D. degree in reliability engineering and systems engineering

from Beihang University and a B.S. degree in mechanical engineering from the Beijing Institute of Mechanical Industry. His current research interests include reliability of electronics, physics of failure, prognostics and health management, failure analysis of electronics, reliability engineering, and integrated design of product reliability and performance.

Dr. Qiang Feng is a reliability engineer and a member of the faculty of School of Reliability and Systems Engineering at Beihang University in Beijing, China. He received his Ph.D. degree in systems engineering and a B.S. degree in mechanical engineering from the Beihang University. His current research interests include reliability engineering, systems engineering, physics of failure, and synthetical design of product reliability maintainability supportability and performance.

Prof. ShengKui Zeng is currently the Vice Director of the School of Reliability and Systems Engineering at Beihang University in Beijing, China. He has over 15 years of research and teaching experience in reliability engineering and systems engineering. He was a visiting researcher with the Center for Advanced Life Cycle Engineering Electronic Products and Systems Consortium at the University of Maryland in 2005. He is the team leader of the KW-ARMS[®] reliability engineering software platform, the co-author of three books, and a recipient of three Chinese ministry-level professional awards. His recent research interests include prognostics and health management, integrated design of reliability and performance, and reliability-based multidisciplinary design optimization.

Symbolic Dynamics and Analysis of Time Series Data for Diagnostics of a dc-dc Forward Converter

Gregory M. Bower¹, Jeffrey Mayer¹, and Karl Reichard²

¹ *The Pennsylvania State University, University Park, PA, 16803*
gmb162@psu.edu
mayer@engr.psu.edu

² *The Applied Research Laboratory, State College, PA, 16801*
kmr5@psu.edu

ABSTRACT

This paper presents a novel approach to diagnosis of dc-dc converters with application to prognosis. The methodology is based on Symbolic Dynamics and Diagnostics. The data derived method builds a statistical baseline of the converter that is used to compare future statistical models of the converter as it degrades. Methods to determine the partitioning and number of partitions for the Symbolic Dynamics algorithm are discussed. In addition, a failure analysis is performed on a dc-dc forward converter to identify components with a high probability of failure. These components are then chosen to be monitored during accelerated testing of the dc-dc forward converter. The methodology is experimentally validated with data recorded from two dc-dc converters under accelerated life testing.*

1. INTRODUCTION

Diagnostics methodologies attempt to determine the current state of health of a system and flag any type of anomalous behavior that could affect the operation of the system. Successful diagnostics can eventually lead to prognostication of a system where prognostication is the prediction of the remaining useful life of the system under monitor (Hess et al., 2005)

The goal of diagnostics and health management in general is to maintain system operability, reduce maintenance costs, and maximize safety. Diagnostics and health management of electronic systems can be obtained by numerous different

methodologies. Most of these can be sorted into either a data driven or model based category. Model based methods, such as the name implies, rely on a physical model representation of the system and the underlying degradation process (Brown et al., 2006). On the other hand, data driven models tend to model the degradation of a system by long term monitoring of the system. This methodology tends to require large data sets in order to train the data driven models used to generate the health measures.

In this paper, we aim to develop a methodology based on Symbolic Dynamics (SD) (Ray, 2004; Rohan, 2006) that can be used to generate diagnostic measures from a degrading dc-dc converter. Symbolic Dynamics has been applied to many systems including inverter fed induction machines (Rohan et al., 2006), fatigue crack diagnosis (Singh et al., 2010), and in nuclear power plant operations (Jin et al., 2011).

In this paper, we used a dc-dc forward converter for our test subject. Data is recorded from an accelerated test of these converters on an hourly basis and is used in the algorithm. It is our intention to expand the results into a prognostic algorithm that can deduce the remaining life of the converter from the current anomaly generated by the SD algorithm.

Symbolic dynamics lends itself well to the area of electronic diagnostics as it is a relatively simple algorithm to implement. In general, the algorithm analyzes the data captures and forms states based on the data. These states are then tracked statistically through time. These states can be designated in numerous ways; that is, the states could directly be related to the data points themselves or represent the duty cycle of the converter.

* Gregory M. Bower et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

This paper is organized as follows. Section 2 presents a background on the Symbolic Diagnostics and Analysis methodology. Section 3 focuses on the dc-dc converter and the accelerated testing of the converter to generate the experimental data used in the algorithm. Section 4 presents the results of the algorithm compared to other electronic system diagnostic metrics. Finally the paper concludes with a short conclusion and future work suggestions.

2. SYMBOLIC DIAGNOSTICS AND ANALYSIS OF TIME SERIES DATA

Diagnostics of the converter is accomplished through the use of symbolic diagnostics and analysis of time series data (Ray, 2004). In this work, we propose a methodology to accomplish diagnostics of a dc-dc converter in real-time with the use of this tool. The method depends on symbolizing the captured time series data, generating states that consists of permutations of generated symbols, and calculating the probability of these states.

In general, the probability of the states change as the converter ages and degrades. Tracking these changes allows for quantification of the degradation in the dc-dc converter over time. These changes can be quantified by comparing the current analysis to a baseline case. Two underlying assumptions must be satisfied in order to use symbolic dynamics. They are:

- 1) The system degradation mechanisms must be dynamically separate from the system dynamics; and
- 2) The system generates monotonically positive anomaly measures.

The first requirement is a two time scale separation argument. If the system dynamics are much faster than the degradation mechanisms, then individual data captures will contain stationary degradation dynamics. The second assumption states that the system does not exhibit self-healing or is repaired. This assumption is flexible in that a non-monotonically increasing anomaly can make diagnosis more difficult but not impossible. A monotonically positive increasing anomaly also pushed the methodology towards a prognostics tool.

The dc-dc converter satisfies assumption one as the system dynamics are monitored through time. The dynamics of the dc-dc converter are based on the switching frequency of the converter, that is, 100 kHz. A short data capture is taken at a faster rate than the switching frequency and is used to determine the current system state of health. During this short interval, the degradation in the converter can be considered stationary. For assumption two, the converter is allowed to age without repair. The degradation in the converter continually increases and with the anomaly quantification metrics, generates an increasing anomaly measure.

The methodology for Symbolic Dynamics begins by first determining the number of symbols to use in the definition of the symbolic sequence and also defining the partitions to assign symbols to time series data points which is closely related to defining the number of symbols. Each data point is assigned to a unique symbol. This step can be considered as a coarse quantization of the time series data.

With the symbolic series now generated, the next step is in determination of states for the algorithm. States are simply defined as groupings of D symbols. Throughout this paper, the choice D , called depth in the algorithm, is chosen to be unity; that is, each symbol results in a state. Once the states are defined, the probabilities of occurrence of the states are used to generate an anomaly in the behavior of the system that is related to degradation. Currently, there are numerous metrics to quantify an anomaly based on these state probabilities.

The algorithm will now be discussed in more detail including the partitioning of the time series data, generation of the symbolic sequence, and the determination of parameters in the algorithm. With the completion of symbolization, the discussion will continue with defining an anomaly based on the statistical model generated from the time series data.

2.1 Choice of Number of Symbols

In order to enable the partitioning of the time series data, the choice of the number of symbols in the algorithm must be determined. Two methods are presented, one for each type of partitioning methodology. The partitioning methods will be discussed in the next section. Each method is based on the entropy of the resultant symbol distributions generated from the partitioning method.

For uniform partitioning, the choice of number of symbols to use is defined by the use of Entropy Efficiency. Entropy Efficiency is given as:

$$E_e = \frac{\sum_{i=1}^N p_i \log_2(p_i)}{\log_2(N)} \quad (1)$$

where p_i is the probability of the i^{th} symbol. The p_i 's are calculated at each iteration of the search for N and represent the probability of each individual partition. The logarithm is taken to the base 2 such that result of entropy is based in bits. The aim is to determine the maximum of Eq. (1) over uniform partition size, N .

Equation (1) can be interpreted in two ways. First, the denominator term acts as a penalty term for larger distributions that is a large choice of N . This enforces computationally a more efficient algorithm.

Secondly, this metric measures the entropy deviation from the ideal entropy given by a uniform distribution (represented in the denominator term). For source distributions that are not known a priori, a good estimate for the source distribution is that from a uniform probability if there are no constraints or assumptions on the underlying generating symbol distribution (Conrad, 2011).

An issue associated with this method that must be kept in mind is if during the process of looking for an optimal number of states is if a state is generated that has a null symbol occurrence probability. This would cause the entropy estimate to be undefined. In this case, the search concludes with the generation of the first null symbol probability.

For ME partitioning, we again turn the use of entropy under a method developed by (Rajagopalan & Ray 2006). This method estimates the number of states through the use of histograms. To estimate the number of states we use the differential entropy of the time series data given as:

$$h(x) = - \int_{-\infty}^{\infty} p(x) \log_2(p(x)) dx \quad (2)$$

where x is the possible values the data can take and $p(x)$ is the probability density of x . Once the entropy for the time series data is estimated, the number of symbols for ME partitioning is given by:

$$N = \operatorname{argmin}_k \{ \log_2(k) - h(x) \geq 0 \} \quad (3)$$

that is, we obtain a distribution whose entropy is greater than or equal to that of the entropy estimate from the time series data. The number of symbols N is chosen to be the minimum k that satisfies Eq. (3).

A difficulty with differential entropy is the ability of this measure to take on a negative value. If this is the case, the algorithm defaults to a selection of two for the number of symbols.

2.2 Partitioning

After the number of partitions has been determined, the next step of the algorithm requires symbolization of the time series data. This step includes the determination of the partitioning structure of the time series data used in the generation of the symbol sequence. This step requires the number of symbols used in the algorithm as well as the partitioning methodology of which includes uniform and Maximum Entropy (ME) partitioning to be determined. The choice of the number of symbols will fix the number of partitions in the algorithm as each partition is assigned a unique symbol as was discussed previously.

The objective of the partitioning is to assign a symbol to each of the time series data points $X \equiv (x_0, x_1, \dots, x_n)$.

Given the set of N symbols, $\Sigma = (s_0, s_1, \dots, s_{N-1})$, each symbol s_i is assigned to one partition P_i , where P is the partitioning of the time series data $P \equiv (P_0, P_1, \dots, P_{N-1})$. Therefore, if $x_i \in P_i, x_i \rightarrow s_i$, that is, we assign s_i to x_i if x_i falls within the bounds of P_i . As mentioned earlier, there are two methods to develop the partitioning P and they are called uniform partitioning and ME partitioning.

Uniform partitioning requires taking the range of the time series data and dividing it into the N mutually-exclusive equally spaced partitions. Each time series data point that falls into one of these N regions is thus assigned a unique symbol.

The other popular method for time series data partitioning is by Maximum Entropy. This partitioning scheme, as hinted by its name, is completed by maximizing the entropy of the resultant symbol occurrence probability. That is, the occurrence probability of the symbols should be uniform in nature.

In order to complete this, the time series data is ordered in magnitude. By grouping the ordered data into subgroups of length X/N , the partitioning structure for ME partitioning is defined. The resultant occurrence probability for these partitions in the baseline case becomes equal. This differs from the results of uniform partitioning as the resultant probabilities are generally not uniform.

In theory, the total number of partitions can range from a simple binary partition to an upper limit defined by the total number of unique samples in X . In the former case, each data point is simply relabeled with a unique symbol.

2.3 Anomaly Generation

With the completion of the partitioning and symbolization, it is left to determine how to quantify an anomaly from changes in the underlying statistical behavior of the system. The deviations in the system are captured through changes in the state occurrence probabilities. In the case of unity depth, D is equal to one, the states that are tracked during life testing are simply the symbol occurrences. In general, if D is not unity, the states of the system consist of permutations of groups of D symbols. In the following, the states are thus the symbol occurrence probabilities as D is set to unity.

In this work, two measures are used to quantify this change and define it as an anomaly A . One is based on a Euclidean distance type measure and the second is based on the Kullback-Leibler divergence (Singh et al., 2010). Both of these measures use the baseline distribution of state probabilities as well as the current distribution to generate an anomaly.

The Euclidean measure is the 2-norm difference between the baseline and current system state probability

distributions. Given the state probability vector p , the Euclidean metric is:

$$A = \|(p_{nominal} - p_i)\|_2 \quad (4)$$

where $p_{nominal}$ is the baseline state probability vector and p_i is the current state probability vector. The baseline SPV is based on the healthy condition of the system such as at the start of use. The other measure implemented in generating an anomaly from the statistical models of the system is the Kullback-Liblier divergence:

$$KL = \sum_{i=1}^N p_i^k \log\left(\frac{p_i^k}{p_i^{nominal}}\right) \quad (5)$$

In (5), the sum is over the total N states in the algorithm while k represents the k^{th} iteration of the algorithm. An anomaly measure is generated from (5) by:

$$A = 0.5 (KL(p^k, p^{nominal}) + KL(p^{nominal}, p^k)) \quad (6)$$

These anomalies are then used to diagnose the current state of the converter as the system degrades from use. From these measures, it is possible to detect degradation or a fault that has occurred in the system.

3. ACCELERATED TESTING OF A DC-DC FORWARD CONVERTER

In order to verify the algorithm, a 50W forward converter was designed, constructed, and placed in an accelerated life test environment. The forward converter used 15 V for input and output 10 V at 5 A nominally. The general circuit

diagram of a forward converter is shown in Figure 1 with the locations of the sensors implemented in the testing. This converter implements the current-mode feedback methodology in addition to output voltage feedback.

It is known that specific components in the dc-dc forward converter are more susceptible to failure than other components. From (Orsagh et al., 2006; Orsagh et al., 2005), the most probable locations of failure for the converter are the MOSFET power switch, the rectifying diodes, and the input and output capacitors.

The accelerated test consisted of placing the converter in an oven to generate a High Temperature Operating Life (HTOL) test. This test is geared to ascertain the usable life of a system by continually running the system at high environmental temperatures. In this case, the converters were continually run at 85°C. This temperature point coincides with the maximum operating temperature of several components in the converter. These components included the Pulse-Width-Modulator (PWM) controller, input/output electrolytic capacitors, and several other integrated circuits.

The high temperature was used to accelerate failures in the dc-dc converter. For example, the electrolytic capacitors contained in the circuit would be directly affected by operating temperature. The higher temperature would cause acceleration in the loss of electrolyte in the capacitor causing wear out (Kulkarni et al., 2009). This in turn would cause an increase in the capacitors equivalent series resistance (ESR).

Additionally, the power MOSFET failure mechanism of Time Dependent Dielectric Breakdown (TDDB) can be accelerated by higher temperatures (Kalgren et al., 2007).

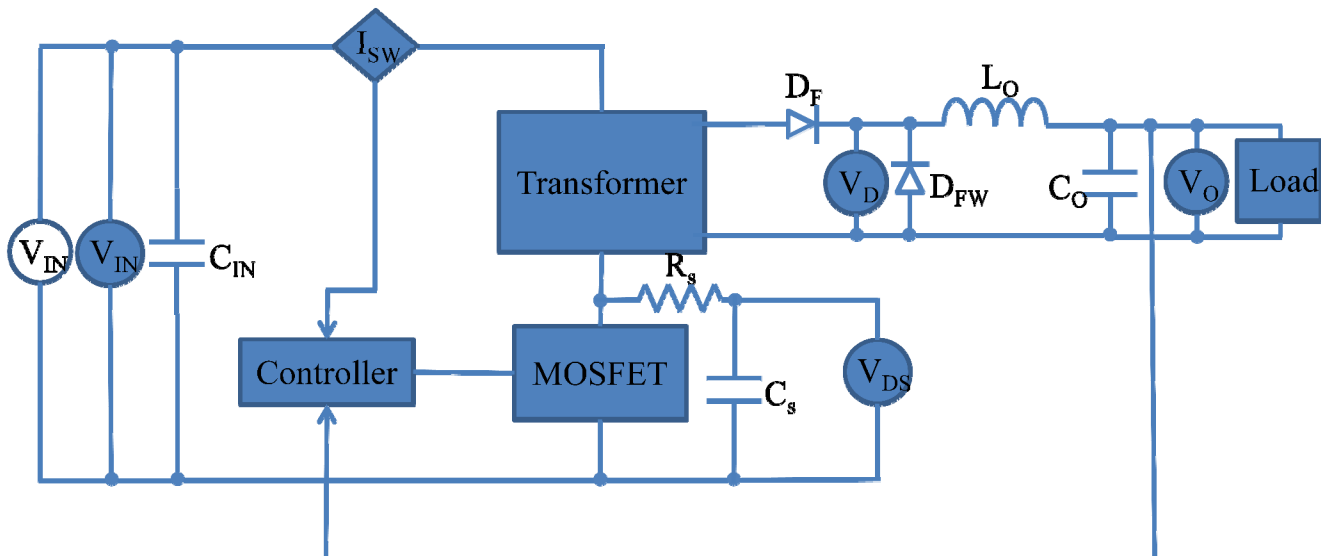


Figure 1: Simplified Diagram of a dc-dc Forward Converter with Sensor Locations

Another failure phenomenon that can be accelerated with higher temperatures is electromigration affecting both MOSFETs and rectifying diodes. These two components are also susceptible to degradation caused by interdiffusion.

3.1 Sensor Placement

Although each component will degrade individually and sensors have been placed at these components to observe degradation, the objective of this research is not to specifically identify the failure mode of a specific component. Instead, the objective is to capture data across locations in the converter that have high probability of degradation and/or failure to generate a diagnostic measure for the system as a whole.

Sensors were placed in the circuit that would maximize the probability of observing degradation in the components mentioned above. In addition, the input voltage was also monitored throughout the testing.

Voltage sensors were placed across the power MOSFET, the freewheeling diode (D_{FW}), and at the output voltage, V_O . In addition, the current signal produced from the current sensor for current-mode feedback was also used in the degradation monitoring. Since this signal was already being monitored for current mode control, it provided an easy means to access the instantaneous switch current waveform.

Thermocouples were placed on the tabs of the TO-220 package for the MOSFET and diode D_{FW} for monitoring as well. During each time series data capture, one sample each was taken of the MOSFET, D_{FW} , ambient, and oven temperatures for monitoring during testing.

Data for the SD algorithm was recorded from these sources at a rate of 800 KS/s. The data acquisition hardware was triggered every hour to record a data snapshot of 0.25s in duration. No anti-aliasing filters were implemented in the data acquisition hardware. The data channels were buffered into NI 9221 analog input modules.

Anti-aliasing filters were not implemented as the filtering function could remove degradation information from the signals. Since the sampling rate is approximately eight times the switching frequency of the converter, the anti-aliasing filters would have filtered too much of the frequency spectrum of the signals. The anti-aliasing filters would remove significant energy from the spectral content of the time series data. Given that our objective is to not recreate the time series data, it is acceptable to have a limited sample rate on large bandwidth signals. More research is currently being performed on the affects of the low rate sampling on the performance of the symbolic dynamics algorithm.

3.2 Life Testing

The forward dc-dc converter was placed into the temperature chamber and allowed to function until failure of the converter. Failure was defined as failure to maintain desired output voltage within 10% of the set point or as the result of complete failure.

For testing, the converter was loaded with a bank of 0.5 Ω resistors used to create a 1 Ω load for the converter. The voltage output of the converter was set at 9.5 V across the 1 Ω load. This resulted in a continuous output power of about 90 W. The converter would then be continually loaded at this power level while under the HTOL testing. Further research will investigate the effects of changing load on the results of the methodology.

The converter had a 24 hour burn in procedure to confirm functionality of the converter and data acquisition systems. This period also allowed the system to reach an operational steady state before the stress testing began. The temperature of the oven during burn in was 65°C. After this period of time, the accelerated testing was started. The temperature of the oven was increased to 85°C at this time. This temperature was selected due to the operational temperature constraints of the onboard electrolytic capacitors and integrated circuits (ICs).

The first converter was operated for 200 hours after the burn in period was completed. At this time, the converter failed by not being able to maintain the desired output voltage. Post failure analysis pointed to the input capacitors as the failed components. Table 1 shows the pre and post test conditions of all electrolytic capacitors in the converter which demonstrate the degradation experienced by the capacitors

Capacitor	Pre-Test		Post-Test	
	C (μ F)	DF	C (μ F)	DF
C1	455	0.049	415	0.487
C2	449	0.05	241	1.24
C3	449	0.047	128	1.96
C4	204	0.057	200	0.057
C5	204	0.058	199	0.058

Table 1: Capacitor Characterization for Converter Test 1

In the table, C1-C3 were the input capacitors (C_{IN} in Figure 1), C4 was the output voltage capacitor (C_O), and C5 was used as a filter for a negative voltage bus in the converter (not shown in Figure 1). DF in the table is the Dissipation Factor of the capacitors and is related to the loss tangent for dielectrics. The higher the DF value results in a larger magnitude of the ESR component of the electrolytic

capacitor generating higher internal power losses in the capacitor. The relationship between DF and ESR is:

$$DF = \frac{ESR}{|X_C|} \quad (7)$$

where X_C is the reactance of the capacitor under the test at the known test frequency.

As is visible in the table, capacitors C2 and C3 suffered severe damage from the testing. This is representative in both the reduction of the capacitance (nominally 470 μ F) and the increase in the DF measure of both capacitors. Confirmation of the failure was completed by restoring the converter to normal functionality by replacement of these capacitors (C1 though C3). Replacement of these capacitors restored the output voltage capability of the converter.

Similarly, a second test was carried out under the same test conditions for failure repeatability with a new converter. This test lasted approximately 1,800 hours at the 85°C test at which point the test temperature was increased by 10°C in order to accelerate the test. At this temperature point, the converter functioned for another 152 hours.

After failure, it was determined that the failure was again the input capacitors of the converter. Table 2 shows the capacitor characterizations before and after the testing.

Capacitor	Pre-Test		Post-Test	
	C (μ F)	DF	C (μ F)	DF
C1	434	0.043	377	0.617
C2	434	0.043	371	0.694
C3	427	0.045	363	0.800
C4	203	0.051	194	0.087
C5	203	0.052	143	0.53

Table 2: Capacitor Characterization for Converter Test 2

Test 1 and test 2 capacitors showed some signs of the top of the canisters bulging. This is most likely related to loss of electrolyte through evaporation due to internally generated heat in the capacitor.

The difference in test lengths is most likely due to component differences in the converters such as those from different lots. The capacitors used in the converters were from the same manufacturer but not from the same production lot. The tables also demonstrate the amount of degradation the capacitors incurred during testing specifically in terms of the dissipation factor. In terms of the data derived method, the difference in test lengths will not negatively affect the performance of the algorithm as will be seen in the upcoming sections.

The other components observed during testing (MOSFET and rectifying diodes) did not show significant changes in parameters after testing. Parameters tested for the MOSFET included $V_{GS,th}$, the gate threshold voltage, approximate $R_{ds,on}$, gate leakage current, and BV_{DSS} , the maximum drain to source voltage. The rectifying diode parameters included V_{FW} , the forward voltage, and the maximum cathode-anode voltage. All of the above parameters recorded minimal changes from pre to post-testing.

4. RESULTS

Once the data collection was completed with the failure of the converters, the SD algorithm was implemented on the captured data sets. The goal of the algorithm is to generate anomalies using the collected data that can be used to determine the state of health of the converter.

The SD algorithm results are compared to features that are commonly used to monitor the health of electronics. The estimated efficiency of an electronic system has been used to determine the current state of health of the system as a loss of efficiency is a sign of system degradation (Orsagh et al., 2005). Efficiency can be monitored through implementation of sensors on the input and output ports of the system to monitor current and voltage. As the components in the system begin to degrade, they tend to have more internal power loss that directly affects the converter's overall efficiency. This degradation can be tracked through the computation of the system's efficiency.

When the testing of the converters was first started, it was not anticipated that an efficiency measure would need to be calculated so input and output currents were not measured. However, input and output voltage was measured and switch current was also monitored. From these three variables plus knowledge of the load enabled efficiency to be estimated from the captured converter signals.

From the captured data, the input current had to be estimated from the captured switch current. This required the duty cycle to be estimated from the data captured from the converter. Once the duty cycle was estimated, the current was scaled by the duty cycle and the mean taken from current data when the switch is ON. This was calculated as:

$$I_{in} = \text{mean}(i_{sw,ON} * D) \quad (8)$$

where $i_{sw,ON}$, is the switch current during the ON interval and D is the duty cycle.

Another feature to be compared to the SD algorithm is related to the output voltage ripple of the converter. As the output capacitor degrades, the ESR of the capacitor tends to

increase causing more output voltage ripple. To attempt to capture this effect, a form factor metric given as:

$$F_F = \frac{V_{O,RMS}}{V_{O,MEAN}} \tag{9}$$

was implemented to track the output ripple characteristics.

Symbolic dynamics was applied to the time series data recorded from the accelerated testing of the converters. Since there was uncertainty in which signals would produce the best results, the algorithm was implemented on all signals using the automatic selection methods discussed in Section 2.2.

From the results of the algorithm implementation, it was discovered that the signals containing the best degradation trending was the diode voltage (DFW in Figure 1). An example of the diode voltage is shown in Figure 2. This data was taken from the first converter test.

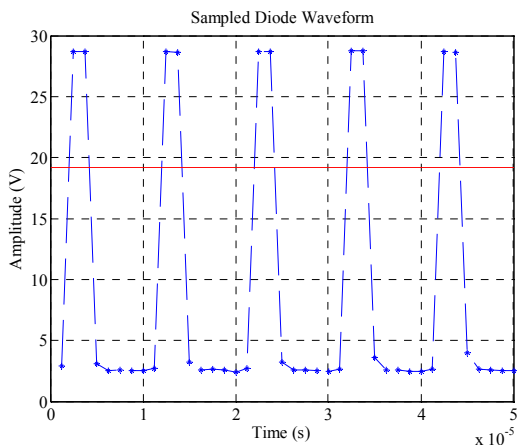


Figure 2: Sample Diode Data with Partitioning (Red) - Healthy

As seen in the figure, the data is sampled at 800 kHz resulting in eight data points per cycle in the waveforms. The binary partitioning implemented in this analysis generates an interesting result. The upper partition probability of occurrence is the duty cycle of the converter. In this case, the algorithm automatically defaults into a duty cycle detector and tracker.

The diode’s voltage works well as the wave shape of the voltage is a pulse waveform in nature. The pulse wave shape enables a direct correlation of duty cycle of the converter to the converter’s current operating condition. The duty cycle of the converter is a good feature to use for converter health. As the converter degrades, in order to maintain the current output power, the duty cycle must be perturbed slightly larger. The duty cycle needs to increase because as the converter degrades the efficiency of the converter also decreases as internal components begin to

become more lossy. The efficiency of test 1 over the complete interval is seen in Figure 3.

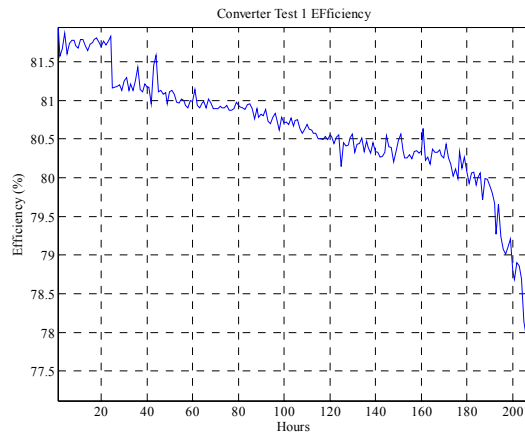


Figure 3: Test 1 Efficiency over Accelerated Testing

The efficiency of the converter decreases throughout the accelerated converter testing. To overcome the additional losses in the converter, the closed loop control perturbs the duty cycle to maintain output power. The duty cycle for test 1 is shown in Figure 4.

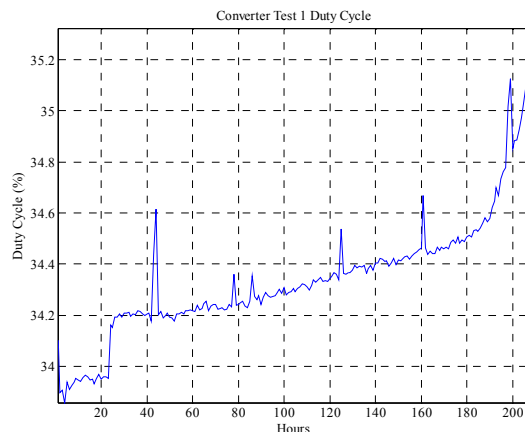


Figure 4: Converter Test 1 Duty Cycle over Accelerated Testing

As the testing progresses, the increasing degradation in the system causes the duty cycle to be increasingly perturbed. From the plot, the converter started at approximately 34% and failed when the duty cycle reached just over 35%.

Figure 5 shows the captured diode data after 200 hours of degradation also from the first test.

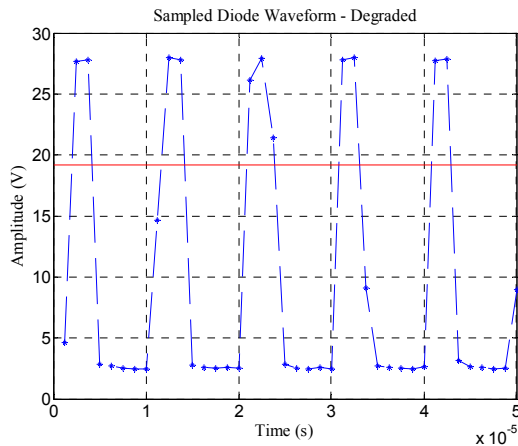


Figure 5: Sample Diode Data with Partitioning (Red) – Degraded

In Figure 5, note that the same partition is used in the example. The partitioning must remain invariant across the lifetime of the system for proper operation. Also note that due to the degradation, there is an additional data point in the upper partition. This in turn increases the upper partition’s occurrence probability which can be interpreted as an increase in the duty cycle of the converter.

4.1 First Converter Test

The results in Figure 6 are from the captured diode data using the binary partitions shown in Figure 2 and 5. The anomaly was generated from the state probability vector where the states are the partitions themselves (depth was set to 1). The anomaly metric used in the figure was from (4). The baselines used in all the cases were from the initial start of the burn in. It is possible to use any point in the test for the baseline. In this case it was convenient to use the first set of captured data.

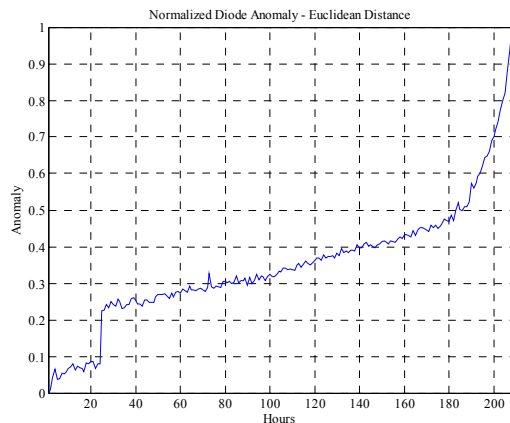


Figure 6: Anomaly Measure Generated from Diode Data – Euclidean Distance Metric – Test 1

In Figure 6, the jump in anomaly at 24 hours was a result from the end of the burn in period leading into the start of the accelerated testing. In general, the anomaly increases steadily until approximately 180 hours into the test where the degradation accelerates rapidly. The last data point was taken just over 200 hours when the converter failed.

The following figure combines the SD anomaly of Figure 6 with those obtained from an efficiency calculation and from the output voltage form factor.

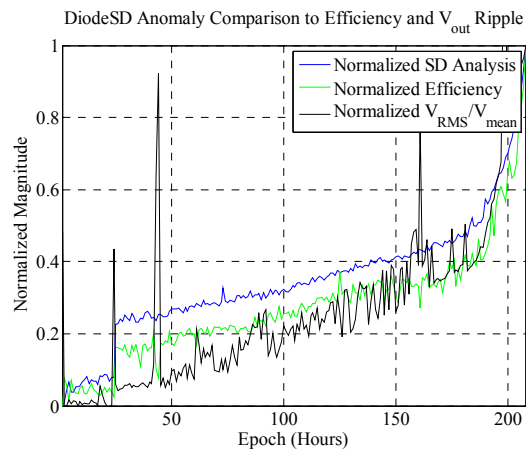


Figure 7: Comparison of Symbolic Dynamics Anomaly, Estimate Converter Efficiency, and Form Factor

As seen in Figure 7, the three measures compare well with one another. All three measures show some effect from the break in period into the accelerated testing. In this example, it is clear that Symbolic Dynamics reproduces the results of the other metrics with minimal effort. Additionally, the SD generated anomaly has less noise as compared to the other two measures over the complete test period. Forward thinking, this result should be positive for use in a prognostics sense with these converters

4.2 Second Converter Test

The testing was repeated with a second converter to reproduce the results seen above. The converter was again tested with a 24 hour burn in period and then left to be operated at 85°C until failure.

This test also resulted in failed input capacitors; however, the complete test lasted approximately 1,800 hours. Symbolic dynamics was again implemented on the diode data and the results are shown in Figure 8.

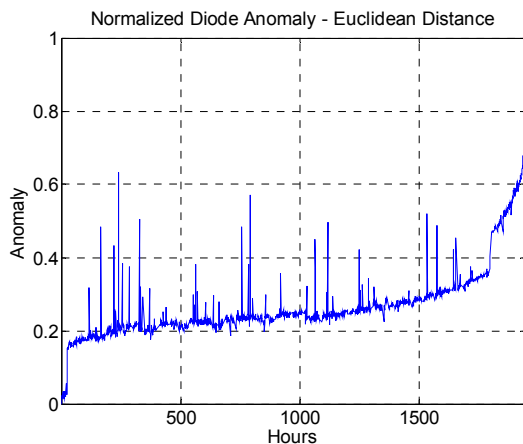


Figure 8: Anomaly Measure Generated from Diode Data – Euclidean Distance Metric – Test 2

In Figure 8, the test lasted significantly longer than the previous test. However, the results are very similar with the anomaly trend increasing rapidly toward failure. The jump in anomaly in the beginning is due to the break in interval while the jump towards the end (around 1,700 hours) was due to a change in the test parameters. At that point, the temperature of the test was increased from 85°C to 95°C in order to further accelerate the testing.

Again, we aim to compare our results to metrics more commonly used to diagnose the health of an electronic system. Using the same data, the efficiency and form factor metrics were calculated and the results are shown in Figure 9.

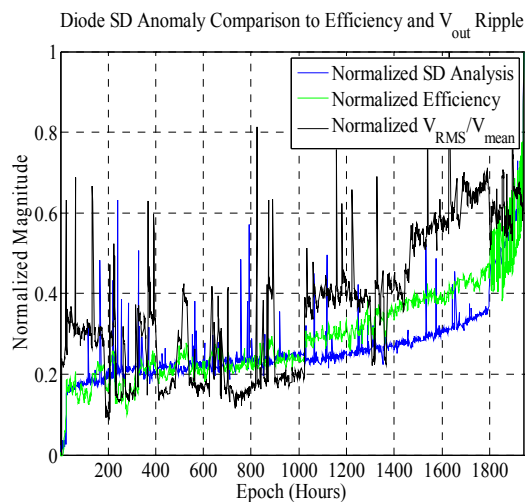


Figure 9: Comparison of Symbolic Dynamics Anomaly, Estimate Converter Efficiency, and Form Factor

As is observable, each of the measures are susceptible to the jump in operating temperature both from the break in period and from the increase in test temperature near the end of the test.

As compared to the results from the first test, the form factor metric does not produce as clear a trend as compared to the SD or the efficiency result. It would be difficult to determine current state of health from this trend.

Efficiency is a consistent metric between the two tests and is relatively easy to calculate. However, it does require one to record the input and output characteristics of the converter during operation whereas the SD methodology only requires the monitoring of one channel. The SD metric in both cases also produces a consistent degradation metric that could be used for diagnostics.

5. CONCLUSION

This paper proposes a data derived approach to monitoring dc-dc converters for degradation during operation. The methodology is based on Symbolic Dynamics that converts the captured time series data into a symbolic series that is analyzed statistically. The statistical results are then used to generate an anomaly based on the current operating conditions of the converter as compared to a known baseline.

The algorithm was tested on data recorded from two dc-dc forward converter tests. The aim was to capture degradation trends from the converters by monitoring the input voltage, switch current, MOSFET drain to source voltage, the output freewheel diode voltage, and the output voltage. It was determined that the diode voltage was the most sensitive to the internal degradation of the converter.

The generated anomaly from the SD algorithm was compared to the overall efficiency of the converter as well as the form factor of the output voltage. The form factor metric aims to capture the change in the output voltage ripple related to degradation of the output electrolytic capacitor.

The results show a consistent trend generated from both the SD anomaly and the efficiency of the converter. The form factor was inconsistent in generating trends between the two tests.

Future work will focus on effects to the algorithm from loading changes as well as further investigation into the effects of the different parameters in the Symbolic Dynamics algorithm. It was also determined that temperature deviations affect the data derived method which requires further investigation. Investigation of using the generated trends for prognostication will also be researched. The trends produced from testing currently have generated trends that we believe are applicable for life prediction.

REFERENCES

- Brown, D.; Kalgren, P.W.; Roemer, M.; Dabney, T. (2006) Electronic Prognostics – A case study using switched-mode power supplies (SMPS). *2006 IEEE Autotestcon*,

pp. 636-642.

- Conrad, K. (2011) Probability Distributions and Maximum Entropy, <http://www.math.uconn.edu/~kconrad/blurbs/analysis/entropypost.pdf>, visited April 12th, 2011.
- Hess, A.; Calvello, G.; Frith, P. (2005) Challenges, Issues, and Lessons Learned Chasing the “Big P”: Real Predictive Prognostics Part 1. *2005 IEEE Aerospace Conference*, pg. 1 – 10.
- Jin, X.; Gou, Y.; Sarkar, S.; Ray, A.; Edwards, R. (2011) Anomaly Detection in Nuclear Power Plants via Symbolic Dynamic Filtering. *IEEE Transactions on Nuclear Science*, vol. 58, no. 1.
- Kalgren, P.W.; Baybutt, M.; Ginart, A.; Minnella, C.; Roemer, M.J.; Dabney, T. (2007) Application of Prognostic Health Management in Digital Electronic Systems. *2007 IEEE Aerospace Conference*, pp. 1-9.
- Kulkarni, C.; Biswas, G.; Koutsoukos, X. (2009) A prognosis case study for electrolytic capacitor degradation in DC-DC converters. *Poster Presentation 2009 PHM Society Conference*.
- Orsagh, R., Brown, D., Roemer, M., Dabnev, T., Hess, A. (2005). Prognostic Health Management for Avionics System Power Supplies. *2005 IEEE Aerospace Conference*, pp. 3585 – 3591.
- Orsagh, R.; Brown, D.; Kalgren, P.; Byington, C.; Hess, A.; Dabney, T. (2006) Prognostic Health Management for Avionic Systems. *2006 IEEE Aerospace Conference*.
- Ray, A. (2004). Symbolic Dynamic Analysis of Complex Systems for Anomaly Detection. *Signal Processing*, v84, pg. 1115 – 1130.
- Rajagopalan, V., Ray, A. (2006). Symbolic Time Series Analysis via Wavelet Based Partitioning. *Signal Processing*, v86, pg. 3309 – 3320.
- Singh, D.S.; Gupta, S.; Ray, A. (2010) Symbolic Dynamic Analysis of Surface Deformation during Fatigue Crack Initialization. *Measurement Science and Technology*, vol. 21, no. 3.
- Rohan, S.; Rajagopalan, V.; Ray, A. (2006) Wavelet-based Symbolic Analysis for Detection of Broken Rotar Bars in Inverter-fed Induction Machines. *2006 American Control Conference*.

Using the Validated FMEA to Update Trouble Shooting Manuals: a Case Study of APU TSM Revision

Chunsheng Yang, Sylvain Létourneau, and Marvin Zaluski

*Institute for Information Technology, National Research Council Canada
Ottawa, Ontario K1A 0R6, Canada*

Chunsheng.Yang@nrc-cnrc.gc.ca, Sylvain.Letourneau@nrc-cnrc.gc.ca, and Marvin.Zaluski@nrc-cnrc.gc.ca

ABSTRACT

Trouble Shooting Manuals (TSMs) provide useful information and guidelines for machinery maintenance, in particular, for fault isolation given a failure mode. TSMs produced by OEMs are usually updated based on feedback or requests from end users. Performing such update is very demanding as it requires collecting information from maintenance practices and integrating the new findings into the troubleshooting procedures. The process is also not fully reliable as some uncertainty could be introduced when collecting user information. In this report, we propose to update or enhance TSM by using validated FMEA (Failure Mode and Effects Analysis), which is a standard method to characterize product and process problems. The proposed approach includes two steps. First, we validate key FMEA parameters such as Failure Rate and Failure Mode Probability through an automated analysis of historical maintenance and operational data. Then, we update the TSM using information from the validated FMEA. Preliminary results from the application of the proposed approach to update the TSM for a commercial APU suggest that the revised TSM provides more accurate information and reliable procedures for fault isolation.*

1. INTRODUCTION

TSMs are useful resources for machinery maintenance, in particular, for fault isolation given a failure mode. Fault isolation in a complex system involves identifying a root contributing component or ranking the contributing components given a failure mode. This is generally complicated and time consuming. The TSM guides the technician through the process by providing a potential list of causes along with procedures to be executed in order to identify the fault(s) and fix the failure. The list contains a set of possible components and corresponding maintenance procedure. However, these components or causes are

typically not ranked and this introduces ambiguity during fault identification. In other words, without the ranking information, the technician has difficulty to decide which component should be first investigated in isolating the contributing component. We believe that enhancing the TSM with an ordered list of components based on experiences from historical maintenance would help increase efficiency. To achieve this objective, we propose to validate FMEAs based on historical operation and maintenance data and then use the validated information to revise and enhance the TSM.

Failure Mode and Effects Analysis (FMEA) models are available for a wide variety of machineries. They provide a foundation for qualitative reliability, maintainability, safety, and logistic analysis by documenting the relationships between failure causes and failure effects. In particular, FMEA models contain useful information such as Severity Class (SC), Failure Rate (FR), and Failure Mode Probability (FMP) for determining the effects of each failure mode on system performance. Our intent is to exploit such information to update and enhance TSM. However, since FMEAs are produced at design time and then hardly validated after deployment of the corresponding system, there is a risk that the information provided is incomplete or no longer accurate. The likelihood for such inaccuracies is particularly high for complex systems such as aircraft engines that operate over a long period time. In such cases, using the initial FMEA information without adequate validation could result in the introduction of irrelevant recommendations. To avoid this issue, the initial FMEA information needs to be validated and then updated as required. In Yang et al. 2009, we proposed a process to validate FMEAs using real-world readily available maintenance and operational data. In particular, we investigated validation of a FMEA for an APU (Auxiliary Power Unit engine). To constrain the study, we focused on components related to the "Inability to Start" failure effect. In this work, we explore the use of the validated FMEA to enhance the sections in the TSM that are related to the same problem (i.e., APU Inability to Start). Our objective is to

* This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

order the contributing components listed in these TSM sections and modify the corresponding procedures based on the parameters from the validated FMEAs.

The next section provides an overview of validation of FMEA using historical operational and maintenance data. Then we present the TSM revision by applying the validated FMEA information. In Section 4, we discuss the results. The final section concludes the paper.

2. OVERVIEW OF VALIDATING FMEA

APU FMEA documents used in this study were provided by the OEM. As usual, the FMEA was created during the design phase. It contains typical FMEA information: failure effect, failure mode (failure identification), failure cause, contributing components, symptoms, functions, corrective actions, Failure Rate (FR), Severity Class (SC), Mean Time between Failures (MTBF), and Failure Mode Probability (FMP). For validation purpose, we focus on key quantitative parameters such as SC, FR, and FMP. The validation process (Yang et al, 2009) combines database retrieval and data mining techniques to automatically adjust the initial values based on actual experiences as recorded within the maintenance database.

Figure 1 illustrates the proposed approach for FMEA validation. The various tasks can be grouped into three main phases:

1. Obtain the failure events from maintenance database given a failure mode in FMEA
2. Gather the relevant data for the failure events and conduct the statistical analysis for APU usage time
3. Update the FMEA parameters using statistical information

2.1 Obtaining Failure Events

The goal is to retrieve information for all relevant failure events or component replacements from the maintenance database that relate to the given failure effect. In this case, we want to retrieve all occurrences of replacement of components that relate to the APU “Inability to Start” effect. The components of interest are the ones identified in the FMEA as contributors to the failure effect “Inability to Start”. As we mentioned in previous section, retrieving these components is a difficult task for a number of reasons: part numbers change over time and we often ended up with several numbering schemes, data entry errors or omission errors, technicians’ personal preference when entering part names when referring to a given component, and sometimes a component is mentioned in the textual description of the repair without being actually replaced. For example, in the database, we found that “ignitor”, “igniter”, “ignitor plug”, “ignition exciter” and “ignition unit” are all use to refer to the component “Igniter”. All of these difficulties need to be taken into account when establishing part names (part description) and part IDs for a given component.

The second step uses the part numbers and part names identified to retrieve all occurrences of replacement of the given part (the so-called failure events) from the maintenance data. This step results in a list of occurrences of part replacements with detailed event information (e.g., repair date, aircraft identification number, and reason for replacement). Further validation is needed to remove duplicates and irrelevant entries from the list of occurrences.

Next, we analyze the maintenance history around each occurrence of replacement in order to get insights on other potentially related fixes (or components). In this work, we considered all APU maintenance repairs in the 60 day interval around each replacement event (i.e., up to 30 days before the given replacement and up to 30 days after the replacement). A number of software tools were developed to help automate the search but manual validation is still needed.

Table 1 shows the preliminary results obtained. The left column lists the components contributing to the failure effect considered (“Inability to Start”) based on the FMEA.

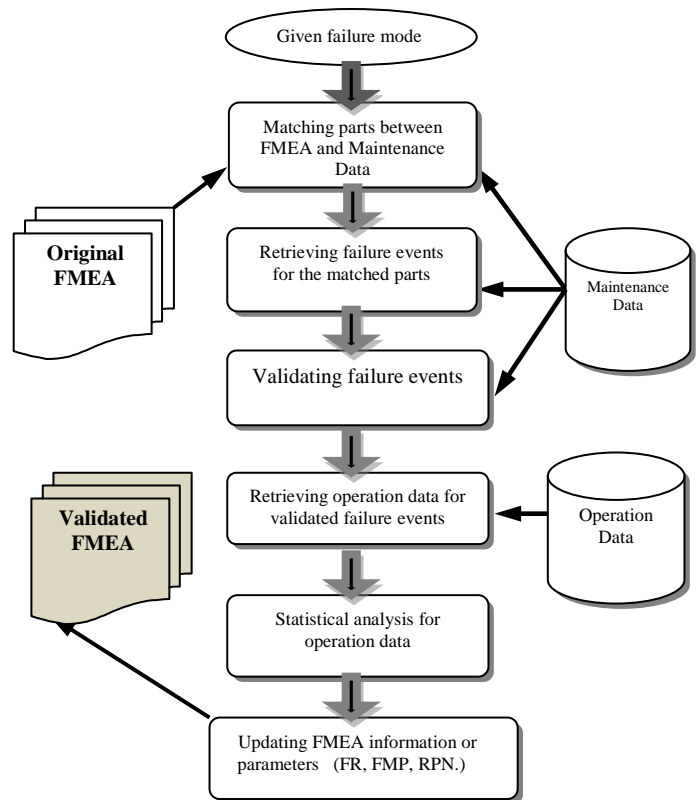


Figure 1. The procedure of FMEA validation

The other three columns show the number of replacement occurrences found using the part ID only and the part name only, respectively. From Table 1, we observe that we have been able to retrieve a significant number of occurrences of

replacement for some FMEA components contributing to the selected failure effect. However, very few or even no replacements have been found for other FMEA contributing components such as Fuel Manifold and O-Ring Seal. This is surprising as the operator's maintenance database covers more than 10 years of operation for a fleet of over 100 aircraft. A couple of hypotheses may be proposed to explain this situation. It is possible that some of the contributing components mentioned in the FMEA simply never failed during the period of maintenance data. Since the FMEA APU and APU used in the study is not the same model, it is also possible that some of the contributing components mentioned in the FMEA do not exist in the APU used in the study.

For the rest of the analysis, we focused on the contributing components which were actually replaced and ignored the other components.

FMEA Part	AMTAC data		
	Identified Instances of Part Replacements (Failures)		
	by Part Number	by Part Description	Total Failures (N _{Fc})
Starter	49	158	207
Igniter	16	140	156
Fuel Control Assembly	46	19	65
Fuel Flow Divider	9	5	14
Low Oil Pressure Switch	1	10	11
Fuel Pump	19	6	25
EGT Thermocouple	0	1	1
Monopole Speed Sensor	1	3	4
Oil Pump Assembly	0	4	4
Isolation Valve	0	0	0
Oil-Ring Seal	0	0	0
Fuel Manifold	0	0	0

Table 1. Instances of replacements for components for failure effect, 'Inability to Start'

2.2 Data Analysis for APU Usage Time

In order to compute statistics about actual failure rate, we need to determine the cumulative usage of the entire fleet of APU over the period covered by the maintenance data. This is done by retrieving the most recent value of the *APU_OPERATING_HOUR* parameter, which is automatically reported as part of the APU starting report, for each APU and then adding all values. For the dataset considered, we obtained a total APU usage of 4,328,083 operating hours (noted as *UT*). In the later section, we use this life consumption of APU engine when updating the FMEA parameters.

2.3 Updating FMEA Parameters

As mentioned before, we are interested in updating quantitative FMEA information, such as FR, FMP, SC, and MTBF. We also considered the "Risk Priority Number" (RPN) (Sellappan and Sivasubramanian 2008, ASENT 2009), which is defined as the product of SC, FMP, and FR. The RPN is a measure used when assessing risk to help identify critical component associated with the failure effect. A large RPN indicates that the given component is more likely to need replacement. The left hand side of Table 2 presents the values for all parameters of interest for each component for which we have been able to retrieve examples of replacements from the maintenance data. Based on RPN, most occurrences of APU "Inability to Start" problems should be resolved by replacing either the "Igniter" or the "Monopole Speed Sensor". However, when considering the number of actual replacements (*N_{Fc}* in Table 1), we notice that the "Starter" comes first, followed by the "Igniter" and the "Fuel Control Assembly". Moreover, the "Monopole Speed Sensor" which was one of the first components to be suspected based on original FMEA is almost never replaced by the maintenance crew (only 4 replacements as reported in Table 1). Such discrepancies between the original FMEA information and real maintenance practice clearly show the need for regular updates of the FMEA information.

We propose to update the FMEA information by relying on data acquired as part of normal operation. First, to update the FR and FMP parameters based on actual maintenance history, we introduce the following equations

$$FR = \frac{N_{Fc}}{UT} \quad \text{--- (1)}$$

$$FMP = \frac{N_{Fc}}{RN} \quad \text{--- (2)}$$

where:

- *N_{Fc}*: The number of replacements of a given component (Table 1);
- *UT*: The total APU usage (in hours) for the entire fleet; it is 4,328,083 hours in this study;
- *RN*: The total number of APU parts replaced during the investigation. It is a sum of *N_{Fc}* in Table 1. In this study, *RN* = 487.

The last four columns in Table 2 show the revised information. FMP and FR are computed from Eq. (1) and Eq.(2) using NFc from Table 1 and UT obtained as described above. RPN is recomputed using the revised parameters. The revised RPN results closely reflect the real maintenance practice. We also add ranking information based on RPN. The larger RPN number is associated with a higher ranking (a smaller value of ranking). The ranking parameter is useful for component ranking during fault identification as described in next section. We believe that the revised information, although quite different from the original number, are more representative of real world practice and therefore potentially more appropriate for decision-based support system to assist the operator in the maintenance of the APUs.

3. TSM REVISION

Troubleshooting is the process of diagnosing the source of a problem. It is used to fix problems with physical components or subsystems in a complex system. The basic theory of troubleshooting is that you start with the most general (and often most obvious) possible problems, and then narrow it down to more specific issues.

In this study, the APU TSM is provided by an OEM to enable the systematic identification, isolation and correction of aircraft warnings and malfunctions reported in flight and on the ground.

Like all TSMs, the provided APU TSM is a highly structured document designed to help identify and

isolate the fault by performing prescribed procedures. There is at least one chapter for each failure effect and each chapter contains 4 sections:

- Possible Causes,
- Job Set-up Information,
- Fault Confirmation, and
- Fault Isolation Procedure (FIP).

Appendix A is an example of the original TSM chapter. Given a failure mode, the Possible Causes section lists the possible components which may contribute to the given failure mode or effect. This list is not ordered and has no priority for each component. Therefore, it is difficult for the end user to decide where to start the investigation. Most mechanics perform troubleshooting based on the symptoms, TSM, and their experiences. They use a sequential trial and error approach with guidance from the TSM until a solution is found. The Job Set-up Information section lists the AMMs (Aircraft Maintenance Manuals), which may relate to the FIPs and provides the detail instructions for installing or removing a contributing component. The Fault Confirmation section advises technicians how to check and test the failure symptoms in order to confirm the failure effects. Finally, the FIP section lists the ordered procedures for fixing failures. Depending on the type of failure, the problem symptoms could lead into a lengthy troubleshooting session especially when addressing intermittent

Component Name	Original APU FMEA Information						Updated FMEA Information			
	SC	FMP (%)	FR	MTBF (hours)	RPN	Old Rank	FMP (%)	FR	RPN	New Rank
Starter	4	1.96	9.75	500,000	0.76	7	41.4	47.61	78.84	1
Igniter	3	16.67	27.78	36,000	13.89	1	31.2	35.88	33.58	2
Fuel Control Assembly	3	16	20	50,000	9.60	3	13	14.95	5.83	3
Fuel Pump	3	0.02	2.0	500,000	0.00	9	5	5.75	0.86	4
Fuel Flow Divider	3	0.8	20	50,000	0.48	8	2.8	3.22	0.27	5
Low Oil Pressure Switch	4	4.44	22.22	45,000	3.95	4	2.2	2.53	0.22	6
Monopole Speed Sensor	3	20.0	20.0	50,000	12.00	2	0.8	0.92	0.02	7
Oil Pump Assembly	3	4.25	17.0	58,824	2.17	5	0.8	0.92	0.02	8
EGT Thermocouple	2	5.0	20.0	50,000	2.00	6	0.2	0.23	0.001	9

Note: (1) Risk Priority Number = SC · FMP · Rate; (2) Failure Rate (FR) is failures in million hours; (3) The shaded columns show the updated parameters.

TABLE 2. Updated parameters for APU FMEA (See Failure Effects “Troubleshooting”) (continued)

failures. To reduce costs and improve the efficiency of fixing a failure, it is expected that TSM can provide relatively accurate and accountable FIPs, such that technicians can quickly isolate the contributing components for fixing the failure.

There exist two issues with TSMs. First, the Possible Causes are not ordered and have no priority information. Second, the FIP may become out of date with respect to the aircraft and ultimately provide an inappropriate procedure for troubleshooting or fault isolation. In this work, we attempted to set an order for the Possible Causes and modify the FIP to reflect the historical maintenance experiences when suggesting a troubleshooting procedure. In particular, we update the Possible causes and FIP by using the “new rank” information from Table 2. We now detail this procedure.

3.1 Procedure of TSM Updating

The developed procedure for updating TSM based on the validated FMEA contains the following three steps.

1. Retrieve the relevant TSM standard chapters for the failure mode or effect of interest.
2. Verify that the order of the possible causes in the TSM corresponds to the ranking obtained with the validated FMEA. In case of discrepancies, update the Possible Causes section so that components are presented in the same order as shown in Table 2.
3. As needed, also align the FIP orders in the TSM with the ranking provided by the validated FMEA.

We repeat these steps for all chapters in the TSM that relate to the failure of interest.

3.2 The Preliminary Results

Following the procedure above, we updated the chapters related to the failure mode “Inability to Start” in APU TSM document. We first retrieved all chapters. There are 17 chapters in APU TSM document. Among them, only ten chapters contain the contributing components which appear

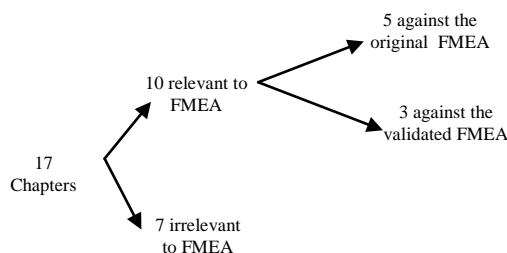


Figure 2. The TSM chapters for “Inability to Start” failure model

in FMEA document. We focused on these ten chapters. Second, we checked the consistency between the TSM and the validated FMEA or the original FMEA for those ten chapters. The Figure 2 shows the result of the ten chapters against the validated FMEA and the original FMEA. For the original FMEA the “old rank” data is used; for the validated FMEA, the ‘new rank’ data is used. Both “rank” data are from Table 2.

Finally, we updated the three chapters where we found discrepancies with the validated FMEA following the steps explained above. Appendix B shows the results of this process when applied to the original TSM chapter shown in Appendix A. After the revision, the order for the Possible Causes became:

- **IGNITER PLUG**
- **GNITION UNIT P10**
- **FUEL CONTROL UNIT**
- **OIL PUMP**
- **FLOW DIVIDER AND DRAIN VALVE ASSY**
- **PRIMARY FUEL NOZZLE AND MANIFOLD**
- **SECONDARY FUEL NOZZLE AND MANIFOLD**
- **ECB (59KD)**

We also revised the sequences of FIPs by changing the procedure of replacing OIL PUMP to follow the new rank of the validated FMEA. We highlighted the changes in *italics* in Appendix B. Such revisions of the TSM FIPs will improve the maintenance efficiency by reducing irrelevant component replacements and also potentially help with planning/scheduling troubleshooting (e.g., right people, right parts). For example, every time the root contributing component is FUEL CONTROL UNIT, the revised TSM will allow the technician to converge to the solution by investigating three components instead of four as initially recommended by TSM FIPs. Since this component fails relatively frequently, this simple change may lead to significant gain in efficiency over time.

4. DISCUSSION

TSM updating depends on the validated FMEA. Most FMEAs are created during the design phase of a system or product and the information may not be accurate enough for practical maintenance decision support system. FMEA should be regularly updated and validated in order to accurately reflect the fleet operation reality. This updated FMEA would provide more reliable and accurate information to enhance the TSM revision.

We only demonstrated the feasibility to update TSM documents using the validated FMEA by showing one failure mode, the ‘Inability to Start’. There are a large number of failure modes in TSM documents. Trying to update the full TSM would represent a significant undertaking. The main challenge comes from the

validation of FMEA. In particular, as noted in the paper, the processing of large amounts of historical maintenance data, which are often characterized by incompleteness and ambiguities, is time consuming and difficult to automate. This might be addressed by integration of even more advanced text processing techniques. An alternative would be to remove free text fields from maintenance data and implement better data validation tool to increase data quality.

There is also a gap between TSM and FMEA documents. For example, we found 7 TSM chapters which could not link to FMEA because the contributing components to the "Inability to Start" effect were completely different. We currently have no explanation for such a gap. Reconciliation would require the participation of the OEM and the end users. As we mentioned in the previous paper (Yang et al, 2009), there is also discrepancies between FMEA document and the operational and maintenance data. All of these create more challenges when updating TSM and validating FMEA.

Other challenges exist in updating TSM that are not related to the data collected from the end users. For example, we have to deal with some business and legal issues. One is the possible effect of the result with respect to the business process within the OEM because updated FMEA and TSM may request the unforeseen changes in the design of the system or component that may enhance the reliability of the system. Also, trade secrets, intellectual property, and competitive advantages can make the OEM reluctant in disclosing its FMEA and design documentation. In turn, this makes it more difficult to validate FMEA and update TSM. Finally, the TSM is considered a legal document that operator must follow in the maintenance. Modifications to this document without OEM consent may have legal ramifications and this issue must be investigated before implementing this procedure into the maintenance organization.

We believe in that the validated FMEA, in particular, the ranking information in Table 2 provides the useful resource for improving fault identification/isolation for a given failure effect or mode. Usually, when a failure has occurred, we have to identify which component is the root cause or to isolate the fault to a specific contributing component in order to schedule a maintenance action. As we introduced, we can use the revised TSM to isolate the root contributing component. However, TSM-based fault isolation procedure is still complicated and time consuming. To further enhance the troubleshooting procedures, we are developing a data mining-based fault isolation technology for PHM systems, which applies the updated FMEA to rank models and uses the operation data prior to failures as input to identify the root contributing component for a given failure mode. Initial results from this work were presented in another paper (Yang, et al, 2010).

5. CONCLUSION

In this paper, we proposed to update TSM by using the updated FMEA which is validated using historical

operational and maintenance data. We conducted the TSM revision for the failure mode of the "Inability to Start" by using the corresponding ranking information from the validated FMEA. The preliminary results obtained suggest that the validated FMEA provides more reliable and accurate information for updating TSM documents. The revised TSM provides more accurate information and reliable procedure for isolating the root components given a failure mode or effect.

ACKNOWLEDGMENT

Many people at the National Research Council Canada have contributed to this work. Special thanks go to Xijia Wu for providing the electronic FMEA documents and to Jeff Bird and Ken McRae for their support and valuable insights.

NOMENCLATURE

<i>FR</i>	failure rate
<i>FMP</i>	failure mode probability
<i>MTBF</i>	mean time between failures
<i>NFc</i>	number of replacements of a Component
<i>RN</i>	total number of APU unit replaced
<i>RPN</i>	risk priority number
<i>SC</i>	severity class of a failure mode
<i>UT</i>	total APU usage time

REFERENCES

- ASENT FMEA Software. (2009). FMEA RPN, available at <http://www.fmea-fmeca.com/fmea-rpn.html>, 2009.
- C. Yang, S. Létourneau, E. Scarlett, and M. Zaluski. (2009). APU FMEA Validation using Operation and Maintenance Data, in *Proceedings of the Annual Conference of the Prognostics and Health Management Society*. San Diego, CA, USA October 2009
- M. Zaluski, N. Japkowicz, and S. Matwin. (2003). Case Authoring from Text and Historical Experiences, in *Proceedings of the Sixteenth Canadian Conference on Artificial Intelligence (AI2003)*, June 11-13, 2003
- N. Sellappan and R. Sivasubramanian. (2008). Modified Method for Evaluation of Risk Priority Number in Design FMEA, *The Icfai Journal of Operations Management*, Vol. 7, No. 1, pp. 43-52, February 2008
- C. Yang, S. Létourneau, M. Zaluski, and E. Scarlett (2010), FMEA Validation and Its Application to Fault Identification, In *Proceedings of the ASME 2010 International Design Engineering Technical Conference & Computer and Information in Engineering Conference*, August 15-18, 2010, Montreal, Quebec, Canada

Dr. Chunsheng Yang is a Senior Research Officer at the Institute for Information Technology of the National Research Council of Canada. Chunsheng is interested in data mining, machine learning, reasoning technologies such as case-based reasoning, rule-based reasoning and hybrid reasoning, multi-agent systems, and distributed computing. Chunsheng received a B.Sc. (Hon.) in Electronic Engineering from Harbin Engineering University, China, an M.Sc. in computer science from Shanghai Jiao Tong University, China, and a Ph.D. from National Hiroshima University, Japan. Chunsheng worked with Fujitsu Inc., Japan, as a Senior Engineer and engaged on the development of ATM Network Management Systems. Chunsheng has been the author for over 50 papers and book chapters published in the referred journals and conference proceedings. Chunsheng was a Program Co-Chair for the 17th International Conference on Industry and Engineering Applications of Artificial Intelligence and Expert Systems. Chunsheng was a guest editor for the International Journal of Applied Intelligence. Chunsheng, as a senior IEEE member, has served Program Committees for many conferences and institutions, and has been a reviewer for many conferences, journals, and organizations, including Applied Intelligence, NSERC, IEEE Trans., ACM KDD, PAKDD, AAMAS, IEA/AIE, etc.

Dr. Sylvain Létourneau a Senior Research Council Officer with the National Research Council of Canada. Sylvain received his BS in computer science and mathematics and his MS in multi-agent systems from Université Laval in Québec City. He obtained his Ph.D. in machine learning from University of Ottawa. Sylvain is now leading the data mining group on Equipment Health Management at the Institute for Information Technology, National Research Council Canada.

Marvin Zaluski is with the Institute for Information Technology at the National Research Council Canada (NRC-IIT). Marvin received his B.Sc. (Hon.) with Honors from the University of Regina in Saskatchewan. Since joining the NRC-IIT in 1996, Marvin has received his Masters in Computer Science at the University of Ottawa in 2003. Marvin is involved in the development of knowledge discovery and knowledge management tools. His current work focuses on the development of data mining-based models for assisting maintenance personnel and managers in predicting faults on board jet aircraft. Marvin has experience working with companies in aerospace and open pit surface mining.

Appendix A: A chapter in TSM for the “inability to Start”

TASK 49-00-00-810-821 **ON A/C 201-234, 251-285, 401-449,

APU AUTO SHUT DOWN - NO FLAME, Ignition System -, or Fuel Control Unit -, or ECB 59KD - Fault (GTCP36-300)

1. Possible Causes

- IGNITER PLUG
- IGNITION UNIT P10
- OIL PUMP
- FUEL CONTROL UNIT
- FLOW DIVIDER AND DRAIN VALVE ASSY
- PRIMARY FUEL NOZZLE AND MANIFOLD
- SECONDARY FUEL NOZZLE AND MANIFOLD
- ECB (59KD)

2. Job Set-up Information

A. Referenced Information

REFERENCE	DESIGNATION
AMM 28-22-00-710-001	Operational Test of the APU Fuel-Pump System on Ground to Purge the Fuel Line
AMM 49-00-00-710-004	Operational Test of the APU (4005KM) (GTCP 36-300)
AMM 49-31-41-000-001	Removal of the Primary Fuel Nozzle and Manifold (8020KM) (GTCP 36-300)
AMM 49-31-41-400-001	Installation of the Primary Fuel Nozzle and Manifold (8020KM) (GTCP 36-300)
AMM 49-32-11-000-001	Removal of the Fuel Control Unit (FCU) (8022KM) (GTCP 36-300)
AMM 49-32-11-400-001	Installation of the Fuel Control Unit (FCU) (8022KM) (GTCP 36-300)
AMM 49-32-12-000-001	Removal of the Flow Divider and Drain Valve Assembly (8023KM) (GTCP 36-300)
AMM 49-32-12-400-001	Installation of the Flow Divider and Drain Valve Assembly (8023KM) (GTCP 36-300)
AMM 49-41-38-000-001	Removal of the Ignition Unit (8030KM) (GTCP 36-300)
AMM 49-41-38-400-001	Installation of the Ignition Unit (8030KM) (GTCP 36-300)
AMM 49-41-41-000-001	Removal of the Igniter Plug (8031KM) (GTCP 36-300)
AMM 49-41-41-400-001	Installation of the Igniter Plug (8031KM) (GTCP 36-300)
AMM 49-41-43-000-001	Removal of the Electrical Lead - Igniter Plug (GTCP 36-300)
AMM 49-41-43-400-001	Installation of the Electrical Lead - Igniter Plug (GTCP 36-300)
AMM 49-61-34-000-001	Removal of the Electronic Control Box (ECB) (59KD) (GTCP 36-300)
AMM 49-61-34-400-001	Installation of the Electronic Control Box (ECB) (59KD) (GTCP 36-300)
AMM 49-91-45-000-001	Removal of the Oil Pump (8080KM) (GTCP 36-300)
AMM 49-91-45-400-001	Installation of the Oil Pump (8080KM) (GTCP 36-300)

3. Fault Confirmation

A. Purging of the APU Fuel Feed-Line and Test

(1) Purge the APU fuel-feed line AMM TASK 28-22-00-710-001.

NOTE : If the fuel supply to the APU is not correct, do the applicable troubleshooting procedure(s) in the Chapter 28.

(2) Do the operational test of the APU AMM TASK 49-00-00-710-004.

4. Fault Isolation

A. If an APU auto shutdown occurs during the APU start sequence and the APU SHUTDOWNS report gives the maintenance message:

NO FLAME - CHECK IGNITION SYSTEM OR FCU OR ECB 59KD:

-do a check at the APU compartment drain-mast for fuel drain.

- (1) If there is no fuel drain:
 - go to step (5).
- (2) If there is fuel drain:
 - replace the IGNITER PLUG
 - AMM TASK 49-41-41-000-001 and AMM TASK 49-41-41-400-001.
- (3) If the fault continues:
 - replace the IGNITER PLUG ELECTRICAL-LEAD
 - AMM TASK 49-41-43-000-001 and AMM TASK 49-41-43-400-001 .
- (4) If the fault continues:
 - replace the IGNITION UNIT P10
 - AMM TASK 49-41-38-000-001 and AMM TASK 49-41-38-400-001.
- (5) If the fault continues:
 - remove the FUEL CONTROL UNIT P19
 - AMM TASK 49-32-11-000-001,

NOTE : TURN THE MANUAL DRIVE SHAFT OF THE STARTER MOTOR WITH A TORQUE WRENCH. THE TORQUE LIMIT IS 29 lbf.ft (3.9318 m.daN) . DO NOT TURN THE SHAFT WITH A TORQUE MORE THAN THE LIMIT. A TORQUE MORE THAN THE LIMIT WILL DAMAGE THE COMPONENT.

-to make sure that the oil pump input-shaft is not broken, turn the manual drive shaft of the starter motor(8KA) in a counterclockwise direction (the direction of the arrow on the housing) and make sure that the oil pump output-shaft (which drives the FCU) turns constantly.

- (a) If the oil pump output-shaft does not turn constantly (the oil pump input-shaft is broken):
 - replace the OIL PUMP
 - AMM TASK 49-91-45-000-001 and AMM TASK 49-91-45-400-001,
 - install a serviceable FUEL CONTROL UNIT P19
 - AMM TASK 49-32-11-400-001.
- 1 If the oil pump output-shaft turns constantly (the oil pump input-shaft is not broken):
 - install a new FUEL CONTROL UNIT
 - AMM TASK 49-32-11-400-001.
- (b) If the fault continues:
 - replace the FLOW DIVIDER AND DRAIN VALVE ASSY
 - AMM TASK 49-32-12-000-001 and AMM TASK 49-32-12-400-001.
- (c) If the fault continues:
 - replace the PRIMARY FUEL NOZZLE AND MANIFOLD
 - AMM TASK 49-31-41-000-001 and AMM TASK 49-31-41-400-001.
 - replace the SECONDARY FUEL NOZZLE AND MANIFOLD
 - AMM TASK 49-31-41-000-001 and AMM TASK 49-31-41-400-001 .
- (d) If the fault continues:
 - replace the ECB (59KD)
 - AMM TASK 49-61-34-000-001 and AMM TASK 49-61-34-400-001.

B. Do the operational test of the APU AMM TASK 49-00-00-710-004.

Revision:2004-11-01 Print Date: 2010-04-01 49-00-00

Appendix B: The revised chapter for the original document in Appendix A.

TASK 49-00-00-810-821 **ON A/C 201-234, 251-285, 401-449,

APU AUTO SHUT DOWN - NO FLAME, Ignition System -, or Fuel Control Unit -, or ECB 59KD - Fault (GTCP36-300)

1. Possible Causes

- **IGNITER PLUG**
- **IGNITION UNIT P10**
- **FUEL CONTROL UNIT**
- **OIL PUMP**
- **FLOW DIVIDER AND DRAIN VALVE ASSY**
- **PRIMARY FUEL NOZZLE AND MANIFOLD**
- **SECONDARY FUEL NOZZLE AND MANIFOLD**
- **ECB (59KD)**



This list is ordered with the rank information from the validated FMEA.

2. Job Set-up Information

A. Referenced Information

REFERENCE	DESIGNATION
AMM 28-22-00-710-001	Operational Test of the APU Fuel-Pump System on Ground to Purge the Fuel Line
AMM 49-00-00-710-004	Operational Test of the APU (4005KM) (GTCP 36-300)
AMM 49-31-41-000-001	Removal of the Primary Fuel Nozzle and Manifold (8020KM) (GTCP 36-300)
AMM 49-31-41-400-001	Installation of the Primary Fuel Nozzle and Manifold (8020KM) (GTCP 36-300)
AMM 49-32-11-000-001	Removal of the Fuel Control Unit (FCU) (8022KM) (GTCP 36-300)
AMM 49-32-11-400-001	Installation of the Fuel Control Unit (FCU) (8022KM) (GTCP 36-300)
AMM 49-32-12-000-001	Removal of the Flow Divider and Drain Valve Assembly (8023KM) (GTCP 36-300)
AMM 49-32-12-400-001	Installation of the Flow Divider and Drain Valve Assembly (8023KM) (GTCP 36-300)
AMM 49-41-38-000-001	Removal of the Ignition Unit (8030KM) (GTCP 36-300)
AMM 49-41-38-400-001	Installation of the Ignition Unit (8030KM) (GTCP 36-300)
AMM 49-41-41-000-001	Removal of the Igniter Plug (8031KM) (GTCP 36-300)
AMM 49-41-41-400-001	Installation of the Igniter Plug (8031KM) (GTCP 36-300)
AMM 49-41-43-000-001	Removal of the Electrical Lead - Igniter Plug (GTCP 36-300)
AMM 49-41-43-400-001	Installation of the Electrical Lead - Igniter Plug (GTCP 36-300)
AMM 49-61-34-000-001	Removal of the Electronic Control Box (ECB) (59KD) (GTCP 36-300)
AMM 49-61-34-400-001	Installation of the Electronic Control Box (ECB) (59KD) (GTCP 36-300)
AMM 49-91-45-000-001	Removal of the Oil Pump (8080KM) (GTCP 36-300)
AMM 49-91-45-400-001	Installation of the Oil Pump (8080KM) (GTCP 36-300)

3. Fault Confirmation

A. Purging of the APU Fuel Feed-Line and Test

(1) Purge the APU fuel-feed line AMM TASK 28-22-00-710-001.

NOTE : If the fuel supply to the APU is not correct, do the applicable troubleshooting procedure(s) in the Chapter 28.

(2) Do the operational test of the APU AMM TASK 49-00-00-710-004.

4. Fault Isolation

A. If an APU auto shutdown occurs during the APU start sequence and the APU SHUTDOWN report gives the maintenance message:

NO FLAME - CHECK IGNITION SYSTEM OR FCU OR ECB 59KD:

-do a check at the APU compartment drain-mast for fuel drain.

- (1) If there is no fuel drain:
 - go to step (5).
- (2) If there is fuel drain:
 - replace the IGNITER PLUG
 - AMM TASK 49-41-41-000-001 and AMM TASK 49-41-41-400-001.
- (3) If the fault continues:
 - replace the IGNITER PLUG ELECTRICAL-LEAD
 - AMM TASK 49-41-43-000-001 and AMM TASK 49-41-43-400-001 .
- (4) If the fault continues:
 - replace the IGNITION UNIT P10
 - AMM TASK 49-41-38-000-001 and AMM TASK 49-41-38-400-001.
- (5) If the fault continues:
 - remove the FUEL CONTROL UNIT P19
 - AMM TASK 49-32-11-000-001,

NOTE : TURN THE MANUAL DRIVE SHAFT OF THE STARTER MOTOR WITH A TORQUE WRENCH. THE TORQUE LIMIT IS 29 lbf.ft (3.9318 m.daN) . DO NOT TURN THE SHAFT WITH A TORQUE MORE THAN THE LIMIT. A TORQUE MORE THAN THE LIMIT WILL DAMAGE THE COMPONENT.

-to make sure that the oil pump input-shaft is not broken, turn the manual drive shaft of the starter motor(8KA) in a counterclockwise direction (the direction of the arrow on the housing) and make sure that the oil pump output-shaft (which drives the FCU) turns constantly.

(a) If the fault continues:

- install a new **FUEL CONTROL UNIT**
- AMM TASK 49-32-11-400-001.

(b) If the fault continues and the oil pump output-shaft does not turn constantly (the oil pump input-shaft is broken):

- replace the **OIL PUMP**
- AMM TASK 49-91-45-000-001 and AMM TASK 49-91-45-400-001,

-If the oil pump output-shaft turns constantly (the oil pump input-shaft is not broken):

- replace the **FLOW DIVIDER AND DRAIN VALVE ASSY**
- AMM TASK 49-32-12-000-001 and AMM TASK 49-32-12-400-001.

This FIP is updated with the rank information from the validated FMEA.

(c) If the fault continues:

- replace the **PRIMARY FUEL NOZZLE AND MANIFOLD**
- AMM TASK 49-31-41-000-001 and AMM TASK 49-31-41-400-001.
- replace the **SECONDARY FUEL NOZZLE AND MANIFOLD**
- AMM TASK 49-31-41-000-001 and AMM TASK 49-31-41-400-001 .

(d) If the fault continues:

- replace the **ECB (59KD)**
- AMM TASK 49-61-34-000-001 and AMM TASK 49-61-34-400-001.

B. Do the operational test of the APU AMM TASK 49-00-00-710-004.

Revision:2011-04-01 Print Date: 2011-04-06 49-00-00

Utilizing Dynamic Fuel Pressure Sensor For Detecting Bearing Spalling and Gear Pump Failure Modes in Cummins Pressure Time (PT) Pumps

J. Scott Pflumm, Jeffrey C. Banks

Applied Research Laboratory, State College, PA, 16801, USA

jsp116@arl.psu.edu
jcb242@arl.psu.edu

ABSTRACT

The objective of this paper is to highlight the results of the fault detection investigation conducted to ascertain the feasibility of exploiting the existing on-board M2/M3 Bradley fuel pressure sensor for the purpose of detecting mechanical bearing spalling and gear pump failure modes of the pressure-time (PT) fuel pump used on the Cummins VTA-903T engine. To investigate this fluid-mechanical cross domain detection approach, a Bradley fuel system test bed was built. Fault tests for four PT pump failure modes were conducted including bearing faults, gear pump fault, idle adjust mis-calibration, and air-fuel control fault. The results of the first two fault tests are summarized in this paper. Due to limited number of pumps available for testing (2), these preliminary findings are not statistically substantiated. With this stated, the findings present a method for investigating the presence of a narrowband frequency-domain-based predictive fault detection capability using the existing pressure sensor installed on the Chassis Modernization and Embedded Diagnostics (CMED) variant Bradley. The test stand based seeded fault analysis was not capable of detecting a 0.080 inch outer raceway bearing spall, but there is preliminary evidence to warrant further study that a nominal 0.001 inch foreign object debris accumulation on the gear teeth of the gear pump might be detectable using a simple kurtosis based calculation using a pressure sensor with a 0-500 Hz dynamic bandwidth.

1. INTRODUCTION

The Program Management Office for the Heavy Brigade Combat Team (PM-HBCT) is leading the development of a Vehicle Health Management System (VHMS) that provides the US Army with an improved diagnostic, predictive and sustainment capability for HBCT platforms including the

M2/M3 Bradley Fighting Vehicle. The common challenge faced by the Program Management Office when evaluating VHMS technology is the requirement to install additional instrumentation or to modify the chassis of an existing fleet of vehicles. These requirements often detrimentally impact the cost-benefit decision to implement VHMS technology. This paper discusses utilizing a dynamic fuel pressure sensor existing on the CMED configuration of the M2/M3 Bradley in order to detect mechanical faults in the PT pump. The objective is to employ existing on-board sensors to extend the vehicle's present VHMS capability. Prior to this investigation a vehicle degrader analysis was conducted in order to ascertain where health monitoring technology would provide the greatest benefit in terms of decreasing diagnosis time, increasing maintenance effectiveness, decreasing No Evidence of Failure (NEOF) rates and facilitating the migration to a 2-tier maintenance system.

2. DEGRADER ANALYSIS

A Reliability Centered Maintenance (RCM) degrader analysis for the M2/M3 Bradley was conducted by the HBCT VHMS program to assess the top degraders of the vehicle's maintainability, reliability and operational availability (Banks, Reichard, Hines, Brought, 2008). The formal RCM process typically consists of regular meetings with subject matter experts (i.e. maintainers, design engineers, logisticians, CBM experts, etc.) to evaluate a system over a significant period of time (i.e. weeks to months). The length of time required for the analysis is dependent upon the complexity of the system and the knowledge level of the subject matter experts participating in the process. It was logistically prohibitive to regularly gather the subject matter experts required to conduct a formal RCM analysis for the U.S. Army's Bradley vehicle. Penn State ARL conducted a modified RCM analysis using the results of the degrader interviews, degrader OEM questionnaire, field service representative reports, technical manuals and engineering judgment. This process provides a systematic approach for the evaluation of the VHMS design,

J.S. Pflumm et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

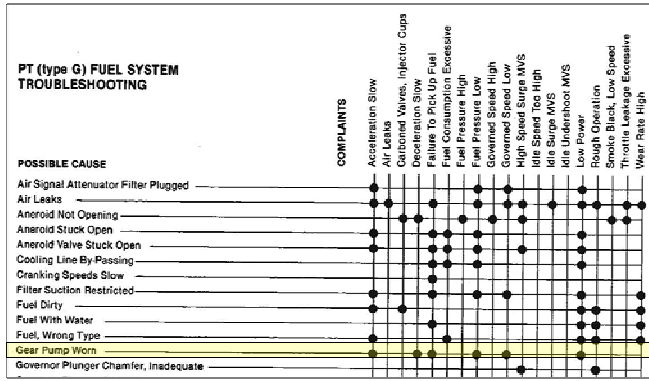


Figure 1. Troubleshooting cause-effect matrix excerpt from Cummins PT Fuel Pump Rebuilding and Calibration Instructions

functional requirements, failure modes and criticality assessment. From this assessment the appropriate and most effective maintenance approaches can be selected for the platform or system based on the operating context of the asset. The specific advantage of utilizing the RCM process when designing an asset health management system is the ability to determine where the implementation of embedded diagnostic technology would provide the greatest benefit in terms of facilitating improved maintainability and increased operational availability of the asset.

The degrader analysis reviewed Field Service Reports for the following 7 Bradley subsystems: Track and Suspension, Gun System, Electrical System, Turret Drive Control Unit, Transmission, Cable and Wiring, and Fuel System. A total of 769 failures were tallied across all subsystems. Four percent (4%) of all 769 failures involved the fuel system. Two percent (2%) of the 769 failures cited the PT pump as the failure mode. Approximately 50% of the 31 cited fuel system related failures involved the PT pump (Banks, Reichard, Hines Brought, 2008). Given the existence of a dynamic fuel pressure sensor located between the PT pump and fuel injectors, and the results of the degrader analysis, the Bradley fuel system was considered a candidate for investigating implementation of VHMS technology. The on-board dynamic fuel pressure sensor is not necessarily capable of detecting all the possible PT pump faults listed in figure 1. The underlying intent of utilizing the existing pressure sensor to support real-time on-board diagnostic capability is to investigate the greatest extent to which the sensor can be exploited to aid the operator/maintainer in terms of alerting the crew of impending fuel system faults, or automatically diagnosing a subcomponent fault that would not be apparent during routine preventive maintenance checks and services.

3. VEHICLE FUEL SYSTEM DESCRIPTION

The Bradley fuel system supports a Cummins VTA-903T water-cooled, eight cylinder diesel engine that rates 600 HP. The fuel system consists of two fuel tanks (upper and

lower), four in-tank fuel pumps, a fuel/water separator, and a PT fuel pump/governor with integrated air-fuel control (AFC) valve.

In general, the fuel flows from the lower tank through two check valves and a main shut-off valve, through the filter/water separator to the PT pump/governor, which sends regulated high pressure fuel to the injectors as shown in figure 2 (Technical Manual, 2006). Control of engine power output and idle speed is accomplished by the engine mounted PT fuel pump/governor with integrated AFC valve. The fuel pressure to the injectors is regulated to between 129 psi - 163 psi depending on the pump version used with the electronic fuel control valve (EFCV). Excess fuel is returned to the tank through a low pressure return line.

The CMED equipped vehicles have a pressure transducer integrated into the fuel system. This sensor has a 0-500 Hz measurement capability and is located at the existing Bradley Standard Test Equipment – Internal Combustion Engine diagnostic test kit (BRADS-ICE) pressure sensor port. The CMED system monitors this pressure transducer during diagnostic/maintenance mode (the vehicle is not operational in this mode) and provides an indication of pump failure based on an insufficient pressure level while the engine is run at a high RPM condition. The limitation is that the AFC valve needs to be manually opened to conduct this procedure. This test can only be conducted when the platform is down for maintenance inspection.

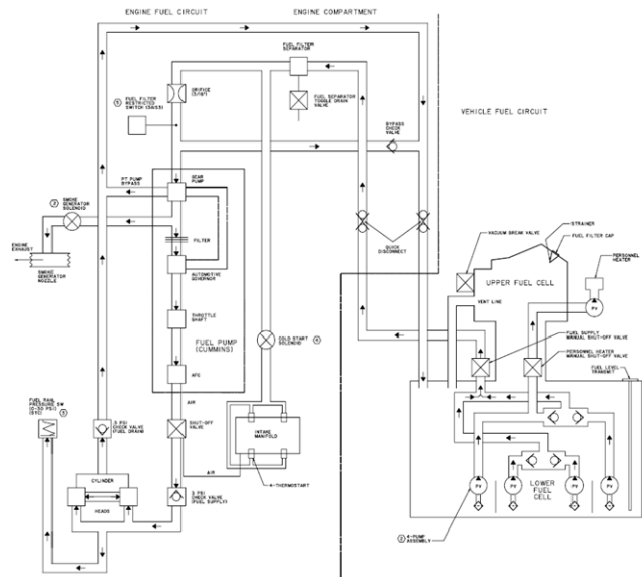


Figure 2. Bradley Fuel System Schematic

(TM 9-2350-294-20-1-1/3, 2000)

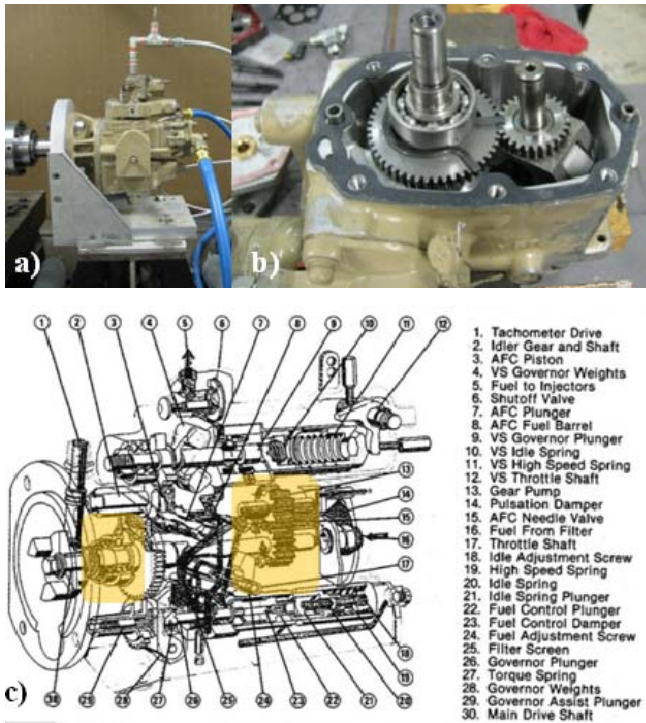


Figure 3. a) PT pump assembled; b) Partially disassembled; c) Schematic

Figure 3 shows three images depicting the PT pump. The bottom image is a schematic of a similar model PT pump, however it shows the relevant components most clearly. The yellow boxes highlight the subcomponents investigated in this analysis. The left-most yellow box highlights the front bearing. The right box highlights the gear pump gears.

4. TEST BED, SENSORS AND DATA ACQUISITION

The test bed consists of one fuel tank (representative of the lower tank), four in-tank fuel pumps, a fuel filter separator, and a PT fuel pump/governor with integrated air-fuel control (AFC) valve. The in-tank pumps, fuel filter separator, PT pump and commercial equivalent of the pressure transducer are configured in the same sequence as they are on the vehicle. On the test bed, fuel enters the in-tank fuel pump and continues through the fuel filter separator to the PT pump which is directly driven by a 30 HP AC electric motor. For simplicity in this phase of the project, the PT pump output, which is intended to supply fuel to the fuel injectors, is routed to a manual adjusted needle valve. The needle valve outlet pressure was set to nominal 160 PSIG and the air supply to the AFC check valve was set to 30 PSIG in accordance with Cummins Calibration Instruction Manuals (Cummins Bulletin Nrs. 3379352-10 and 3379084-02, 1980 and TM 9-2350-294-20-1-3). Validation data is maintained and available for review. The PT pump leakage and fuel supplied to the needle valve are returned directly into the sump tank to form a continuous closed loop non-combusting fuel circuit as

shown in figure 4. For fault induction and component isolation purposes, the valves, flow meters, pressure transducers and thermocouples are located at the following locations:

- After the in-tank fuel pump and before the fuel filter separator
- After the fuel filter separator and before the PT pump
- After the PT pump and before the needle valve

A 16 channel National Instruments PXI based data acquisition system with 100 kSamples/second per channel capability is used to support data gathering. This data acquisition system collects user-triggered 10 second snapshots of the voltage and current sensors for monitoring the in-tank pump power, ICP tri-axial accelerometer mounted on the PT pump, fluid pressure and flow at each of the aforementioned three locations in the fluid circuit, as well as a torque cell measuring both torque and speed of the drive shaft to the PT pump.

The analysis emphasizes the findings of the pressure sensor in terms of its capability as an embedded diagnostic tool. The commercial version of this pressure sensor, which we used, provides higher bandwidth and is therefore capable of higher resolution data processing techniques that are otherwise not available with the existing on-board 0-500 Hz pressure sensor. Our goal is to limit our analysis to what is implementable given the operational bandwidth of this 0-500 Hz on-board pressure transducer.

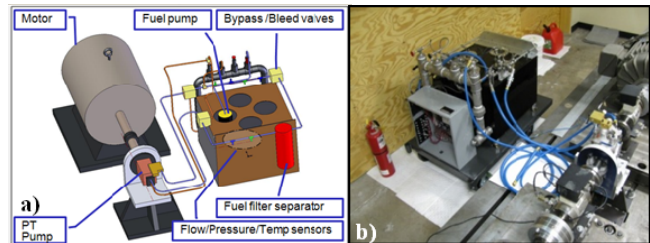


Figure 4. a) Drawing Bradley Fuel System Test Bed b) Photo

5. TESTING OVERVIEW

The test approach can be summarized briefly as follows: first investigate failure modes that do not result in permanent damage to the pump, then progress deeper into failure modes that permanently damage the pump or its subcomponents. The rationale for this approach is due to having only 2 pumps available on which to conduct testing. Initial testing of both pumps focused on baseline characterization of steady state and transient conditions across specific run speed setpoints conditions.

With respect to the second pump, testing focused primarily on permanently damaging subcomponents, inserting these various seeded fault subcomponents into the pump in place

of their no-fault component counterpart, and then conducting testing to determine if the fault could be detected by means of the pressure sensor. These tests focused first on seeding faults in the front bearing and then secondly by seeding a fault in the gear pump.

The ideal test plan would utilize a large sample size of actual pumps degraded over the course of real world operational service. We did not have access to such a sample set of pumps. Furthermore, the degrader analysis only identified the PT pump as a single monolithic component in terms of its responsibility as a contributor to overall fuel system failures. Due to data access limitations, the degrader analysis did not provide further details as to which subcomponent(s) within the PT pump were responsible for the pump's overall failure. Absent this information, the next best alternative was to simulate faulted pumps by seeding faults in the pump's subcomponents which anecdotally or hypothetically explained the pump's reported failure during fielded operation. The seeded faults were chosen to represent faults that were either reported by maintainers or potentially problematic given operating conditions. With this said, given the pumps are intended for use with JP8 fuel, and as such, the fuel system is required to be capable of operation with fuel containing impurities and particulate such as fine sand/dirt common in desert environments, two failure modes were chosen for test stand investigation: (1) bearing spalling and (2) accumulation of a particulate coating on the teeth of the gears within the gear pump sub-assembly. The following subsections will document the testing and analysis of: (1) simulated bearing spalling fault and, (2) simulated gear tooth particulate accumulation.

As indicated earlier in this report, the primary objective was to utilize the capability of the existing pressure sensor that is integrated into CMED variant Bradley platforms with the least complex and computationally intensive condition indicators. The rationale with respect to the evaluation progression was to start with the basic root mean square condition indicator and progress to more advanced statistical features if RMS is not effective for these faults and this application.

5.1 Bearing Fault Test Discussion

To summarize the motivation for the spalling fault tests, the rationale for investigating a simulated spall in the bearing raceway is based on the hypothesis that fine sand particles, which have infiltrated the fuel system, make their way into the bearing raceway. The repetitive impact of the roller elements passing over the particle(s) as they lay in the raceway results in pitting/spalling which increases in size over time. The subsequent imperfections in the surface of the raceway might potentially degrade the bearings performance and thereby the PT pumps performance. The purpose of the study is to investigate whether such a fault

can be detected by the on-board pressure sensor located post-PT pump on the M2 CMED Bradley vehicle. This test initially set out to specifically investigate whether the bearing fault frequency known as the Ball Pass Frequency – Outer Race (BPFO) could be detected in the pressure spectrum. For completeness, the other three fault frequencies are referenced as follows: Ball Pass Frequency – Inner Race (BPFI), Ball Spin Frequency (BSF), Fundamental Train Frequency (FTF).

The analytical equations for these bearing fault frequencies displayed in subsequent plots are as follows according to (White, 1995):

$$BPFO = \frac{n}{2} \left(1 - \left(\frac{B_d}{P_d} \right) \cos \theta \right) RPM \quad (1)$$

$$BPFI = \frac{n}{2} \left(1 + \left(\frac{B_d}{P_d} \right) \cos \theta \right) RPM \quad (2)$$

$$BSF = \frac{P_d}{2B_d} \left(1 - \left(\frac{B_d}{P_d} \right)^2 (\cos \theta)^2 \right) RPM \quad (3)$$

$$FTF = \frac{1}{2} \left(1 - \left(\frac{B_d}{P_d} \right) \cos \theta \right) RPM \quad (4)$$

The front bearing of the PT pump is a type NSK 6203. According to the manufacturer, the following specifications are:

Number of rolling elements (n):	8
Ball diameter (B_d):	6.746 mm
Pitch diameter of the bearing (P_d):	29 mm
Contact angle (θ):	0°

While the figures below include the fault frequencies in the subplot titles, we did not observe spectral peaks at these frequencies when analyzing the datasets.

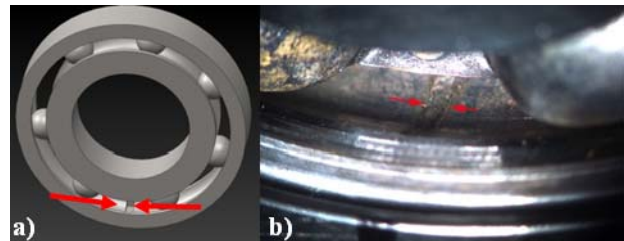


Figure 5. a) Electrolytically etched bearing outer raceway schematic format, b) Actual 0.030 inch etch

Electrolytic etching was selected instead of electric discharge machining (EDM) as the method of choice to create faults in the bearing raceway for this study because of the non-uniform spall edge it produced. This ad hoc technique was used specifically because of its ability to create non-uniform edges in the bearing raceway. No literature reference source was used as a precedent for this specific etching application. Electrolytic etching is essentially a controlled electrolytic corrosion process described in standard chemistry texts. Electrolytic corrosion is defined as a process of accelerated corrosion resulting

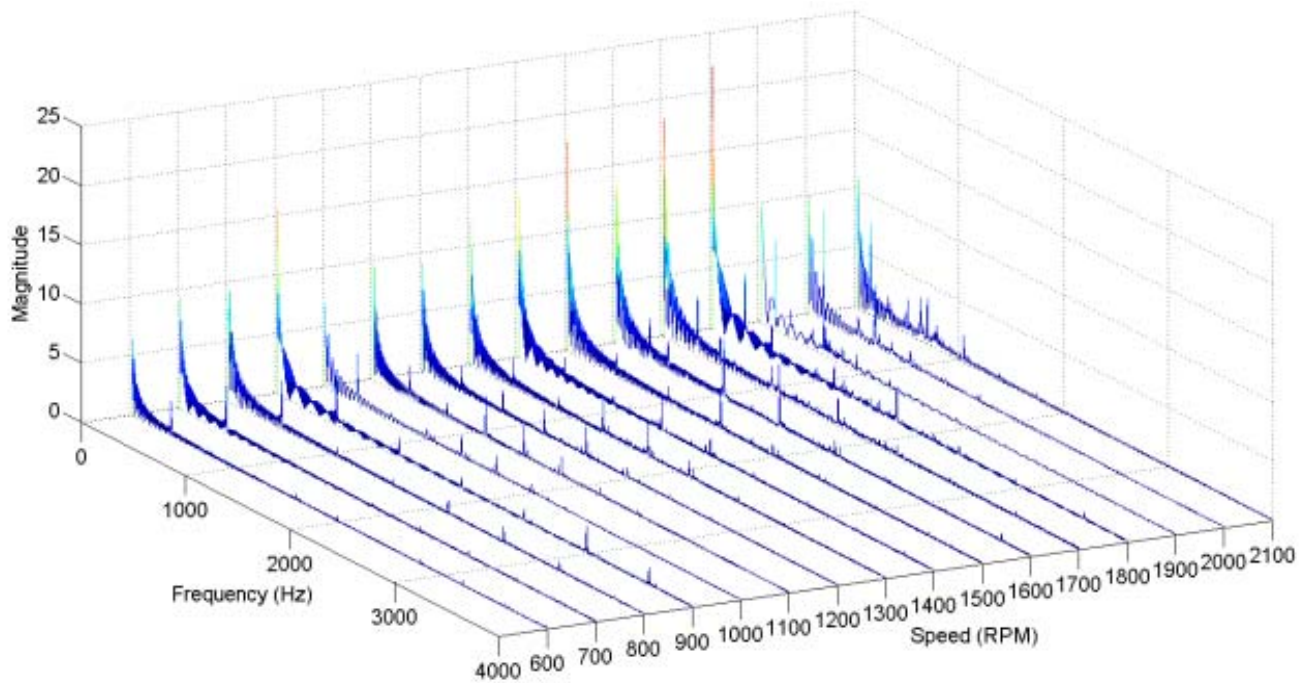


Figure 6. Representative waterfall plot depicting selected frequency domain spectrum of pressure transducer over the 600-2100 RPM speed setpoint conditions for a no-fault bearing test

from an electric current between two metals coupled in an electrolytic environment. (DTSFE, 2011). In this case the electrolytic environment was a saline solution.

The simulated spall testing, hereafter referred to as etched bearing or etched fault testing, was conducted with two bearings, each with a nominal etch width of 0.030 inch and 0.080 inch, respectively. Figure 5 depicts the faulted bearing in schematic form as well as a photo of the actual etched 0.030 inch faulted bearing. The chosen etch widths of 0.030 and 0.080 inch were dictated primarily by the precision limits of the equipment used to create the etch. A spall width of 0.080 inch is likely an over-exaggeration compared to a spall that may occur in a field environment. For initial investigation purposes our intent was to validate whether a bearing defect could be detected using the existing pressure transducer on the CMED platform. The test plan used to investigate whether the etched fault could be detected via the pressure transducers utilized 16 speed setpoint conditions ranging from 600-2100 RPM at wide open throttle along with four zero (0) RPM setpoint conditions used to characterize background ambient noise conditions when the pump main shaft was not rotating. Inspection of the background ambient noise data indicated potential electromagnetic interference due to the adjacent AC motor controller. While the methodology is valid, the statistical limitations due to pump sample size combined with the potential EMI limit the extent to which significance can be ascribed to the spectrum analysis discussed in the following sections.

Figure 6 shows the spectra for 16 run speed setpoint conditions. Similar waterfall plots showing all 16 speed setpoints were generated for the 0.030 inch and 0.080 inch etched bearing tests for both pressure transducer and accelerometer signals.

The next step of the analysis was to determine whether the pressure spectra contained visual/qualitative features indicating the presence of the etch in the bearing. To this end the initial approach taken was to overlay the spectra of the no-fault, 0.030 inch width and 0.080 inch width etched fault test at each speed setpoint and view the spectra across the 0-500 Hz operational bandwidth of the dynamic pressure transducer. Figure 7 illustrates a representative result of this overlay for one run speed setpoint condition at 1500 RPM. While there are distinguishable differences between the spectra, there was no specific distinguishable and repeatable feature observed across all run speed setpoints.

Recognizing the limits of visual inspection with respect to spectrum plots of this nature, an alternate approach was undertaken by which a root means square (RMS) calculation was computed for each no-fault, 0.030 inch and 0.080 inch fault spectrum across the entire spectrum range, respectively. Figure 8 depicts the results of this approach for four selected run speeds along with the RMS values for each fault condition displayed in the legend.

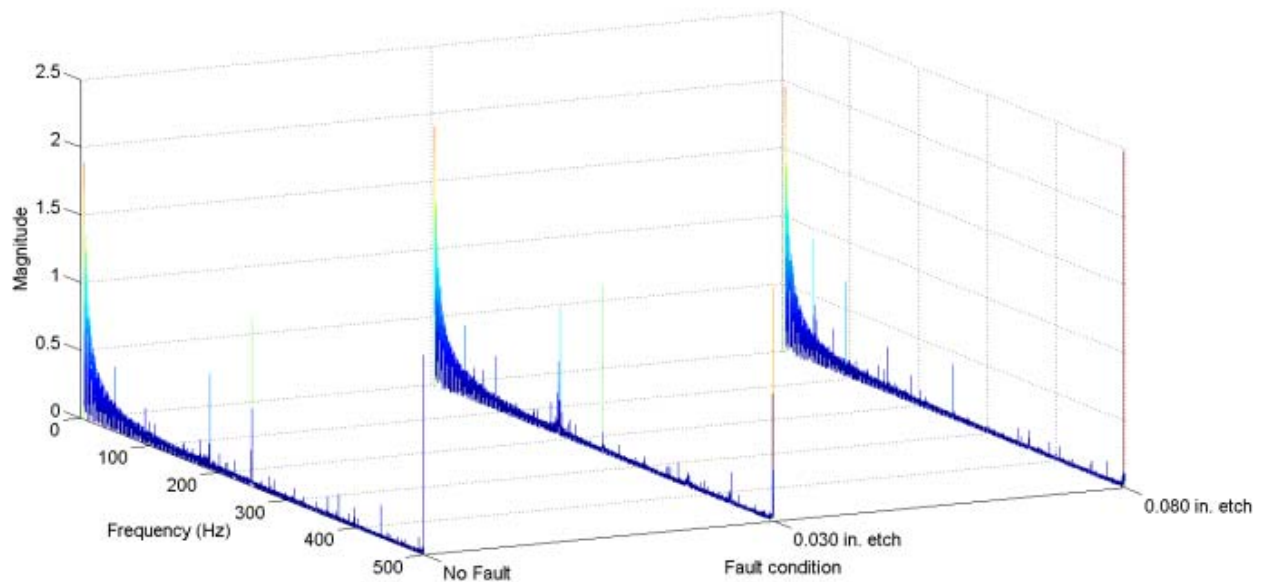


Figure 7. Waterfall comparison overlay of pressure spectrums for no-fault, 0.030 inch width and 0.080 inch width etched fault at nominal run speed 1500 RPM

It is observed that at three of these specific run speeds (600, 1000, and 1500 RPM), a simple RMS calculation produced condition indicator (CI) values that increased as etch width increased from no-fault through 0.030 inch to 0.080 inch width. Notice that the RMS values for the bottom plot (nominal run speed = 2000 RPM) did not support this trending correlation observed in the top three subplots. The bar graph in figure displays these RMS values in a graphic representation.

Note that at the 2000 RPM speed setpoint, the no-fault RMS value is greater than the 0.030 inch etch RMS value. This deviation from the trend was not an isolated outlier. The general RMS trending relationship observed between 600-2000 RPM across fault sizes was not repeatable for all speed setpoint conditions. In general, the RMS feature did not provide a consistent and highly sensitive condition indicator for bearing fault predictive detection.

Because the RMS based approach did not yield consistent results in terms of correlating RMS value to etch width for all speeds, we surmised that the potential reason for not detecting this correlation was perhaps due, at least in part, to the fact that utilizing the entire operational bandwidth of the pressure sensor may have been effectively masking RMS changes occurring within specific frequency ranges.

To address this point, we returned to the radial vibration data and observed the accelerometer's frequency spectrum at each run speed setpoint condition. The objective was to identify spectra peaks that; (1) remained constant across all run speed setpoint conditions, or (2) that predictably changed with respect to run speed, or (3) that somehow appeared to correlate with the width of the etched bearing fault. We

would then use these peaks as a reference, or marker frequency about which we would isolate our RMS calculation of the pressure spectrum for the no-fault, 0.030 inch and 0.080 inch width etch fault tests. In general, using vibration spectrum peaks as a 'frequency marker' did not produce consistently repeatable RMS trend results.

This inconsistency in the RMS calculation even when using a vibration frequency marker as described above led us to investigate whether there were any frequencies, or frequency bands at which the RMS value of vibration spectrum increased with increasing etched fault width AND simultaneously, at these frequencies (or frequency bands), the RMS value of the pressure spectrum also increased with increasing etched fault width. Figure 10 shows a binary plot illustrating the results of this study for selected run speed setpoint conditions.

In long hand explanation, the algorithm plots either a 'TRUE' or 'FALSE' depending on whether the following condition is satisfied:

The plot is 'TRUE' or 'High', if:

- (1) The vibration data is consistent in terms of the RMS value of the 0.080 inch etch spectrum is greater than the 0.030 inch etch spectrum;
- (2) AND the RMS value for the 0.030 inch etch spectrum is in turn greater than the no fault spectrum;
- (3) AND simultaneously the pressure data is consistent in terms of the RMS value of the 0.080 inch etch spectrum is greater than the 0.030 inch etch spectrum;

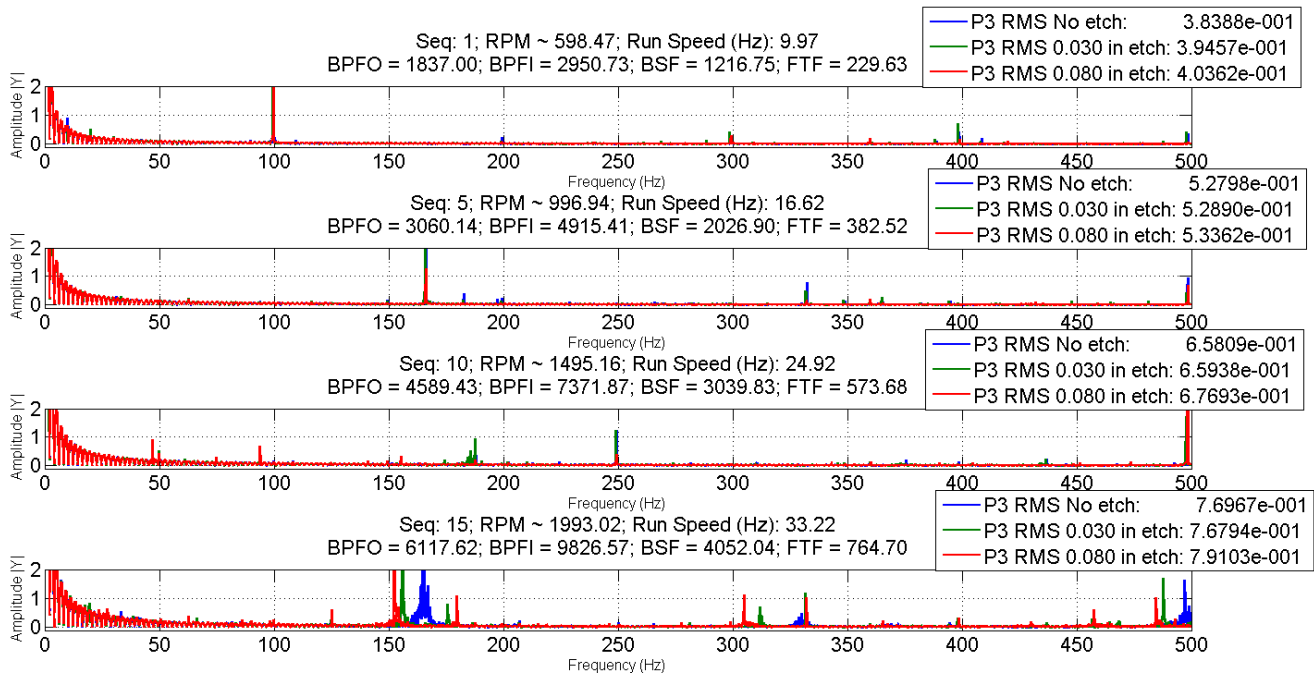


Figure 8. RMS calculations for no-fault, 0.030 inch width and 0.080 inch width etched fault spectrums for four selected run speed

- (4) AND the RMS value for the 0.030 inch etch spectrum is in turn greater than the no fault spectrum.

Though the resolution is difficult to visually resolve across the entire 0-500 Hz bandwidth, there are frequency regions which the ‘density’ of TRUE conditions are relatively greater than for other regions. Even when comparing across multiple run speed setpoint conditions, there appears to be frequency bands that have corresponding TRUE vs. FALSE regions. However there does not appear to be a frequency band that is consistent across all run speed setpoint conditions.

Had a consistent frequency band been identified across all speeds, this frequency range would have been the focal point for developing a fault detection algorithm, specific to this bearing spalling fault.

To investigate the feasibility of implementing the RMS fault detection approach using the existing broad band vehicle sensor data, figure 11 distills in bar graph format the 0-500 Hz RMS results for the respective pressure spectra at multiple runs speeds. It should be noted that the results are not consistently repeatable across all run speeds.

Based on our limited data set, there appears to be speed regimes in which the RMS value increases with fault severity, as highlighted in figure 11. In general the RMS calculation has a number of limitations. For example it is

common to observe significant variation in RMS values among pumps, engines and transmissions all of the same model type. It is therefore prudent that any condition monitoring algorithm incorporate multiple calculation approaches as a means of corroborating conclusions and minimizing false positive results.

There exist a number of techniques collectively referred to as statistical signal processing techniques or features that can also be applied. Four additional features were applied to the pressure signal data; Total Energy, Crest Factor,

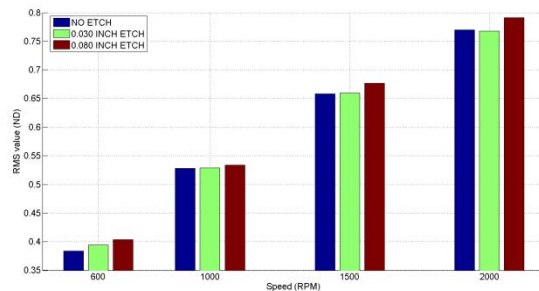


Figure 9. Bar graph plot of RMS calculations for no-fault, 0.030 inch width and 0.080 inch width etched fault spectrums at four selected run speeds

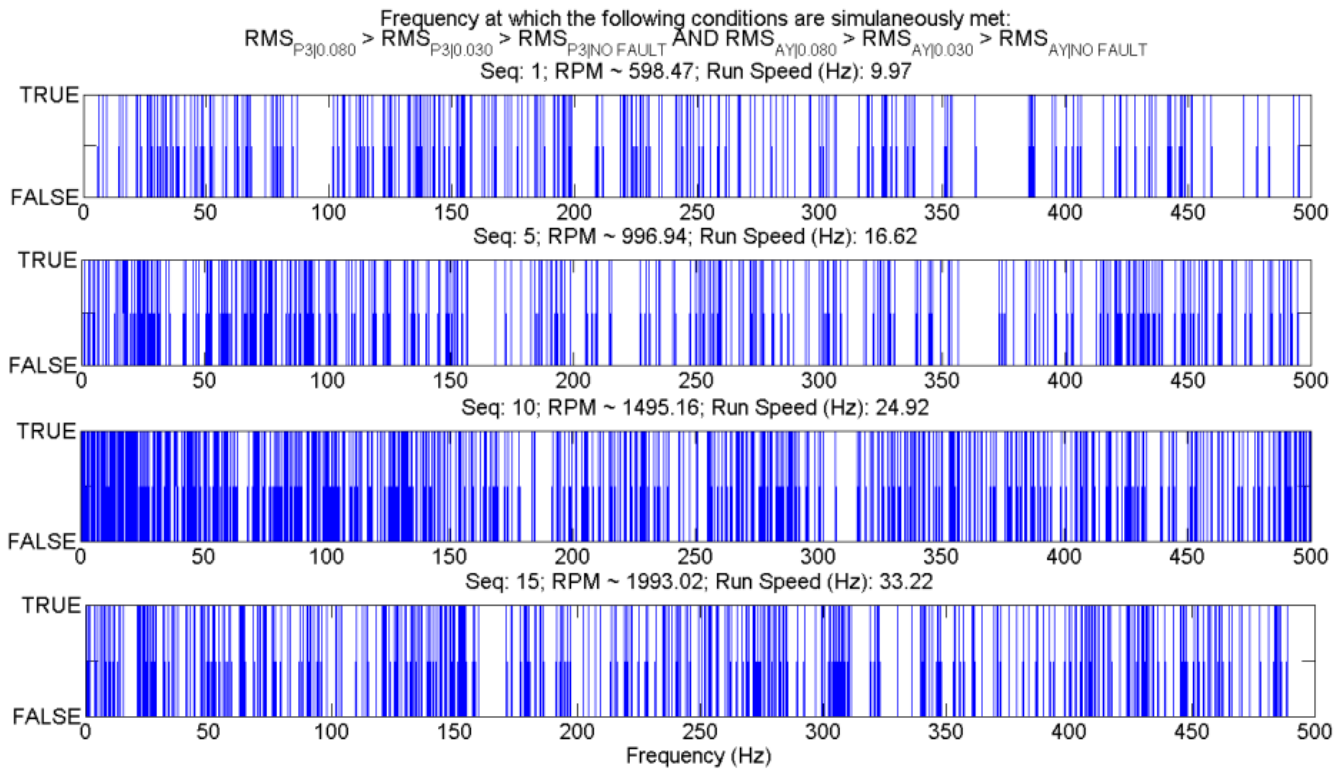


Figure 10. Identifying frequency bands for which the RMS value of the pressure spectrum increased with increasing etched fault width AND simultaneously at these frequencies the RMS value of the vibration spectrum also increased with increasing etched fault width

Kurtosis, and Kurtosis Interstitial Envelope. The results of the four methods also did not show consistent quantitative trends with respect to etch width. Representative of the inconsistency among the feature findings is figure 12. Figure 12 illustrates that the results of applying kurtosis did not provide the ability to predictively detect the various states of bearing condition consistently across every speed range, but it did provide a more sensitive predictive fault detection capability at one more of the speed setpoints relative to the RMS approach.

5.2 Etched Bearing Test Conclusion

The results of the of the bearing fault detection investigation indicated the RMS pressure spectrum from 0-500 Hz, as a condition indicator, has a low fault sensitivity and it is only effective for less than half of the speed setpoints that were tested. Specifically, the RMS calculation was not consistent across all 16 run speed setpoint conditions in the sense that the RMS values correlated with increasing etch width. Select individual speed setpoint conditions appeared to be consistent in terms of increasing RMS value vis-a-vis etch width but this was not consistent across all speeds. The kurtosis pressure spectrum from 0-500 Hz provided a predictive condition indicator with a slightly improved fault sensitivity as compared to the RMS feature. Though it is effective for more of the speed setpoints tested it is not effective for the entire speed range. In comparison, the RMS accelerometer

spectrum from 0-500 Hz provided a predictive condition indicator with the highest fault sensitivity as compared to the other features.

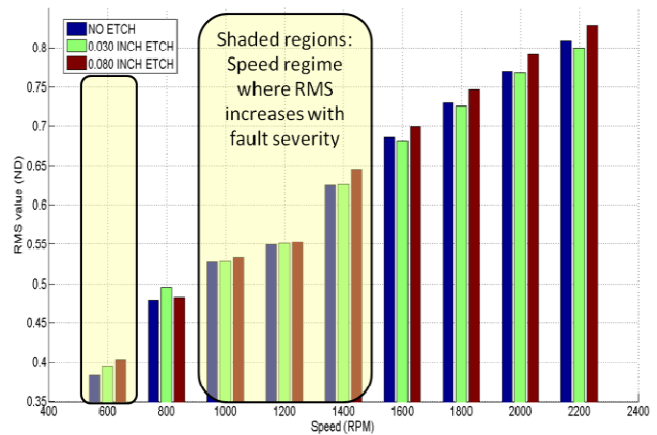


Figure 11. Bar graph summary of 0-500 Hz bandwidth RMS calculation for fuel pressure spectrum with bearing fault for nine run speed set point conditions

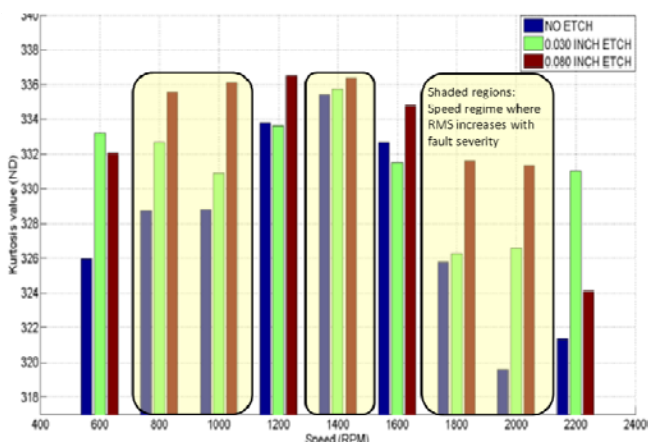


Figure 12. Bar graph summary of 0-500 Hz bandwidth Kurtosis calculation for fuel pressure spectrum with bearing fault for nine run speed set point conditions

5.3 Gear Fault Test Discussion

The aim of the gear fault testing was to investigate whether accumulation of particulate on the teeth of the gear pump sub-assembly could be detected by the downstream pressure transducer. As mentioned above, one individual maintainer/technician noted he was aware of at least one pump failing due to the pump shaft breaking because particulate accumulating on the gear teeth caused the gear pump to bind under load. Without reproducible evidence, this fault is difficult to conceptualize as a failure mode in practice. It is plausible that a larger piece of foreign-object-debris entered the gear pump and thus resulted in the shaft breaking due to the gears binding versus the failure resulting from the gradual accumulation of particulate on the gear teeth. Nevertheless, with this in mind, the following discussion documents our efforts to investigate this accumulation of particulate on the gear teeth as a potentially detectable failure mode.

Figure 13 shows the gear pump sub-assembly and the aluminum coating applied to one gear tooth for the simulated fault test. To re-iterate, the ideal test would utilize fielded gear pumps with particulate accumulated on the gear teeth or perhaps more likely in the gear's space width (teeth valleys). The gear fault testing incorporated only four run speed setpoint conditions compared to the 16 used for the subsequent bearing fault testing. The four run speeds were 600, 1300, 1700, 2100 RPM. These run speeds correspond to the four calibration setpoint conditions we obtained from the Cummins factory that calibrated our original PT pump.

The bar graph data in figure 14 summarizes the broadband RMS calculation of the fuel pressure spectrum for the no gear fault case (no coating) and the fault case (~0.001 inch coating) for the four run speed setpoint conditions.

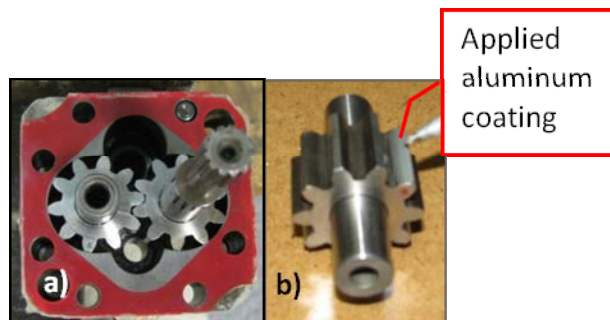


Figure 13. a) Gear pump sub-assembly, b) Metal coating applied to one gear tooth for simulated fault test

The data in figure 14 shows that this basic method does provide a low sensitivity condition indicator at the 1700 RPM and 2100 RPM pump speed setpoints but it does not provide an indication at 600 RPM and 1300 RPM. Based on these results the next step in the analysis was directed toward a narrowband frequency evaluation.

The spectrum characteristic we focused on was the distinct peak observed at 10X run speed or the 10th order as indicated by the yellow shaded box in figure 15. This frequency corresponds to the gear pump's Gear Mesh Frequency (GMF), which is equal to the shaft speed in Hertz multiplied by the number of teeth on the gear mounted on that same shaft.

With this stated, we continued the line of inquiry to examine whether a similarly consistent peak and RMS value would be detected in the pressure signal. Based on visual inspection, across this broad order range from 0 to 30 orders, there appeared to be two 'relatively' consistent indications of the gear fault in the pressure spectrum. The next step in the study focused on discrete frequency analysis as indicated by the two ordered frequencies marked in yellow as potential condition indicators in figure 15.

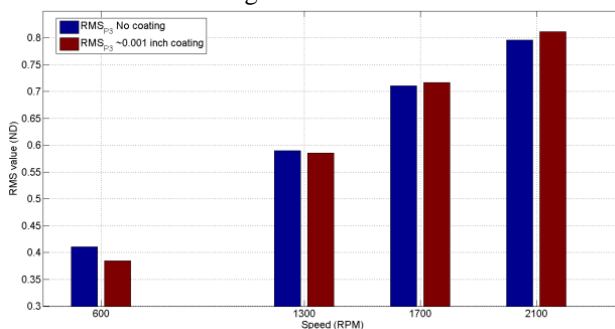


Figure 14. Bar graph summary of the broadband 0-500Hz RMS calculation of fuel pressure spectrum for the no gear fault case (no coating) and the fault case (~0.001 inch coating) for the four run speed set point conditions

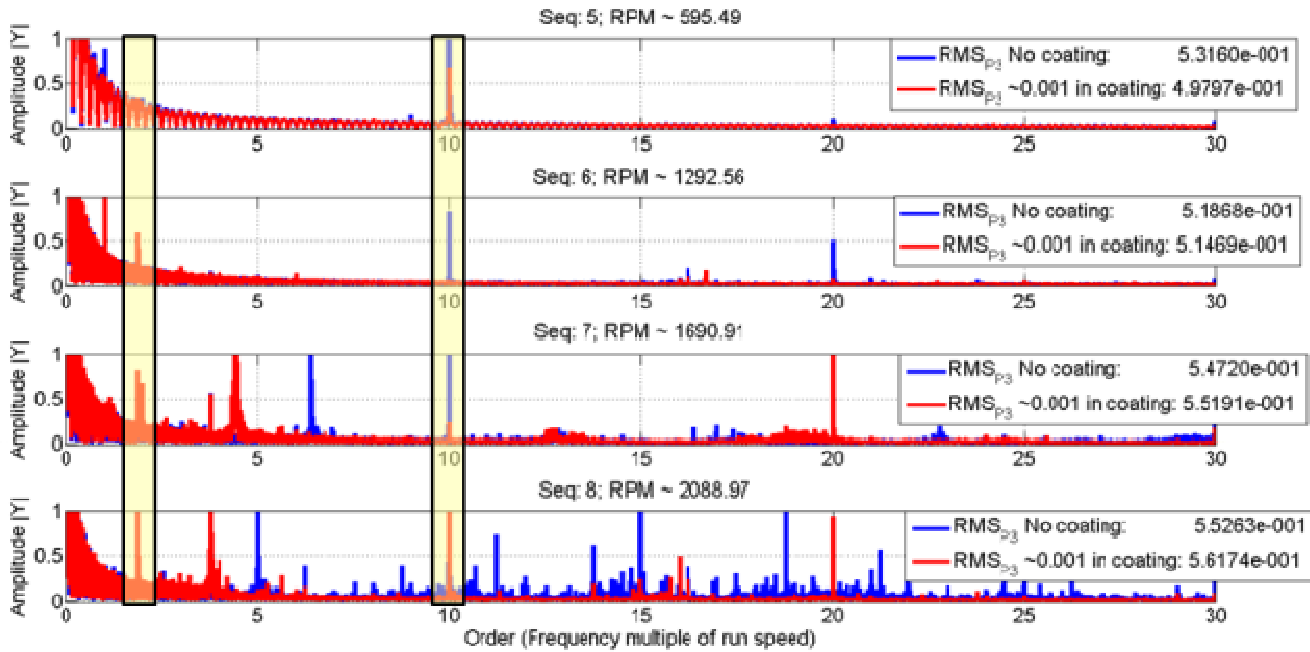


Figure 15. Fuel pressure frequency spectrum in the order domain with gear fault for the four run speed set point conditions

As stated above, because the 10th order frequency appears to be the most favorable ordered frequency marker, we narrowed our ordered frequency band over which we performed the RMS calculation to the 9th through 11th order range in order to determine whether a quantitative fault indicator may be more pronounced when only a specific band of the entire pressure spectrum is used for the calculation. However the RMS of the nominal 10th order frequency from the pressure sensor does not provide a positive correlation with the fault. It was therefore ruled out as a predictive condition indicator.

Based on visual inspection, the data indicates that there does appear to be a second discrete frequency in the pressure data that correlates to the seeded gear fault condition. This frequency is at approximately 1.8 orders as indicated in figure 15. The plot in figure 16 shows the positive correlation in the narrowband RMS calculation and the no fault versus seeded fault cases at the 1.8 order of run speed frequency. This finding suggests this discrete ordered frequency might potentially be useful as a predictive condition indicator. Pending further analysis using a statistically significant number of pumps, we emphasize two points: (1) There is no precedent or physical explanation providing the rationale to focus on the 1.8th order of run speed; This order was selected based on manual inspection of the spectrum. The objective was to identify orders at which the RMS results demonstrated a positive correlation with respect to presence of the gear tooth coating; (2) It is not yet confirmed whether the correlation at this run speed order is consistent for a larger sample set of pumps.

5.4 Gear Fault Test Conclusion

The results of the gear fault detection investigation are not decisive but did not rule out the potential for an effective predictive gear fault detection capability using the existing pressure sensor installed on the CMED variant Bradley.

The preliminary analysis conducted using the accelerometer data showed that the broadband (0-500 Hz) RMS vibration spectrum does not provide a predictive condition indicator. The discrete frequency analysis indicated that the 10th order, which is the gear mesh frequency for the gear pump, did not prove effective when utilized with the pressure sensor. The next step in the analysis led to an assessment of other effective discrete frequency indicators that could be used with the pressure sensor. It was determined that a discrete frequency at the 1.8th order could potentially be utilized as an

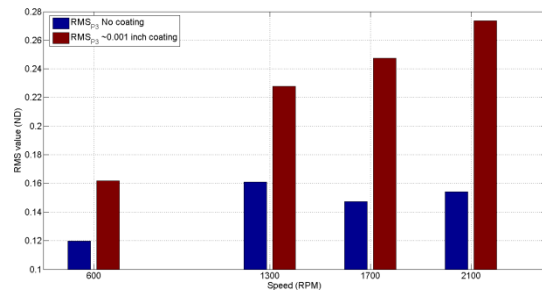


Figure 16. Bar graph summary of nominal 1.8th order RMS calculation of fuel pressure spectrum with gear fault for the four run speed set point conditions

effective predictive condition indicator for the gear fault case. The relationship between this frequency and the gear fault has not been determined at this point. Among other considerations, the effect of the pressure pulses generated by the fuel injectors requires study in order to rule out the potential it may possess to mask the characteristic spectral signature of this fault. Further testing would be conducted to validate this condition indicator.

6. CONCLUSION

This paper documents the test stand setup and analysis methodology used to investigate the feasibility of detecting seeded gear and bearing faults in a PT pump using an existing on-board M2 Bradley fuel pressure sensor with a dynamic bandwidth of 0-500 Hz. The results are not statistically valid, nor are the results consistent at all speed setpoint conditions. With this stated, there is limited evidence suggesting that it may be feasible to detect a 0.001 inch particulate accumulation on the gear pump teeth using narrowband RMS based quantification methods. Four additional statistical signal processing features yielded no more consistent results across the range of speed setpoints examined. The inconsistencies associated with run speed are not fully understood. Further study would enable confirmation as to whether structural frequencies associated with the pump or the test stand configuration may be contributing factors. The data processing employed in this study utilized a 100 kHz sampling rate to acquire the pressure signal and a one pulse per revolution tachometer signal for time synchronous averaging. Limitations to the data processing are dependent on the intended on or off-board end-use implementation. On/off-board implementation must address electro-magnetic interference and signal pre-conditioning considerations along with associated cable shielding, computer processing, data storage and user interface requirements. Given the relatively gradual lead up to PT pump failure given bearing and gear faults of this type, further study would investigate the cost effective feasibility of detecting such bearing and gear faults at idle speed using an at-platform maintenance-bay diagnostic software tool approach. While on-board detection at higher speeds may be feasible in theory, the additional cost of data acquisition, instrumentation and sensor/DAQ maintenance along with the complexity associated with noise in a field environment, may not justify maintenance and logistics costs.

ACKNOWLEDGEMENT

This work was supported by the NAVSEA Contract Number N00024-D-02-D-6604, Delivery Order Number 0356. The content of the information does not necessarily reflect the position or policy of NAVSEA, and no official endorsement should be inferred.

REFERENCES

- Banks, J.C., Reichard, K.M., Hines, J.A., Brought, M.S. (2008). Platform Degradation Analysis for the Design and Development of Vehicle Health Management Systems. *IEEE Prognostics and Health Management (PHM) Conference*, October 6-9, Denver, CO. doi: 10.1109/PHM.2008.4711468
- Cummins Component Shop Manual for Cummins PT Fuel Pump Rebuilding and Calibration Instructions (1980) Bulletin No. 3379084-02
- Cummins Fuel Pump PT (type G) Calibration Values Bulletin No. 3379352-10
- DTSFE (2011) Electrolytic corrosion <http://www.dtsfe.com/faq/pdf/electolytic%20corrosion.pdf>
- Hines, J.H., Bennett, L., Ligetti, C., Banks, J.C., Nestler, S. (2009) Cost-Benefit Analysis Trade-Space Tool as a Design-Aid for the U.S. Army Vehicle Health Management System (VHMS) Program, *Prognostics and Health Management (PHM) Society 2009 Conference*, September 27-October 1, San Diego, CA.
- Technical Manual 9-2350-294-20-1-1/3 (2006), Technical manual unit maintenance manual Fighting Vehicle, Infantry, M2A2 2350-01-248-7619, Department of the Army, December 1, 2006
- White, D.G. (1995) *Introduction to Machine Vibration*: Bainbridge Island, WA, DLI Engineering Corp.: Part number 8569, version 1.76 (p. 110)



Jeffrey C. Banks is the Department Head of Complex Systems Engineering & Monitoring. His education includes a B.S.M.E. from Villanova University, Villanova, PA and a M.S. in Acoustics from The Pennsylvania State University, University Park, PA. He has 16+ years experience in applying advanced signal processing techniques, intelligent systems technology, and embedded diagnostics / prognostics tools to condition monitoring applications for the US Navy, US Marine Corps, US Army, NASA and Industry. His research engineer duties include developing machinery health management systems and diagnostic and prognostic technology for various DoD assets including the U.S. Marine Corps Expeditionary Fighting Vehicle (EFV), Light Armored Vehicle (LAV), AV-8B Harrier, U.S. Army Heavy Expanded Mobility Tactical Truck (HEMTT), Family of Medium tactical vehicles (FMTV) and U.S. Navy CVN class aircraft carriers. He has developed and delivered two short courses at all NASA facilities in the areas of Condition Based Maintenance (CBM) and Reliability Centered Maintenance (RCM). Additional responsibilities include conducting Failure Modes, Effects and Criticality Analysis (FMECA) for a variety of complex systems and platforms including aircraft engines and combat ground vehicles. He has also designed and developed diagnostic

instrumentation systems for machinery failure test beds and field data acquisition systems. He has first authored and published more than 18 papers in conference publications, and technical reports.

Before coming to Penn State ARL, he worked for the Mead Central Research Laboratory as a research engineer in the paper industry. He led over 100 diagnostic surveys at 9 separate paper manufacturing facilities and his primary functions were to troubleshoot machinery maintenance problems and process variability issues and develop unique predictive failure technology. He also developed and implemented vibration analysis surveys to evaluate resonance issues and condition of rotating elements and structural systems used in the paper industry.

J. Scott Pflumm is a research engineer at the Applied Research Laboratory. He received his B.S.M.E and M.S.M.E in Mechanical Engineering from Penn State University in 2002 and 2005 respectively. He served 3 years in the U.S. Army. His current work at ARL is in support of test and analysis of failure modes for ground vehicles at the sub-system and component level.

A Discussion of the Prognostics and Health Management Aspects of Embedded Condition Monitoring Systems

Roger I. Grosvenor¹ and Paul W. Prickett¹

¹*School of Engineering, Cardiff University, Cardiff, South Glamorgan, CF24 3AA, Wales, UK*

*grosvenor@cf.ac.uk
prickett@cf.ac.uk*

ABSTRACT

This paper presents a review of embedded condition monitoring research carried out at Cardiff University. A variety of application areas are described, along with a discussion of the evolution of the hardware platforms used. The current operating philosophies of the Intelligent Process Monitoring and Management (IPMM) research group and the deployed hierarchical and distributed architectures are described. The paper sets out to discuss the on-going trend towards such monitoring systems needing to provide more than fault detection and diagnostic capabilities. System requirements such as tracking operational settings, performance and efficiency measures and providing limp-home facilities are seen to be consistent with prognostics and health management ideals. The paper concludes with a discussion of new and future developments and applications.

1. INTRODUCTION

The Intelligent Process Monitoring and Management (IPMM) research group at the Cardiff School of Engineering has 20+ years experience of condition monitoring research. The following sections describe industrially related application areas and track the evolution of technologies and approaches. The associated discussions describe how modern monitoring systems must provide far more than fault detection and diagnosis.

Originally the IPMM research concentrated on machine tool applications used heavily sensor-based techniques, using PC platforms and interfaces, and working with large companies.

With technologies changes, and following an ERDF funded project aimed at SMEs in south Wales, distributed, microcontroller-based systems became the main area of research.

* This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Table 1. Condition Monitoring System Concepts

Distributed,	data acquisition / monitoring nodes linked by a CAN-bus network
Hierarchical,	3 tier approach – higher levels provide data fusion and robust decisions
Remote,	deployed systems linked to remote base via Internet
Intelligent	minimised data communications and storage
Low-Cost	8-bit microcontrollers used, 96% of faults detected locally.....
Monitoring Systems only PC is server-side and used to provide higher level analysis... of remaining 4 % of faults.

There was an accompanying diversification of application areas. The machine tool work continued, but with PLC controlled systems, process and environmental / energy systems added to the range of monitoring applications.

Table 1 reflects on the main concepts employed with the latest generations of the microcontroller-based monitoring systems.

The low cost and ease of use of these systems led to their application to a range of monitoring functions for machine tools and process plant, for example as reported by Siddiqui et al (2010) and Siddiqui et al (2007) respectively. Initially limitations, in terms of processing capabilities restricted their application. However, the current generation of microcontroller devices, such as those now deployed has largely overcome such limitations. As will be discussed later, a generic microcontroller platform is often adopted as a starting point.

The monitoring systems deployed by the IPMM group are not exclusively based upon the described microcontroller platform. For example, SCADA based systems have been used for both the monitoring of water treatment plants and

cooling towers within a power plant. In other applications, where higher processing capabilities are typically required, PC based systems have been deployed. Examples include the monitoring of crop shear and roughing mill operations in a steel plant, where sensory inputs were constrained to being provided by acoustic microphones. In other applications the microcontroller-based systems have been used to pre-process data, with the aim of reducing the data processing, communications and archiving tasks on predominantly PC based systems. .

2. MACHINE TOOL MONITORING

Earlier research, as summarised by Drake et al (1995), concentrated on the data acquisition and signal processing aspects of machine tool monitoring. In parallel, many of the constituent sub-systems were researched, in a prioritised manner derived from industrial reliability information, from a fault detection and diagnostic viewpoint. Examples included axis drives (Rashid & Grosvenor, 1997), tool changer & coolant sub-systems (Frankowiak et al, 2005) and the cutting tools (Amer et al, 2007).

Hess et al (2005) described the constituent functions and processes for Prognostics and Health Management (PHM) systems. In addition the timely and correct acquisition of signals is a vital element of any monitoring and/or PHM system. The approach within the IPMM research is argued to be consistent with these guiding themes. For example, with the machine tool research, the primary aims have been to reduce the downtime of such high capital cost, high utilisation machines. The challenge is to provide sufficient lead time / warning to the operator of progressive faults and to handle the fault detection and isolation of 'hard' (catastrophic) faults with sufficient fault library coverage. A higher level of fault information is then made available to the service / maintenance teams to assist their corrective actions. Further, techniques such as Overall Equipment Effectiveness (OEE) may be used to provide a longer term tracking of the health and performance of the machines. In the context of machine tool monitoring the provision of a scaled indication of the feasibility of continued use is a useful feature. Should the operator immediately halt the machining process, or is there a possibility to complete the existing job or batch (perhaps at reduced cutting speeds and feedrates), or can the machine be run until the next convenient maintenance opportunity?

Rather what has evolved has been the data acquisition, signal processing and computing platforms utilised, along with consideration of the number of additional sensors to be fitted to the machine for monitoring purposes. The data acquisition system (DAS), from the early 1990's work mentioned above, was based on a PC platform and utilised a large number of analog and digital inputs and custom designed interface cards to form the DAS. During remotely sited industrial trials up to 21 additional fitted sensor

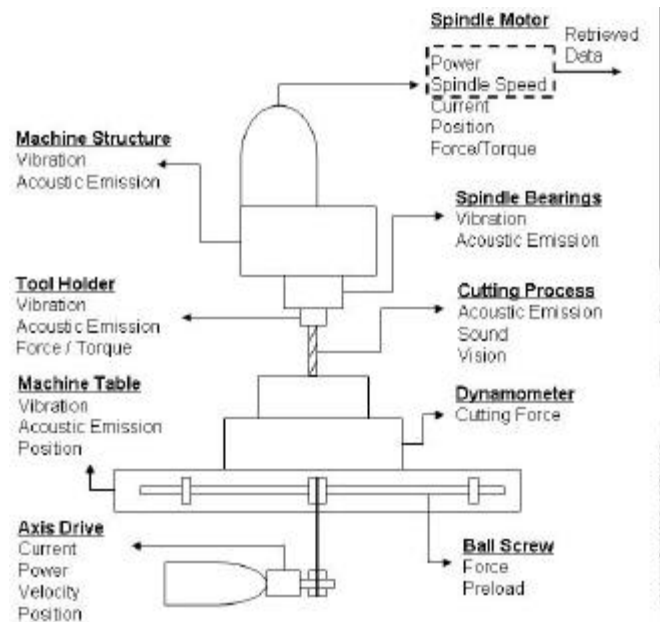


Figure 1. Potential Machine Tool Monitoring Measurements

signals, along with 14 signals from sensors that pre-existed on the machine and 46 digital signals were used. The digital signals derived from the CNC and from limit switches etc were used, with a database defined series of diagnostic tests, to provide consistency for trend comparisons and for determining best matches to the established fault library.

Also, to eliminate the variability from all machining operations whilst providing on-line monitoring of the machine tools themselves the database configured tests to capture diagnostic information during all periods whilst the machine was on / moving but not actually cutting metal. The pre-internet enabled communications retrieval from the remote locations and the then limited PC storage capabilities also required a variety of (database configurable) signal processing / data reduction methods. Figure 1 provides a summary of many of the potential measurements that can and have been used in machine tool monitoring.

For the more recent distributed and embedded monitoring systems, and making use of the increased processing power and communications protocols, single chip microcontrollers have been utilised. The number of additional sensors has been dramatically reduced. Continued use is made of any suitable sensors pre-fitted on the machine tool, with their potential for monitoring typically being assessed during an initial auditing phase. Carefully designed monitoring tests then often infer fault conditions from a collection of inputs, which are acquired from the lowest level of the microcontroller based nodes. The next higher level node coordinates and provides more robust decision making from the available information. In a more general sense, Jacazio et al (2010) have reported on the role of logical and robust decision making elements of sensor-based PHM systems.

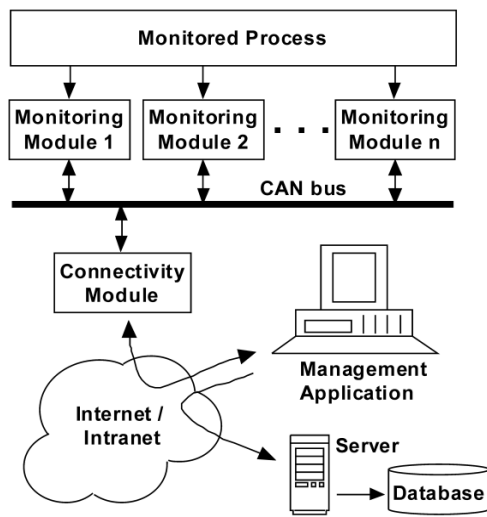


Figure 2. Monitoring System Architecture

The proposed and relatively simple monitoring algorithms are then developed and tested. Research machine tools and / or representative scaled physical models of sub-systems are used to deliberately introduce typical faults. For the case of cutting forces and tool wear / breakage detection the effectiveness of using inferred measures, from motor currents for example, is tested against higher cost dynamometers during this development phase.

Table 1 shown previously describes the overall monitoring system parameters and Figure 2 provides the remote monitoring architecture. A number of (PIC) microcontroller monitoring nodes are deployed on the machine (or process) to be monitored. These are typically capable of detecting 80% of all faults, mainly trivial, low level hard faults. These are connected to each other and to the 2nd level microcontroller via a CAN bus communications protocol. The CAN bus protocol is heavily used and was developed for automotive applications, and its robust performance in harsh and noisy environments make it ideal for machine monitoring applications.

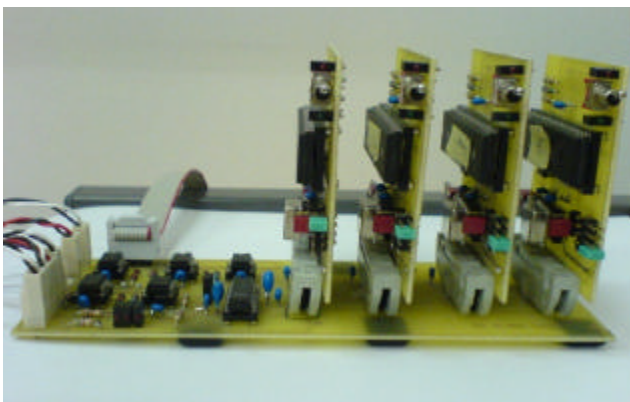


Figure 3. Monitoring Modules Hardware

The level 2 node, as stated, co-ordinates the information from the other nodes and typically 80% (of the 20% not detectable at the lowest level nodes) of remaining fault coverage is provided. These will require more sophisticated diagnostic methods compared to the previously described low level hard faults. More typically the early detection of faults whose level of severity increases with age would need to be detected and diagnosed at this node. The node also provides the internet based communications (at low level UDP protocols) back to the PC-based server. For the 4% or so of faults requiring higher processing capabilities and algorithms the monitored data may be streamed back.

Figure 3 shows the hardware developed, in this case for a batch process application (Ahsan et al, 2006). The 4 vertical circuit boards, each with a PIC microcontroller, are connected to analog signals measuring flowrate, temperature, liquid level and pump power, and are the monitoring modules. The horizontal circuit board includes the connectivity module and CAN bus communications (brought physically close together in this implementation).

2.1 Petri Nets

The group has used Petri Net techniques for a variety of applications, including machine tool monitoring. Initially, and in line with Petri's original concepts, the Petri Nets were used as a graphical user interface. In simple terms, the operator could view the dynamic flow of coloured tokens around the defined Petri Net diagram. A review of the use of Petri Nets in monitoring applications is provided by Frankowiak et al (2009). The approach was then adapted to provide the context and consistency of the defined monitoring tests and to reduce the amount of re-programming of the microcontroller nodes when deployed on new applications. Frankowiak et al (2009) concluded that the extensions provided, to conventional Petri Net representations, facilitated the interfacing and handling of real-life process signals. The addition of thresholded analog inputs and other constructions more suited to monitoring rather than control of sequential process was deemed to be vital to the evolution of low-cost monitoring systems.

The coding of the particular Petri net representations was demonstrated for a machine tool changer, a conveyor based assembly process and for a hydraulic press. The coding of the look-up tables within the microcontroller programs allowed for a selective approach to which sequence transition data were recorded and transmitted. This enabled both OEE calculations and the population of dynamic web page displays at the server-side PC. In particular the recording of start and end transitions enabled cycle time calculations. The counting of branched states, for example representing good or bad assemblies (for the conveyor application) enabled a quality measure. The third constituent of OEE calculations was then provided by the time-out / alarms of the Petri Net transitions in faulty conditions. The

look-up tables were achievable within the memory constraints of the microcontroller hardware. Each microcontroller node could be interfaced to up to 24 digital inputs, 4 analog inputs and 2 pulse train inputs and could provide 1 digital output.

2.2 Microcontrollers

The PIC microcontrollers used are simple 8 or 16 bit single chip devices, whose features and capabilities have advanced with time. They conveniently handle inputs and outputs and have a variety of communications protocols. Originally the use of CAN bus communications required an additional transceiver chip alongside the PIC. These days PIC devices are available with built-in CAN bus facilities. The simple PIC devices do not have extensive capabilities for mathematical manipulations and/or diagnostic algorithms, although comparison of inputs to pre-determined threshold levels are readily implemented. The considerations for more advanced signal analysis, such as frequency analysis will be discussed here as an example.

Amer et al (2007) reported on the use of sweeping filters for machine tool condition monitoring. A PIC 18 series microcontroller was deployed as one of the monitoring module nodes. It was used to control an analog programmable filter, in the stated application to detect breakage of a 4-toothed milling cutter. The limitations were such that, in effect, a 32 point Fast Fourier Transform (FFT) of spindle load signal on the milling machine was achieved. The filter was swept through the determined and appropriate range of frequencies and enabled the PIC to accumulate sufficient data, with the available timeframe, to determine a limited resolution frequency spectrum. The system was developed through a series of cutting trials, with a range of set machining conditions and for 3 and 2 tooth cutters, in addition to the 4 tooth cutters. The approach was successful, when considered within the context of the first level diagnostics within the hierarchical monitoring system.

The detection of the breakage of milling cutters is a challenging task. In a survey of health management user objectives, Wheeler et al (2010), included considerations of diagnostics and diagnostic metrics. They included detection rates, detection accuracy and detection response time as desirable objectives. For milling breakage detection there is a premium on the detection response time, particularly for high value, long cycle time, minimally supervised machining jobs. The use of better resolution and more sophisticated FFT that was then enable by the next generation of microcontrollers, known as dsPICs, was reported by Siddiqui et al (2007). Figure 4 shows the structure of the dsPIC system. The dsPIC is a 24 bit device and has digital signal processing (DSP) commands along with the established PIC I/O handling and communications. It also has higher resolution analog signal acquisition and

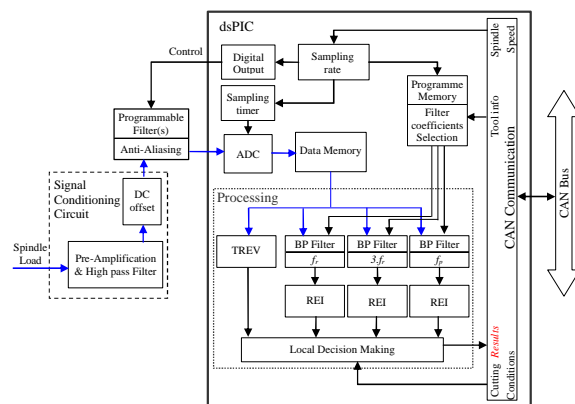


Figure 4. Schematic of dsPIC Monitoring System

in-built FFT routines. Siddiqui et al (2007) implemented an overlap FFT processing scheme in order to address the demanding detection response time requirement. The reported results showed that a robust detection of sudden tool breakage could be achieved within 1.5 revolutions of the spindle (and cutter) post failure. The efficient coding of the software and algorithms meant that detection could be achieved for spindle speeds up to 3000 r.p.m. The monitoring system was designed to be relatively immune to false alarms, even under a range of machining conditions. These included break-in and break-out (these often trigger false alarms in such monitoring systems), variable depth of cut and a range of (operator selected) spindle speeds. For the latter case the sample rate of the dsPIC was changed under software control in order to 'track' the intended and particular frequencies of interest.

Further the derived states of the frequency components, at the spindle rotational frequency (f_r), at 3 times this frequency ($3f_r$) and at the tooth passing frequency (f_p) were fed into a decision maker. The other parameters used by the decision maker were a Tool Rotation Energy Variation (TREV) and a Relative Energy Index (REI). Table 2 summarises the decision making logic. If a clear categorization was not directly possible then either further frames of data could be captured and processed or the raw data could be passed up the monitoring hierarchy for more advanced frequency analysis.

The other aspect of the milling cutter monitoring system that may be of relevance to PHM approaches is the estimation of tool life. Often milling cutters are deployed with a (usually) conservative estimate of expected lifetime. The parameters used with the dsPIC monitoring system are also used to calculate the accumulated usage time of the cutting tool. The energy based monitoring calculations further enable the usage time to be considered in combination with a measure of how hard the tool was used and when it was actually in use, cutting metal. This provides refinements compared to simple logging of the calendar age of the tool or machine-on

Table 2. Decision Making Table

Mean Freq	Magnitude ²			Pattern	Decision
	f_r	$3f_r$	f_b		
0	0	0	0	0	Healthy
0	0	0	1	0	Blunt Tool
0	1	X	X	0	Unexpected : request advanced diagnosis
0	X	1	X	0	Unexpected : request advanced diagnosis
1	0	0	0	0	Wait for next Frame
1	0	0	0	1	Chipped Tool
1	1	X	X	1	Broken Tooth
1	X	1	X	1	Broken Tooth
1	X	X	1	1	Broken Tooth

hours. Potentially such lifetime profiles could be used towards the end of the useful life to determine whether a particular job could be finished, for example at reduced machining rates, with the existing tool.

3. INDUSTRIAL MACHINE / PROCESS MONITORING

3.1 Embedded Monitoring Applications

The previously described microcontroller systems were also deployed, as stated in section 2.1, to monitor a conveyor based assembly process and a hydraulic press. Both of these are good examples of industrial processes whose sequence and logic is controlled by a Programmable Logic Controller (PLC). The conveyor assembly monitoring system was interfaced to 14 digital signals, utilizing both inputs to and outputs from the programmed PLC. The defining Petri Net structure had 63 transitions and was predominately a branched structure. This reflected the various outcomes at the sorting, assembly, overflow and accept/reject stages of the process. The Petri Net was configured to enable OEE calculations and the remote tracking via dynamic web pages of the assembly process performance. Figure 5 shows one example of such performance tracking. The pie chart reflects the counts of well assembled parts, incorrect assemblies or parts and reprocessed parts. The numbers preceding the counted occurrences are the respective Petri Net transition numbers.

For the hydraulic press application 22 digital signals from the PLC were used along with 3 analog signals used to measure the motor currents on each axis of motion. The Petri Net representation had 29 transitions and the only branching required depended on whether a left hand or right hand pallet was selected for pressing actions by the press

PIC-ConveyorOutput

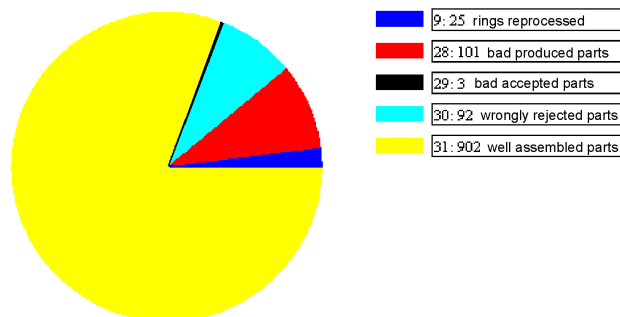


Figure 5. Dynamic Webpage Example for Assembly Process Application.

operator. The movement of the vertical axis provided a good example of where context based (provided by the Petri net structure) thresholding of signals was required. The vertical motor currents, for normal fault free operations were different for upwards and downwards movements. The monitoring system was again configured to provide cycle times, loading times and fault diagnostics.

Prickett et al (2010) reported on the monitoring of pneumatic systems, such as linear actuators and grippers. These are widely used in the automotive, manufacturing and food packaging industries. The dsPIC microcontroller system was used to detect the presence of parts and identified their size in real time during gripping operations. Key timing in measured pressure response profiles during a gripping cycle were identified. A modelled 3D surface that described the actuator movement and the effect of air supply pressure and stroke length was then utilised. The timings could then be used to confirm that the correctly sized component had been gripped and that it had not slipped or had been dropped during the actuation cycle.

3.2 Monitoring Applications Using Other Platforms

Sharif and Grosvenor (1997) used a PC based system in the monitoring of pneumatically actuated process control valves. In this application the monitoring system was used to complement the built-in diagnostics and test cycles of the valve's digital position controller. A test rig was established and the fault diagnosis capabilities were assessed following the introduction of simulated faults. It was reported that a range of fault conditions and their levels could be detected with the addition of 1 extra pressure transducer. The faults were deemed to be representative of harsh and arid type pipeline conditions. The faults were vent hole blockages, diaphragm ageing & cracking and damage to the valve stem

seal due to accumulated deposits on the valve stem. The problem of internal leakage through the valve was separately investigated and was found to require the addition of acoustic emission sensors.

Eyers et al (2005) considered the monitoring of a robotic welding station. The industrial application began with the deployment of a commercial system that interfaced up to 4 sensor signals from the machine on the factory shopfloor to an office based location. This device used Bluetooth class 1 communications but was found to be inflexible in terms of file storage. This rendered the viewing of longer term trends difficult and required large file storage capacity. The developed PC-based monitoring system was accordingly focussed on intelligent data management and reporting. Web-based OEE statistics were generated and a 99% reduction in the daily traffic of monitored information was achieved. Significant differences in completed welding operations across the 3 shifts per day were observed and the industrial partner was then able to instigate performance improvement measures. The shift-by-shift reports and the weekly and monthly trends were reported via a number of mechanisms and technologies.

4. OTHER MONITORING APPLICATIONS

Edwards et al (2006) considered monitoring techniques for determining lamp condition in lighting applications. The proposed approach required the measurement of a combination of lamp characteristics in order to accurately determine remaining life. Testing was carried out with filament lamps, low pressure discharge lamps and UV sterilization lamps. In the case of filament lamps it was found that strong correlations existed between initial characteristics and lamp life. A short duration (30 seconds) test of each lamp could then be used to predict the remaining useful life. A multi channel PC based test rig was used to test multiple lamps and to gather the data used to establish the correlations.

Davies et al (2009) used a SCADA based system to obtain PLC information for water treatment plant and cooling tower applications. The water treatment plant monitoring mainly consisted of detecting pump and piping blockages and of determining the performance of the programmed schedule of filter bed backwashing actions. The Citect SCADA software that was utilised acted as an OPC server and was hosted on a PC platform. The initial detection of potential faults was triggered if a raised speed request from the PLC continuous PID control loop was detected. This could indicate the controller 'working harder' to maintain the set flowrate of water for treatment in the filter bed. This could be in reaction to either single or combined blockages, of the upstream or downstream pipework or could indicate that the filter is in need of backwashing. A diagnostic program then ran and manipulated the speed request signal to create a test cycle. The flow and pressure signal profiles

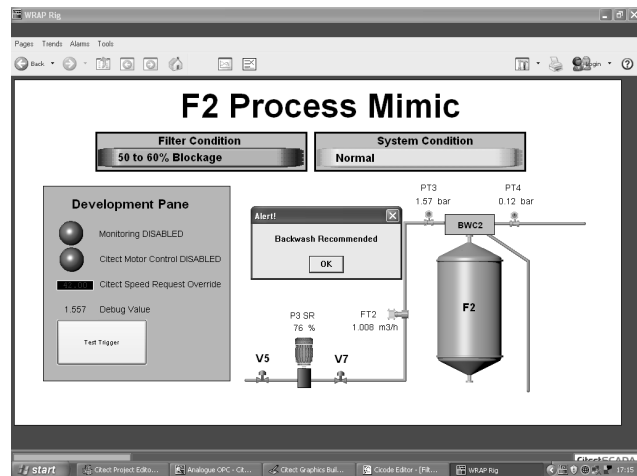


Figure 6. Process Mimic Screen for Water Treatment Plant.

obtained were then used to determine the fault conditions. The diagnostic program could estimate blockage levels for both single and multiple fault scenarios and could distinguish pump and pipework blockages from filter bed fouling. The triggering of filter backwashing was implemented when required. The system however was also used to optimise the duration between scheduled backwash operations. The operator was provided with a process mimic summary screen, an example of which is shown in Figure 6.

For the cooling tower application the monitoring system was also required to track the chemical dosing regime. The system also helped to co-ordinate and optimise the selective operation of 3 cooling towers in varying operational conditions. The system also provided accurate real time information on the energy usage and efficiencies and provided the manager with a financial costing screen.

5. DISCUSSION OF FUTURE APPLICATIONS AND PHM TECHNIQUES

5.1 New and Future Monitoring Applications

One example relates to the emerging technologies for tidal turbines and renewable power generation. In many cases the proposed monitoring schemes are deemed to be analogous to those deployed on wind turbines. Owen et al (2010), for example, have reported on a multi-mode structural health monitoring system for wind turbine blades and components. Certainly in considering typical generic designs that are emerging for tidal turbines there are, at the sub-system level many similar components to wind turbines. The operating conditions and medium are vastly different. The IPMM group is considering the monitoring and PHM requirements that are likely to be embedded within tidal turbines. As a starting point a Failure Modes and Effects Analysis (FMEA) provides a vehicle for the systematic

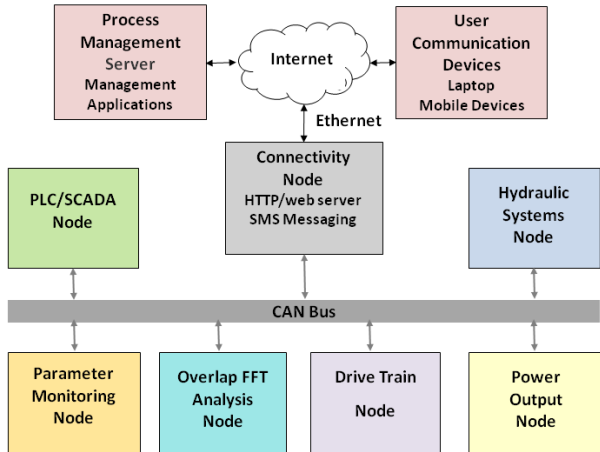


Figure 7. Representation of the Marine Tidal Turbine Condition Monitoring Architecture

analysis of potential failure modes to reduce and if possible prevent failures. An effective FMEA can identify critical points within the design, manufacture, installation and operation of components, characterise failure modes, actions also direct the specification and configuration of condition monitoring systems that can support the successful operation of marine tidal turbines. Values for the severity, occurrence and detection ratings for constituent sub-systems are multiplied to produce risk priority numbers (RPN). The group plans to implement an embedded monitoring system and to initially test and develop the system on scale models of the turbines. These are being used in water flume testing for the validation of computational fluid mechanics (CFD) mathematical models. An outline of the system architecture is shown in Figure 7. A representation of the main constituents of a generic tidal turbine that will require monitoring is provided in Figure 8.

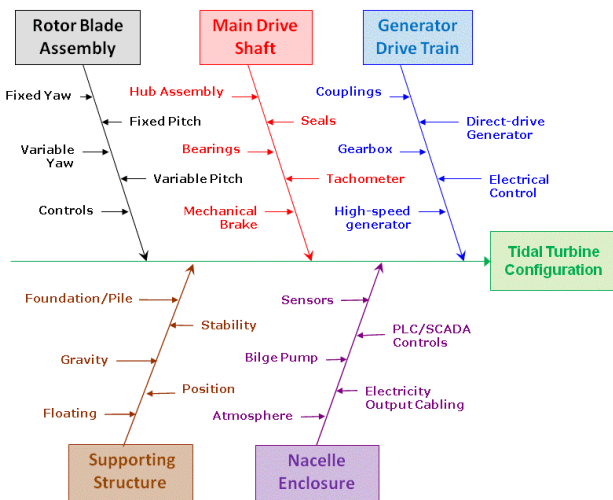


Figure 8. Representation of Possible Turbine Configurations.

5.2 Discussion of the Need for Condition Monitoring Systems to Embrace PHM Techniques.

The authors believe that the modern and future generations of monitoring systems need to provide more than just condition monitoring and fault diagnostic functions. It is hoped that the range of monitoring applications reported in this paper already contain some of the PHM philosophies and techniques. The dsPIC based embedded and distributed monitoring architectures reported are believed to provide a potential platform for future developments.

It is hoped that the experiences reported may be of use to other PHM practitioners when they initially consider which approaches and platforms for their applications.

When working with multiple distributed applications and/or small resource limited organizations the lower cost microcontroller platforms and the selective use of key additional sensors will almost inevitably be a constraint or be of fundamental importance. Further consideration should be applied to which of the available range of techniques is most appropriate. For example, the simple comparison of an analog signal level to a set threshold will have low microcontroller resource implications. It would not be over demanding in terms of sample rates, processing power or data storage and communications. In other cases, for example machine tool or rotating machinery applications, the ability to provide FFT processing would almost certainly be a vital requirement. A more detailed analysis of the microcontroller resources would be needed.

The IPMM group is, as stated, applying such considerations to the monitoring of future generations of tidal stream turbines. It is envisaged that ruggedized commercially available modules, such as the compactRIO system from National Instruments will be investigated in conjunction with some of the reported microcontroller modules.

It has been reported that PHM involves interdisciplinary research with a broad range of application areas. The detection of impending faults remains a key objective and in a wider PHM system allows logistical decision making. This along with the concept of transforming data into information and onwards to decisions is consistent with the condition monitoring approaches reported in this paper.

There is potential to expand the monitoring research to have more explicit links to reliability predictions and to more fully consider lifetime management of components and systems.

REFERENCES

- Ahsan, Q., Grosvenor, R.I. & Prickett, P.W. (2006) Distributed On-Line System for Process Plant Monitoring. *Proc I.Mech.E, J.Process Mech.Eng.*, vol 220 (e2), pp 61 – 77. doi: 10.1243/09544089JPME53.
- Amer, W., Grosvenor, R. & Prickett, P. (2007) Machine Tool Condition Monitoring Using Sweeping Filter Techniques. *Proc. I.Mech.E. J. Systems & Control Eng.*, vol 221 (1), pp 103 – 117. doi: 10.1243/09596518JSCE133.
- Davies, G.R., Grosvenor, R.I., Prickett, P.W. & Lee, C. (2009) An Approach to the Detection and Characterisation of Faults in a Water Treatment Plant. *Proc. Comadem*, (pp 553 – 558), June 9-14, San Sebastian. ISSN 9788493206468.
- Drake, P.R., Grosvenor, R.I. & Jennings, A.D. (1995) Review of Data Acquisition System Developed for the MIRAM Project. *Conf. Proc. Sensors & Their Applications VII*, (pp 433 – 437), Sept 10-13, Dublin. ISBN 0 7503 0331 X
- Edwards, P.M., Grosvenor, R.I. & Prickett, P.W. (2006) A Review of Lamp Condition Monitoring Techniques. *Lighting Journal*, vol 71 (5), pp 31 – 38. ISSN 0950-4559.
- Eyers, D.R., Grosvenor, R.I. & Prickett, P.W. (2005) Welding Station Condition Monitoring using Bluetooth Enabled Sensors and Intelligent Data Management. *Sensors & their Applications XIII*, (pp 143 – 148), Sept 6-6, Greenwich. doi: 10.1088/1742-6596/15/1/024.
- Frankowiak, M., Grosvenor, R. & Prickett, P. (2005) A Review of the Evolution of Microcontroller-Based machine and Process Monitoring. *International Journal of Machine Tools & Manufacture*, vol. 45, pp 573 – 582. doi: 10.1016/j.ijmactools.2004.08.018.
- Franowiak, M.R., Grosvenor, R.I. & Prickett, P.W. (2009) Microcontroller-based Process Monitoring Using Petri-Nets. *EURASIP Journal on Embedded Systems*, ID 282708, pp 1 – 12. doi: 10.1155/2009/282708.
- Hess, A., Calvello, G. & Frith, P. (2005) Challenges, Issues, and Lessons Learned Chasing the “Big P”: Real Predictive Prognostics Part1. *Aerospace Conf. IEEE*, (pp 3610 – 3619) March 5-12. Big Sky MT. doi: 10.1109/AERO.2005.1559666.
- Jacazio, G., Risso, D., Sorli, M. & Tomassini, L. (2010) Advanced Diagnostics of Position Sensors for the Actuation Systems of High-Speed Tilting Trains. *Conf. Prog.&Health Mangt. Soc.* Oct 10 -16, Oregon.
- Owen, R.B., Inman, D.J. & Ha, D.S. (2010) A Multi-Mode Structural Health monitoring System for Wind Turbine Blades and Components. *Conf. Prog.&Health Mangt. Soc.* Oct 10 -16, Oregon.
- Prickett, P.W., Grosvenor, R.I. & Alyami, M. (2010) Microcontroller-Based Monitoring of Pneumatic Systems. *5th IFAC Symp. Mechatronic Systems*, (pp 614 – 619) Sept 13-15, Cambridge MA.
- Rashid, M. & Grosvenor, R.I. (1997) Fault diagnosis of Ballscrew Systems Through Neural Networks. *Proc. Comadem '97*, (pp 142 – 151), June 9-11, Helsinki. ISBN 1 901892131.
- Sharif, M.A. & Grosvenor, R.I. (1998) Fault Diagnosis in Industrial Control Valves and Actuators. *Proc. IEEE Instn&Meast.Tech Conf*, (pp 770 – 778), May 18-21, Minnesota. ISSN 0 7803 4797 8/98.
- Siddiqui, R.A., Amer, W., Ahsan, Q., Grosvenor, R.I. & Prickett, P.W. (2007) Multi-band Infinite Impulse Response Filtering using Microcontrollers for e-Monitoring Applications. *Microprocessors and Microsystems* 31, pp 370-380. doi: 10.1016/j.micpro.2007.02.007.
- Siddiqui, R.A., Grosvenor, R. & Prickett, P. (2010) An Overview of a Microcontroller based approach to Intelligent Machine Tool Monitoring. *Lecture Notes in Computer Science*, 6277, 371-380. doi: 10.1007/978-3-642-15390-7_38.
- Wheeler, K.R., Kurtoglu, T. & Poll, S.D. (2010) A Survey of Health Management User Objectives in Aerospace Systems Related to Diagnostic and Prognostic Metrics. *Int. J. Prog. & Health. Mangt.*

Dr Roger I. Grosvenor

Educated at Cardiff University obtaining a BEng Mechanical Engineering degree (1978), a MEng via fluid dynamics research (1981) and a PhD on the topic of in-process measurement of machined components (1994). He is currently a reader in systems engineering at Cardiff University and has been employed as a lecturer there since 1984. He has published 90 journal and conference papers, mainly on the topic of machine and process condition monitoring. He is co-director of the Intelligent Process monitoring & management (IPMM) centre. He is a chartered engineer and a member of the Institute of Measurement and Control.

Paul W. Prickett

Educated at Cardiff University obtaining a BEng degree in Mechanical Engineering (1979). He is a senior lecturer in Cardiff School of Engineering and is also a co-director of the IPMM centre (established in 1998). He has directed the work of sixteen researchers and has supervised seven PhD students and two EngD students. He has published over 70 papers on the topic of machine and process condition monitoring. He is presented at numerous conferences and acts as a referee for journals and funding bodies in his field of research. He is a chartered engineer and a member of the Institution of Mechanical Engineers.

A new method of bearing fault diagnostics in complex rotating machines using multi-sensor mixtured hidden Markov models

Z. S. Chen¹, Y. M. Yang¹, Z. Hu¹, Z. X. Ge¹

¹ *Institute of Mechatronic Engineering, College of Mechatronic Engineering and Automation, National University of Defense Technology, Changsha, Hunan, P. R. China, 410073*

czs_study@sina.com

yangyongmin@yahoo.com

ABSTRACT

Vibration signals from complex rotating machines are often non-Gaussian and non-stationary, so it is difficult to accurately detect faults of a bearing inside using a single sensor. This paper introduces a new bearing fault diagnostics scheme in complex rotating machines using multi-sensor mixtured hidden Markov model (MSMHMM) of vibration signals. Vibration signals of each sensor will be considered as the mixture of non-Gaussian sources, which can depict non-Gaussian observation sequences well. Then its parameter learning procedure is given in detail based on EM algorithm. In the end the new method was tested with experimental data collected from a helicopter gearbox and the results are very exciting.

1. INTRODUCTION

Today's industry uses increasingly complex rotating machines, some with extremely demanding performance criteria. Machine failures are significantly contributed to both safety incidents and maintenance costs. The root cause of faults in complex rotating machines is often faulty bearings. A bearing condition monitoring system is therefore necessary to prevent major breakdowns due to progression of undetected faults. Over the past tens years, much research has been focused on vibration-based fault diagnostics techniques (Paul and Darryll, 2005). For complex rotating machines, however, it is still difficult to achieve a high degree of accuracy in classifying faults of a bearing inside due to the complexity of vibration signals.

Hidden Markov Model (HMM) has been a dominant method in speech recognition since 1960s and becomes very popular in the late 1980s and 1990s (Rabiner, 1989). The structure of HMM is useful for modeling a sequence that has a hidden stochastic process. It has become popular in various areas like signal analysis and pattern recognition, such as speech processing and

medical diagnostics. Recently, HMMs have been introduced into mechanical diagnostic areas and many HMMs were proposed and extended successfully for mechanical systems monitoring and diagnostics (Baruah and Chinnam, 2005; Leea, et al., 2004; Bunks, et al., 2000). In practice, it is an important issue how to select an appropriate HMM model. Most existing HMM-based fault diagnostic methods mainly assume that each state generates observations according to a Gaussian or Gaussian mixture model (Baruah and Chinnam, 2005; Leea, et al., 2004; Bunks, et al., 2000; Wang, et al., 2009). Also these methods often use a single sensor system to perform condition monitoring and diagnostics. Whereas vibration signals of complex rotating machines are often known to be highly non-Gaussian and non-stationary (Bouillaut and Sidahmed, 2001), such as a helicopter gearbox. Thus classical HMMs with Gaussian or Gaussian-mixtured observations have serious limitations for bearing fault diagnostics in complex rotating machines.

Obviously, a multi-sensor fault diagnostic system can overcome the limitations of a single sensor system and has improved performance. So our motivation is to build a novel HMM with non-Gaussian observations based on multi-sensor signals and then use it for bearing fault diagnostics in complex rotating machines. Vibration signals from a sensor on complex rotating machines can be looked as emanating from a number of sources caused by these components within it. This naturally fits an independent component analysis (ICA) process (Lee, et al., 2000). By this way, this paper will present a multi-sensor mixtured hidden Markov model (MSMHMM) for bearing fault diagnostics, which is improved on classical HMMs with mixtured non-Gaussian observation models.

2. DEFINITION OF MSMHMM

For a Gaussian observation model, the observation O_t at time t is assumed to be generated from a Gaussian

process, which is a scalar value corresponding to a single sensor. While for a multi-sensor system with N sensors, the observation \mathbf{O}_t at time t will be a vector, i.e. $\mathbf{O}_t = [x_{t1}, x_{t2}, \dots, x_{tN}]^T$. As mentioned before, signals from each sensor on a helicopter gearbox can be considered to be mixed by M sources caused by its inner components. In this paper a linear mixing process is considered. Denoting \mathbf{W}_k as the mixing matrix at state k and $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_M]^T$ as M sources, the observation vector at time t for state k can be calculated according to an independent component analysis process as follows,

$$\mathbf{O}_t^k = \mathbf{W}_k \mathbf{S}_t \quad (1)$$

Where \mathbf{W}_k is the $N \times M$ mixing matrix, $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_M$ are statistically independent. For the sake of simplicity, we only consider $N = M$ in this paper and \mathbf{W}_k is a square matrix. Then we have

$$\mathbf{S}_t = \mathbf{V}_k \mathbf{O}_t^k \quad (2)$$

Where \mathbf{V}_k is called as the unmixing matrix and $\mathbf{V}_k = (\mathbf{W}_k)^{-1}$.

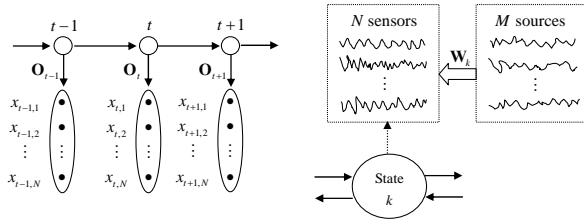


Figure 1: A graphical MSMHMM

A standard MSMHMM is shown as a graphical model in Fig. 1. Then based on the maximum likelihood framework of an independent component analysis process, the multivariate probability of the multi-sensor observation vector can be calculated from the source densities as follows (W. D. Penny, 1998),

$$P(\mathbf{O}_t^k) = \frac{P(\mathbf{S}_t)}{|J_k|} \quad (3)$$

Where $|J_k| = \det(\mathbf{W}_k) = 1/\det(\mathbf{V}_k)$, $P(\mathbf{S}_t) = \prod_{i=1}^M P(s_{it}^k)$ and

$$s_{it}^k = \sum_{j=1}^M \mathbf{V}_{kij} x_{tj}.$$

Then Eq.(3) can be transformed as

$$\log P(\mathbf{O}_t^k) = \log \frac{P(\mathbf{S}_t)}{|J_k|} = -\log |\det(\mathbf{W}_k)| + \sum_{i=1}^M \log P(s_{it}^k) \quad (4)$$

It can be easily seen from Eq.(4) that the probability density of each observation sequence is determined by the probability density of source components. Thus in practice, we should choose proper non-Gaussian source

density models to represent non-Gaussian observation sequence, such as vibration signals of helicopter gearboxes. Assuming that non-Gaussian source density model at state k is depicted by the parameter set $\{\theta_k\}$, a multi-sensor mixed hidden Markov model can be built by the complete parameter set as follows,

$$\lambda_{\text{MSMHMM}} = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{W}, \boldsymbol{\theta}) \quad (5)$$

Next we need to train the MSMHMM before using it, which refers to the estimation of parameters: $\boldsymbol{\pi}$, \mathbf{A} , \mathbf{W} and $\boldsymbol{\theta}$.

3. MSMHMM PARAMETERS LEARNING BASED ON EM ALGORITHM

Actually a MSMHMM is improved on a standard HMM, so its parameters learning frame is similar to that of a standard HMM. Thus expectation maximization (EM) algorithm can also be used for MSMHMM parameters learning. That is to say, it needs to maximize, $E(\lambda_{\text{MSMHMM}}, \hat{\lambda}_{\text{MSMHMM}})$, the expectation of the joint log likelihood of an observation sequence $\mathbf{O} = [\mathbf{O}_1, \mathbf{O}_2, \dots, \mathbf{O}_T]$ and hidden state sequence \mathbf{Q} . Here (W. D. Penny, 1998),

$$\begin{aligned} E(\lambda_{\text{MSMHMM}}, \hat{\lambda}_{\text{MSMHMM}}) &= \sum_{\mathbf{Q}} P_{\lambda_{\text{MSMHMM}}}(\mathbf{Q} | \mathbf{O}) \log P_{\lambda_{\text{MSMHMM}}}(q_1) \\ &+ \sum_{\mathbf{Q}} P_{\lambda_{\text{MSMHMM}}}(\mathbf{Q} | \mathbf{O}) \sum_{t=2}^T \log P_{\lambda_{\text{MSMHMM}}}(q_t | q_{t-1}) \\ &+ \sum_{\mathbf{Q}} P_{\lambda_{\text{MSMHMM}}}(\mathbf{Q} | \mathbf{O}) \sum_{t=1}^T \log P_{\lambda_{\text{MSMHMM}}}(\mathbf{O}_t | q_t) \end{aligned} \quad (6)$$

Obviously, Eq.(6) composes of three terms which can be used to train MSMHMM model parameters respectively: the first term for the initial state probabilities ($\boldsymbol{\pi}$), the second term for the state transition probabilities (\mathbf{A}) and the third one for the observation model parameters, i.e. the mixing matrix (\mathbf{W}) and source density parameters ($\boldsymbol{\theta}$).

3.1 Initial state probabilities learning ($\boldsymbol{\pi}$)

The initial state probabilities π_i can be updated by maximizing the first term, $\sum_{\mathbf{Q}} P_{\lambda_{\text{MSMHMM}}}(\mathbf{Q} | \mathbf{O}) \log P_{\lambda_{\text{MSMHMM}}}(q_1)$. Furthermore we have,

$$\begin{aligned} \sum_{\mathbf{Q}} P_{\lambda_{\text{MSMHMM}}}(\mathbf{Q} | \mathbf{O}) \log P_{\lambda_{\text{MSMHMM}}}(q_1) &= \sum_{i=1}^K P_{\lambda_{\text{MSMHMM}}}(q_1 = i | \mathbf{O}) \log P_{\lambda_{\text{MSMHMM}}}(i) \\ &= \sum_{i=1}^K \gamma_1(i) \log \hat{\pi}_i \end{aligned} \quad (7)$$

Where the constraints are as follows:

$$\sum_{i=1}^K \hat{\pi}_i = 1, \sum_{i=1}^K \gamma_1(i) = 1.$$

By maximizing Eq.(7), we can get the final update formula as

$$\hat{\pi}_i = \gamma_1(i) \quad (8)$$

Where $\gamma_1(i)$ can be calculated using the forward-backward algorithm.

3.2 State transition probabilities learning (A)

The state transition probabilities \mathbf{A} can be updated by maximizing the second term,

$\sum_{\mathbf{Q}} P_{\lambda_{\text{MSMHMM}}}(\mathbf{Q} | \mathbf{O}) \sum_{t=2}^T \log P_{\lambda_{\text{MSMHMM}}}(q_t | q_{t-1})$. Furthermore we have,

$$\begin{aligned} & \sum_{\mathbf{Q}} P_{\lambda_{\text{MSMHMM}}}(\mathbf{Q} | \mathbf{O}) \sum_{t=2}^T \log P_{\lambda_{\text{MSMHMM}}}(q_t | q_{t-1}) \\ &= \sum_{i=1}^K \sum_{j=1}^K \sum_{t=2}^T P_{\lambda_{\text{MSMHMM}}}(q_t = j, q_{t-1} = i | \mathbf{O}) \log P_{\lambda_{\text{MSMHMM}}}(q_t | q_{t-1}) \\ &= \sum_{i=1}^K \sum_{j=1}^K \left(\sum_{t=1}^{T-1} \xi_t(i, j) \right) \log(a_{ij}) \end{aligned} \quad (9)$$

$$\text{where } \xi_t(i, j) = \frac{\alpha_t(i) a_{ij} b_j(\mathbf{O}_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^K \sum_{j=1}^K \alpha_t(i) a_{ij} b_j(\mathbf{O}_{t+1}) \beta_{t+1}(j)}$$

By maximizing Eq.(9), we can get the final update formula as

$$\hat{a}_{ij} = \frac{\sum_{t=1}^T \xi_t(i, j)}{\sum_{i=1}^K \gamma_t(i)} \quad (10)$$

3.3 Mixing matrix (W) and source density parameters (θ) learning

The observation model parameters, i.e. the mixing matrix (\mathbf{W}) and source density parameters (θ), can be updated by maximizing the third term,

$\sum_{\mathbf{Q}} P_{\lambda_{\text{MSMHMM}}}(\mathbf{Q} | \mathbf{O}) \sum_{t=1}^T \log P_{\lambda_{\text{MSMHMM}}}(\mathbf{O}_t | q_t)$. However,

the update process is determined by the observation model. In this paper in order to represent non-Gaussian vibration signals of a helicopter gearbox, we need to choose proper non-Gaussian source models in MSMHMM. (S. J. Roberts, 1998) has pointed out that a signal consisting of multiple sinusoids has a multimodal probability density function (PDF) and

generalized autoregressive (GAR) source models can provide better unmixing than generalized exponential (GE) source models for multimodal PDFs sources. On the other hand, as we all know that a rotating machine works under periodic motions and its vibration source are often multi-frequencies sinusoids, so GAR source models will be used in this paper.

A GAR source model is shown as follows,

$$s_{it}^k = -\sum_{d=1}^p c_i^k [d] s_{i(t-d)}^k + e_{it}^k, 1 \leq i \leq N, 1 \leq t \leq T, 1 \leq k \leq K \quad (11)$$

Where $c_i^k[\cdot]$ are the GAR coefficients for the i th source at state k and denoted as \mathbf{c}_i^k , e_{it}^k is a non-Gaussian additive noise and p is the model order. In practice, e_{it}^k denotes the GAR prediction error and can be calculated as,

$$\begin{aligned} e_{it}^k &= s_{it}^k - \hat{s}_{it}^k \\ \hat{s}_{it}^k &= -\sum_{d=1}^p c_i^k [d] \hat{s}_{i(t-d)}^k \end{aligned} \quad (12)$$

Then each GAR source density at state k is (S. J. Roberts, 1998)

$$P(s_{it}^k) = \frac{R_i^k (\beta_i^k)^{1/R_i^k}}{2\Gamma(1/R_i^k)} \exp(-\beta_i^k |e_{it}^k|)^{R_i^k} \quad (13)$$

Where $\Gamma(\cdot)$ is the gamma function, R_i^k , β_i^k are the two density parameter for i th source at state k .

So in this paper source density parameters (θ) composes of $\{p, \mathbf{c}_i^k, R_i^k, \beta_i^k\}$. Next we will train these parameters according to the third term in Eq.(6). Furthermore we have,

$$\begin{aligned} & \sum_{\mathbf{Q}} P_{\lambda_{\text{MSMHMM}}}(\mathbf{Q} | \mathbf{O}) \sum_{t=1}^T \log P_{\lambda_{\text{MSMHMM}}}(\mathbf{O}_t | q_t) \\ &= \sum_{k=1}^K \sum_{t=1}^T P_{\lambda_{\text{MSMHMM}}}(q_t = k | \mathbf{O}) \log P_{\lambda_{\text{MSMHMM}}}(\mathbf{O}_t | k) \end{aligned} \quad (14)$$

$$= \sum_{k=1}^K \sum_{t=1}^T \gamma_t[k] \log P_{\lambda_{\text{MSMHMM}}}(\mathbf{O}_t | k)$$

Where $\log P_{\lambda_{\text{MSMHMM}}}(\mathbf{O}_t | k)$ can be calculated by Eq.(4). That is,

$$\begin{aligned} \log P_{\lambda_{\text{MSMHMM}}}(\mathbf{O}_t | k) &= \log P_{\lambda_{\text{MSMHMM}}}(\mathbf{O}_t^k) \\ &= -\log |\det(\mathbf{W}_k)| + \sum_{i=1}^N \log P(s_{it}^k) \end{aligned} \quad (15)$$

By substituting Eq.(12), (13), (15) into Eq.(14), updating of $\{\mathbf{W}, \mathbf{c}_i^k, R_i^k, \beta_i^k\}$ can be derived by

differentiating Eq.(14) on W_{ij}^k , \mathbf{c}_i^k , R_i^k , β_i^k respectively.

Besides of $\boldsymbol{\pi}$, \mathbf{A} , \mathbf{W} and $\boldsymbol{\theta}$, there are some other parameters needed to be determined, including the number of sources, N, the number of states, K, and the order of GAR, p. How to select these parameters is a problem to be solved, which will not be discussed deeply in this paper.

By now, the algorithm of MSMHMM parameters learning can be implemented by Matlab software.

4. A CASE OF BEARING FAULT DIAGNOSTICS IN A HELICOPTER GEARBOX

In the experiment, a bearing in a helicopter gearbox is selected and two classical faults are seeded on it, i.e. rolling element fault and outer race fault, shown in Fig. 2. Then vibration signals are collected from five sensors under normal and faulty conditions respectively. The sampling frequency is 10 KHz at each channel.

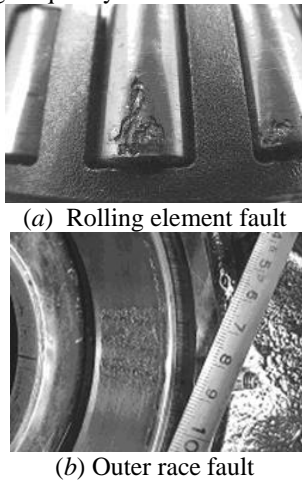


Figure 2: Two kinds of faults on the bearing

4.1 MSMHMMs training

In the scheme of MSMHMM-based fault diagnostics, firstly it needs to determine the number of sources, the number of states and the order of GAR. Because the gearbox consists of five main components in this paper, the number of sources is selected as N=5 here. Then five vibration sensors for observations are mounted on the gearbox. The number of states is selected as K=4 and the order of GAR is selected as p=6 artificially. The length of observation sequence is selected as T=512.

By initializing initial probabilities, $\boldsymbol{\pi}_{K \times 1}$, transition matrix, $\mathbf{A}_{K \times K}$, mixing matrix, $\mathbf{W}_{K \times N \times N}$, source density parameters, $\mathbf{c}_{K \times N \times p}$, $R_{K \times N}$, $\beta_{K \times N}$, different MSMHMMs under three conditions are trained based

on 10 training samples respectively. After training, we can get three MSMHMMs (MSMHMM1 for normal, MSMHMM2 for rolling element fault and MSMHMM3 for outer race fault) and the corresponding state sequences are shown in Fig. 3.

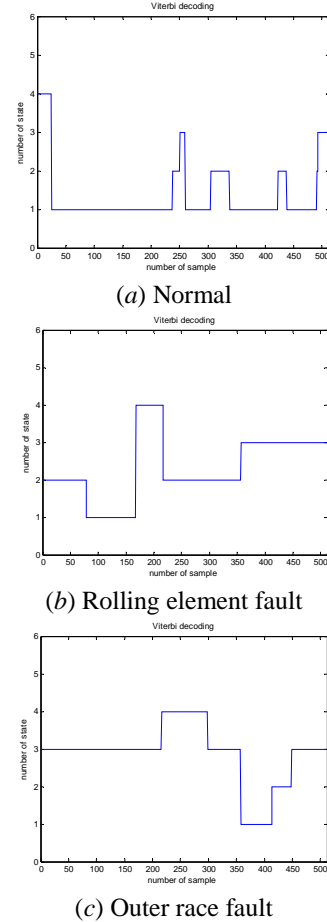


Figure 3: State sequences for different bearing conditions after training

4.2 MSMHMMs-based bearing faults identification

After three MSMHMMs has been built and trained, we can use them to isolate different conditions using testing samples. The number of testing samples under each condition is selected as 15. Then each MSMHMM is used to analyze normal, rolling element fault and outer race fault samples to test its classification ability respectively, and then the corresponding results are shown as Fig. 4~Fig. 6. In Fig. 4, MSMHMM₁ is used and the maximum log-likelihood corresponds to normal condition, so MSMHMM₁ identify health condition of the bearing accurately. Similar results can be obtained in Fig. 5 and Fig. 6. Thus it demonstrates that MSMHMMs can identify faults in the helicopter gearbox accurately.

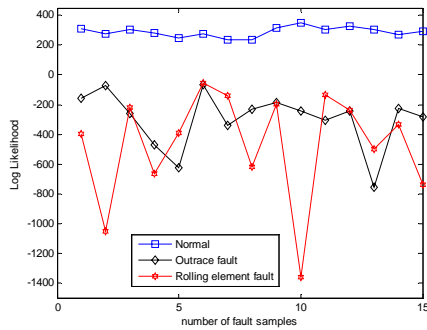


Figure 4: Identified results based on MSMHMM₁

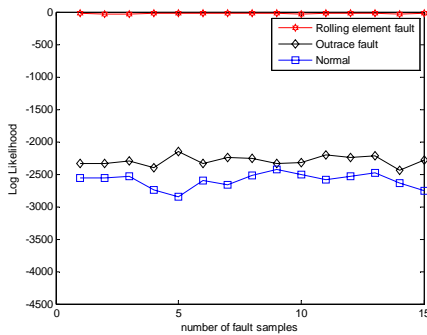


Figure 5: Identified results based on MSMHMM₂

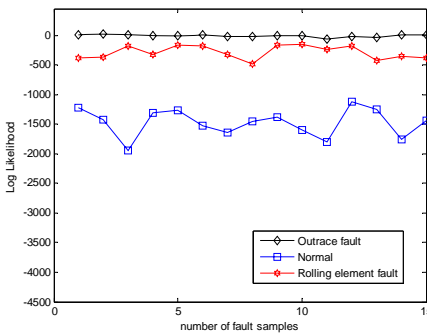


Figure 6: Identified results based on MSMHMM₃

In order to testify that Gaussian observation HMM (GHMM) may not fit for bearing fault diagnostics in the helicopter gearbox, we will use the above training samples to build and train three GHMMs (GHMM₁ for normal, GHMM₂ for rolling element fault and GHMM₃ for out race fault), where the number of states is also selected as $K=4$. Then three GHMMs are used to analyze normal, rolling element fault and out race fault testing samples, and then the corresponding results are shown as Fig. 7~Fig. 9 respectively. It can be seen that GHMMs cannot identify the anticipated condition and provide mistaken results. The reason may be that observation sequences from the helicopter gearbox are truly non-Gaussian and non-stationary. Also we can find the log-likelihood values in Fig. 7~Fig. 9 fluctuate more than those in Fig. 4~Fig. 6. The

reason may be that the observation sequences are non-stationary. Thus it testifies that the proposed MSMHMM is a better tool than traditional GHMM for bearing fault diagnostics.

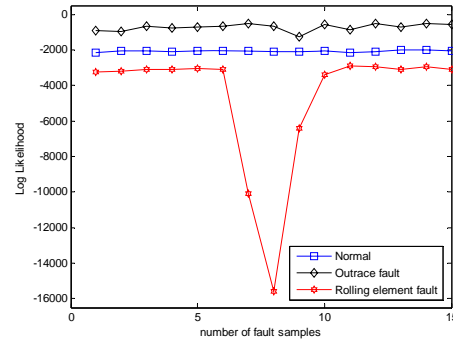


Figure 7: Identified results based on GHMM₁

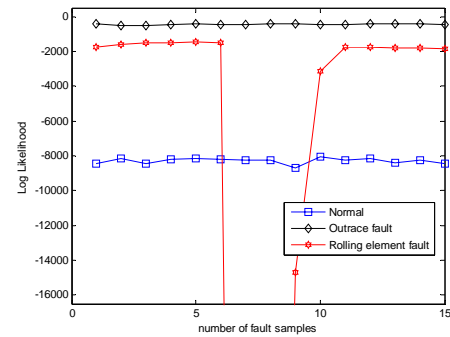


Figure 8: Identified results based on GHMM₂

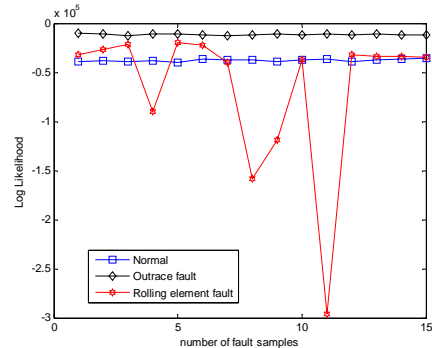


Figure 9: Identified results based on GHMM₃

5. CONCLUSION

This paper has presented a MSMHMM-based bearing fault diagnostics method for complex rotating machines using multi-sensor observation signals. Each sensor signals was considered as the mixture of non-Gaussian sources, which can depict non-Gaussian observation sequences well. Then its parameter learning algorithm was proposed based on EM algorithm. In the end through the experimental study on a bearing in a helicopter gearbox, we have testified that MSMHMMs can identify bearing faults more accurately than

traditional GHMMs. Furthermore, the proposed MSMHMMs can be extended for fault diagnostics of other complex rotating machines.

Future work will include how to determine the number of states and the order of GAR models in MSMHMMs theoretically, which may be solved by understanding particular mechanical systems and their working processes.

ACKNOWLEDGMENT

The authors sincerely appreciate the funding provided to this research by the National Nature Science Foundation of China under grant number 50805142. Also the authors would like to thank Dr. W. D. Penny for his help and contribution to this work.

REFERENCES

- D. S. Paul, J. P. Darryll (2005). A review of vibration-based techniques for helicopter transmission diagnostics, *Journal of Sound and Vibration*, vol. 282, pp. 475–508.
- L.R. Rabiner (1989). A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of IEEE*, vol. 77, pp. 257–285.
- P. Baruah, R. B. Chinnam (2005). HMMs for diagnostics and prognostics in machining processes, *International Journal of Production Research*, vol. 43, pp. 1275–1293.
- J. M. Leea, S. J. Kima, Y. Hwang, et al.(2004). Diagnosis of mechanical fault signals using continuous hidden Markov model, *Journal of Sound and Vibration*, vol. 276, pp.1065–1080.
- C. Bunks, D. McCathy, T. Al-Ani (2000). Condition-based maintenance of machines using hidden Markov models, *Mechanical Systems and Signal Processing*, vol. 14, pp. 597–612.
- F. G. Wang, Y. B. Li , Z. G. Luo (2009). Fault classification of rolling bearing based on reconstructed phase space and Gaussian mixture model, *Journal of Sound and Vibration*, vol. 323, pp. 1077–1089.
- L. Bouillaut, M. Sidahmed (2001). Helicopter gearbox vibrations: cyclostationary analysis or bilinear approach, *International symposium on signal processing and its application*, pp. 13–16.
- T-W. Lee, M. Girolami, A.J. Bell, et al.(2000). A unifying information-theoretic framework for independent component analysis, *Computers & Mathematics with Applications*, vol. 31, pp.1–21.
- W. D. Penny, S. J. Roberts (1998). Hidden Markov models with extended observation densities, *Technical Report TR-98-15* .
- R. Everson, S. J. Roberts (1998). Independent Component Analysis: A flexible non-linearity and decorrelating manifold approach. *Proceedings of IEEE conference on neural network and signal processing*, pp. 33–42.

Z. S. Chen was born in Anhui Province, China, on Aug. 13, 1977. He received the B.E. and Ph.D. degrees in Mechatronic Engineering from the National University of Defense Technology, P. R. China in 1999 and 2004, respectively. From 2008, he worked as an associate professor in Institute of Mechatronic Engineering, National University of Defense Technology. Current research interests include condition monitoring and fault diagnosis, mechanical signal processing and data fusion, vibration energy harvesting, etc.



A Prognostic Health Management Based Framework for Fault-Tolerant Control

Douglas W. Brown¹, and George J. Vachtsevanos²

^{1,2} *Georgia Institute of Technology, Atlanta, GA 30332, USA*

dbrown31@gatech.edu

gfv@ece.gatech.edu

ABSTRACT

This paper presents one approach in developing a PHM-based reconfigurable controls framework. A low-level reconfigurable controller is defined as a time-varying multi-objective criterion function and appropriate constraints to determine optimal set-point reconfiguration. A set of necessary conditions are established to ensure the stability and boundedness of the composite system. In addition, the error bounds corresponding to long-term state-space prediction are examined. From these error bounds, the point estimate and corresponding uncertainty boundaries for the remaining useful life (RUL) estimate are obtained. Finally, results are obtained for an avionics grade triplex-redundant electro-mechanical actuator (EMA) with a specific fault mode; insulation breakdown between winding turns in a brushless DC (BLDC) motor is used as a test case for the fault-mode.

1 INTRODUCTION

The emergence of complex and autonomous systems, such as modern aircraft, unmanned aerial vehicles (UAVs) and automated industrial processes is driving the development and implementation of new control technologies aimed at accommodating incipient failures to maintain system operation during an emergency. A prognostics health management (PHM) based fault-tolerant control architecture can increase safety and reliability by detecting and accommodating impending failures thereby minimizing the occurrence of unexpected, costly and possibly life-threatening mission failures; reduce unnecessary maintenance actions; and extend system availability / reliability.

The primary motivation for this research topic emerged over

Douglas W. Brown and George J. Vachtsevanos. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

the need for improved reliability and performance for safety critical systems, particularly in aerospace related applications. Fatal accidents in the worldwide commercial jet fleet during the years 1987-2005 were due primarily to (i) controlled flight into terrain, (ii) loss-of-control in flight and (iii) system/component failure or malfunction (Darby, 2006). In a coordinated effort to improve aviation safety, industry and government worked together to reduce the number of fatal commercial aircraft accidents, which dropped by 65% during the period of 1996-2007 (Wald, 2007). As a result of this effort, accidents due to controlled flight into terrain have been virtually eliminated through the addition of various safeguards, but the same cannot be said for accidents due to loss-of-control in flight and system/component failure or malfunctions. System/component failure and malfunctions are recognized as contributing factors to aircraft loss-of-control in flight, so safeguarding against such events will reduce the number of fatal accidents in the two top accident categories (ii) and (iii) respectively.

The remainder of this document is organized as follows. Section 2 presents a literature review for fault detection and diagnosis, long-term prognosis predictions and fault tolerant control strategies. Section 3 defines the FTC architecture. Section 4 studies the stability and boundedness of the reconfigured system and the RUL prediction. Section 5 provides general design guidelines and demonstrates the reconfigurable control algorithms on an EMA. Finally, Section 6 summarizes the findings and future work.

2 LITERATURE REVIEW

According to the NASA ASP IVHM program, the following enabling technologies are necessary before prognosis based control can be considered: fault detection, fault diagnosis and failure prognosis (Srivastava, Mah, & Meyer, 2008). The section concludes with a brief overview of FTC strategies.

2.1 Fault Detection and Diagnosis (FDD)

Over the past three decades, the growing demand for reliability, maintainability, and survivability in dynamic systems has drawn significant research in FDD. Historically, FDD has been used in FTC to retrieve fault information from the system for use in a control recovery strategy and procedure, which is commonly referred to as reconfiguration. Preliminary research by Jiang & Patterson (Jiang & Zhao, 1997; Patterson, 1997) demonstrated that state estimation based schemes are most suitable for fault detection since they are inherently fast and cause a very short time delay in real-time decision making. However, the information from state estimation based algorithms may not be detailed enough for subsequent control system reconfiguration. Work presented by Wu and Zhang (N. E. Wu, Zhang, & Zhou, 2000; Y. M. Zhang & Jiang, 2002) recommends that parameter estimation schemes be used for control reconfiguration and state estimation based schemes for FDD. A unified approach to state estimation/prediction and parameter estimation/identification for FDD using particle filtering was thoroughly studied by M. Orchard (Orchard, 2007).

2.2 Failure Prognosis & Long-Term Prediction

The term prognosis has been used widely in medical practice to imply the foretelling of the probable course of a disease. In the industrial and manufacturing fields, prognosis is interpreted to answer the question, "What is the RUL of a machine or component once an impending failure condition is detected, isolated, and identified?" Within the context of this work, prognosis is defined as (Vachtsevanos, Lewis, Roemer, Hess, & Wu, 2006),

Definition 1 (Prognosis). The ability to predict accurately the RUL of a failing component or subsystem.

Definitions for failure, probability of failure and RUL must be well established before continuing the discussion on prognosis. First, the notion of a failure is defined.

Definition 2 (Failure). An event that corresponds to the fault-dimension, L , entering an unwanted range, or hazard-zone. The hazard-zone is defined by the upper and lower bounds, H_{ub} and H_{lb} , respectively.

The boundaries of the hazard zone are design parameters related to the false-alarm rate (type I error). It should be recognized any discussion regarding a failure over a future time horizon $t > t_0$ is stochastic in nature. Instead, the probability of failure should be used.

Definition 3 (Probability of Failure). The probability of a failure occurring at some time t , represented as,

$$p_{\text{failure}}(t) = p(H_{lb} \leq L(t) \leq H_{ub}), \quad (1)$$

where p is a probability density function (pdf).

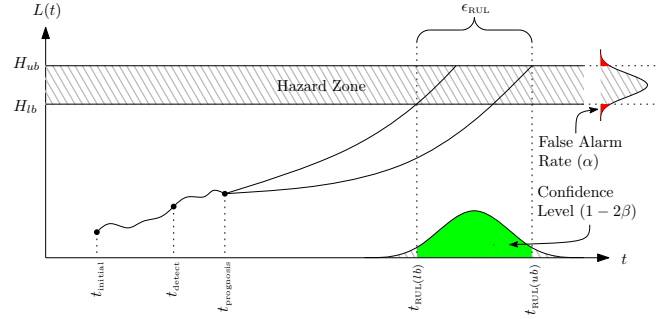


Figure 1. Predicted fault growth curves, hazard zone and corresponding projection on the time-axis.

Finally, its often convenient to describe the minimum time-horizon (or RUL) corresponding to a failure with a particular level of certainty, represented by the symbol $t_{\text{RUL}(lb)}$.

Definition 4 (Remaining Useful Life (RUL)). The amount of time before a failure occurs at the initial time of prediction, t_0 . The time corresponding to the probability of failure can be expressed as,

$$t_{\text{RUL}(lb)}(t_0) \triangleq \min(t^*) \quad \text{s.t.} \quad p_{\text{failure}}(t^*|t_0) \geq \beta, \quad (2)$$

where $t^* \in (t_0, \infty)$ and $0 < \beta < 1$. The symbols t_0 and β refer to the initial prediction time and the type-II error associated with the prediction accordingly.

Sometimes the term confidence level is used instead of the type-II error, which is defined next.

Definition 5 (Confidence Level (CL)). Let the upper RUL boundary, $t_{\text{RUL}(ub)}$, predicted at time t_0 be defined as,

$$t_{\text{RUL}(ub)}(t_0) \triangleq \min(t^*) \quad \text{s.t.} \quad p_{\text{failure}}(t^*|t_0) \leq 1 - \beta. \quad (3)$$

where $t^* \in (t_0, \infty)$. Then the CL is defined by the following probability,

$$\text{CL} = \int_{t_{\text{RUL}(lb)}}^{t_{\text{RUL}(ub)}} p_{\text{failure}}(t^*|t_0) dt^*. \quad (4)$$

Additionally, CL is related to β by,

$$\text{CL} = 1 - 2\beta. \quad (5)$$

Figure 1 illustrates the predicted fault growth of a system where a fault is detected at time t_{detect} and a prediction of the RUL is made at time $t_{\text{prognosis}}$. The boundaries of the hazard-zone are defined by H_{lb} and H_{ub} . The probability that a failure occurs outside this boundary is defined as the false-alarm rate, α . The time corresponding to each predicted fault trajectory in the hazard-zone is represented as a distribution on the time-axis. The upper and lower RUL boundary values that encompass a CL of $1 - 2\beta$ are represented as $t_{\text{RUL}(ub)}$ and $t_{\text{RUL}(lb)}$, accordingly. The width of the corresponding confidence interval is defined as,

$$\epsilon_{\text{RUL}} \triangleq t_{\text{RUL}(ub)} - t_{\text{RUL}(lb)}. \quad (6)$$

Several approaches to prognosis have been investigated in recent years, such as model-based (Yu & Harris, 2001; Paris & Erodogan, 1963), data-driven (Schwabacher, 2005), hybrid methods and particle filtering (Orchard, 2007; Orchard, Kacprzynski, Goebel, Saha, & Vachrsevanos, 2009).

2.3 Fault-Tolerant Control (FTC) Strategies

Modern systems rely on sophisticated controllers to meet increased performance and safety requirements. A conventional feedback control design for a complex system may result in unsatisfactory performance, or even instability, in the event of malfunctions in actuators, sensors or other system components. To overcome such weaknesses, new approaches to control system design have been developed in order to tolerate component malfunctions while maintaining desirable stability and performance properties. According to Y. Zhang (Y. Zhang & Jiang, 2003), FTC is defined as,

Definition 6 (Fault-Tolerant Control (FTC) Systems). Control systems that possess the ability to accommodate system component failures automatically [while] maintaining overall system stability and acceptable performance.

Traditionally, FTC systems are classified into two categories: passive and active (Y. Zhang & Jiang, 2008).

2.3.1 Passive Fault-Tolerant Control Systems (PFTCS)

Historically, when fault tolerance was an issue, controllers were designed targeting selected faults with specific control actions to mitigate the risk of impending failures (Isermann, 1984). Within such passive approaches, no fault information is required and robust control techniques are employed to ensure the closed-loop system remains insensitive to specific anticipated faults (Zhenyu & Hicks, 2006). The most common and widely studied PFTCS is robust control. Although PFTCS are widely used, they lack an active reconfiguration of the control law thus disallowing use of any external information such as FDD and prognostics.

2.3.2 Active Fault-Tolerant Control Systems (AFTCS)

AFTCS react to system component failures by reconfiguring control actions to maintain stability and acceptable system performance. AFTCS FTC methodologies typically have two main objectives: FDD and control reconfiguration (Rausch, Goebel, Eklund, & Brunell, 2007). Several authors have reported on the problem of FDD (Filippetti, Franceschini, Tassoni, & Vas, 2000; Kleer & Williams, 1987). In such control systems, the controller compensates for the effects of faults either by selecting a pre-computed control law or by synthesizing a new control scheme on-line (Skormin, Apone, & Dunphy, 1994; Willsky, 1976). An AFTCS consist of a reconfigurable controller, a FDD scheme and a reconfiguration mechanism (B. Wu, Abhinav, Khawaja, & Panagiotis, 2004). Types of AFTCS include adaptive robust control

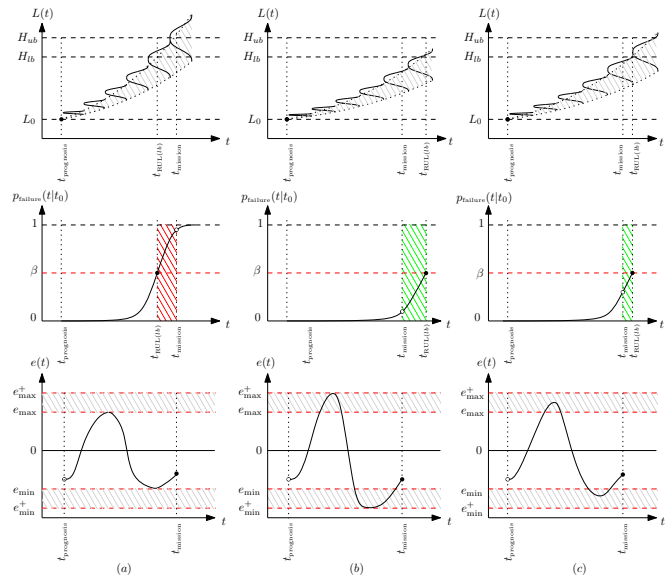


Figure 2. Conceptual plots for the fault dimension, L , probability of failure predicted at time $t_{prediction}$ and the tracking error, e , versus time for three different reconfiguration scenarios: (a) $t_{RUL(lb)} < t_{mission}$, (b) $t_{RUL(lb)} > t_{mission}$ but the control is overcompensated leading to an unnecessary increase in tracking error and (c) $t_{RUL(lb)} > t_{mission}$ and the tracking error is minimized.

(Saberi, Stoorvogel, Sannuti, & Niemann, 2000), expert control (Isermann, 1997; Levis, 1987; N. E. Wu, 1997), optimal control (Bogdanov, Chiu, Gokdere, & Vian, 2006; Garcia, Prett, & Morari, 1989; Kwon, Bruckstein, & Kailath, 1983), and hybrid control. Particular interest is the work by (Bogdanov et al., 2006) and (Monaco, D.G., & Bateman, 2004) which introduces prognostic information into a control law using model predictive control. For systems where on-line computation is feasible, MPC has proved quite successful (Richalet, 1993; Richalet, Rault, Testud, & Papon, 1978). Monaco *et al.* (Monaco et al., 2004) demonstrated an MPC based framework used to retrofit the F/A-18 fleet support flight control computer (FSFCC) with an adaptive flight controller.

3 CONTROL ARCHITECTURE

The problem of incorporating prognosis in a control system can be approached in a variety of ways. The efficacy of any one approach depends on the problem formulation and the specific application. Therefore, fixed performance criteria are necessary to compare any two designs. In the scope of this work, the controller performance criteria are determined by the ability to prevent a failure while minimizing the impact on overall system performance over a well-defined time horizon.

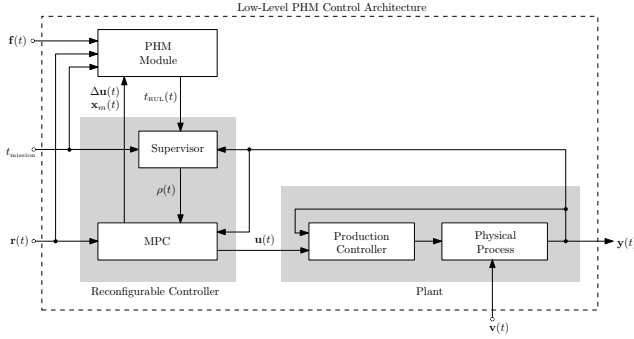


Figure 3. Reconfigurable controller illustrating the internal MPC controller and supervisor elements in addition to the external PHM module and connected plant.

Let these criteria be evaluated by the cost function, Θ ,

$$\Theta(\Phi, \Psi) = \Phi(t_f) + \int_{t_0}^{t_f} \Psi(e(t)) dt, \quad (7)$$

where the terminal cost is defined as,

$$\Phi = \begin{cases} \infty & : p_{\text{failure}}(t_f) \geq \beta, \\ 0 & : \text{otherwise.} \end{cases} \quad (8)$$

The mappings Φ and Ψ represent the cost associated with the performance and the final damage. The symbols L and e refer to the fault dimension, and tracking error, accordingly. In the scope of this work, the tracking error is defined as,

$$e \triangleq \mathbf{u} - \mathbf{y} \quad (9)$$

3.1 Qualitative Example

Consider the plots in Figure 2 for the fault dimension, L , probability of failure, p_{failure} and tracking error, e , versus time for three different scenarios. The illustration of this example is simplified by considering a single-input single-output (SISO) case. Let the symbols e_{\min} and e_{\max} be constant boundaries for the tracking error and $t_{\text{RUL}(lb)}$ represent the lower confidence bound of the RUL. In scenario (a) the performance criteria is not relaxed and the RUL is not achieved. That is, the probability of failure exceeds β before time t_{mission} . In scenario (b) the performance criteria is relaxed, more specifically e_{\min} and e_{\max} are extended to e_{\min}^+ and e_{\max}^+ , to achieve the RUL. However, the performance criteria is relaxed by more than what is actually necessary. In scenario (c) the performance criteria is relaxed such that the RUL requirement is satisfied, but not as much as scenario (b).

3.2 Control Architecture

The main elements of the control architecture are depicted in Figure 3 on page 4. The control architecture is comprised of the plant (physical process and production controller), reconfigurable controller and a PHM module. Initially, the

production controller is utilized with no modification while the PHM module continuously monitors the system for one (or more) fault mode(s). Once a fault is detected, the RUL is evaluated by the PHM module. If the estimated RUL is greater than the desired RUL, no action is taken. During this period the RUL is re-evaluated periodically. However, if the estimated RUL is less than the desired RUL, a reconfiguration action is triggered. The reconfigurable controller relaxes constraints on the error boundaries by adjusting the weight matrices in the MPC cost function. This continues until either the RUL is satisfied or the weight matrices can no longer be adapted. The remainder of this section presents a detailed description of each module.

3.2.1 Plant (Nominal System)

The plant consists of the production controller and physical process with a control input, \mathbf{u} , internal state, \mathbf{x} , measured disturbance, \mathbf{v} and output response \mathbf{y} . Prognosis based control can only be considered once it's established the RUL of the plant can be directly controlled and observed. As a result, two important questions arise, "Under what conditions can the RUL of the plant be controlled? ... observed?" These questions are answered by well defined criteria for RUL controllability and RUL observability given in Definitions 7 and 8, accordingly.

Definition 7 (RUL Controllability). A system is RUL controllable at time t_0 if there exists a control input, $\mathbf{u}(t) \in \mathcal{U}$ on the interval $t \in [t_0, t_f]$ such that any initial RUL $t_{\text{RUL}}(t_0)$ can be driven to any desired RUL value, $t_{\text{RUL}}(t_f) \in \mathcal{T}_{\text{RUL}}$.

Definition 8 (RUL Observability). A system is RUL observable at time t_0 if for any initial state in the state space $\mathbf{x}(t_0) \in \mathcal{X}$ and a given control input $\mathbf{u}(t) \in \mathcal{U}$ defined on the interval $t \in [t_0, t_f]$ the RUL, t_{RUL} , can be determined for $[t_0, t_f]$.

Remark 1. If the conditions for RUL controllability and observability are simultaneously satisfied, then the system is said to be RUL stabilizable

Definition 9 (RUL Stabilizable). A system is RUL stabilizable if for any initial state in the state space $\mathbf{x}(t_0) \in \mathcal{X}$ and any control input $\mathbf{u}(t) \in \mathcal{U}$ defined on the interval $t \in [t_0, t_f]$ the plant is simultaneously RUL controllable and RUL observable.

3.2.2 Reconfigurable Controller

The two elements of the reconfigurable controller include the low-level supervisor and the MPC controller.

Low-Level Supervisor – A logical unit used to continuously monitor the output of the MPC controller to ensure it meets the desired RUL and set-point requirements. More specifically, if the measured RUL, t_{RUL} , is greater than the mission time, t_{mission} , then no reconfiguration is necessary;

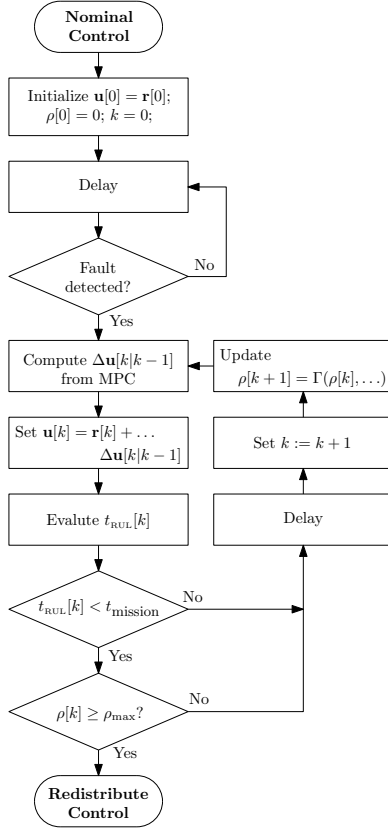


Figure 4. Flowchart of the low-level supervisor.

otherwise new acceptable minimum and maximum allowable tracking errors, e_{\min}^+ and e_{\max}^+ , are adopted. Then, the adjusted set-point, $\Delta \mathbf{u}$, and modeled state estimate, \mathbf{x}_m , are passed to the PHM module to estimate, t_{RUL} . This estimate is used as an input to the adaptation function, Γ , to update the adaptation parameter, ρ . The cost function is updated using a new value for ρ at time-instant k . When the estimated RUL, t_{RUL} , is less than the desired mission time, t_{mission} , the adaptation parameter, ρ , increases, otherwise it decreases. This process is re-iterated until $\rho \geq \rho_{\max}$. When this occurs, the controller makes no further adaptation attempts. An outline of this process is shown as a flowchart in Figure 4.

Model Predictive Controller (MPC) – used to make adjustments to the control signal, \mathbf{u} , thereby altering the internal states, \mathbf{x} , and causing the RUL to increase. In the scope of this work, constraints are imposed on the maximum allowable tracking error, \mathbf{e} . Foreshadowing briefly to the next chapter, it can be proven if \mathbf{e} is constrained by $\mathbf{e}_{\min} \leq \mathbf{e}(t) \leq \mathbf{e}_{\max}$ for $\forall t \in [t_0, t_{\text{RUL}}]$, then $\Delta \mathbf{u}$ must belong to \mathcal{U}_δ ,

$$\mathcal{U}_\delta \in \{ \Delta \mathbf{u}_{\min} \leq \Delta \mathbf{u}(t) \leq \Delta \mathbf{u}_{\max} \mid \forall t \in [t_0, t_{\text{RUL}}] \}, \quad (10)$$

where,

$$\begin{cases} \Delta \mathbf{u}_{\min} &= \mathbf{e}_{\min}^+ - \mathbf{e}_{\min}, \\ \Delta \mathbf{u}_{\max} &= \mathbf{e}_{\max}^+ - \mathbf{e}_{\max}. \end{cases} \quad (11)$$

Now, the optimal set-point adjustment $\Delta \mathbf{u}$ is found by minimizing the quadratic cost function,

$$J(\mathbf{x}, \Delta \mathbf{u}) = \min_{\Delta \mathbf{u} \in \mathcal{U}_\delta} \left\{ \int_{t_0}^{t_0+T} [(\mathbf{x}^* - \mathbf{x})^\top (\mathbf{Q}\rho[k]) \cdots (\mathbf{x}^* - \mathbf{x}) + \Delta \mathbf{u}^\top \mathbf{R} \Delta \mathbf{u}] dt \right\}, \quad (12)$$

where \mathbf{x}^* is the desired state-space value. The weight matrices \mathbf{Q} and \mathbf{R} are of size $n_x \times n_x$ and $n_r \times n_r$, respectively.

PHM Module – In the scope of this work, the PHM module is external to the reconfigurable controller. In general, the prognostic control input to the PHM model includes the reference signal, \mathbf{r} , modeled state estimate, \mathbf{x}_m , and set-point adjustment $\Delta \mathbf{u}$.

4 STABILITY AND UNCERTAINTY ANALYSIS

The qualitative overview of the reconfigurable control architecture in the section provides a basis for a quantitative study of set-point reconfiguration with respect to stability and boundedness.

4.1 Reference Model

The MPC requires a reference model of the plant to predict the future set-point adjustments for control reconfiguration. Ideally, the reference model is equivalent to the non-linear plant dynamics. However, using a linear reference model reduces the complexity of the optimal control problem and guarantees a solution exists by optimizing over a convex set. The linear reference model is written in state-space form as,

$$\begin{cases} \dot{\mathbf{x}}_m(t) &= \mathbf{A}_m \mathbf{x}_m + \mathbf{B}_{m,r} \mathbf{r}(t) + \mathbf{B}_{m,v} \mathbf{v}(t), \\ \mathbf{y}_m(t) &= \mathbf{C}_m \mathbf{x}_m + \mathbf{D}_{m,v} \mathbf{v}(t), \end{cases} \quad (13)$$

where $\mathbf{x}_m(0) = \mathbf{x}_{m0}$ and \mathbf{A}_m , $\mathbf{B}_{m,r}$, $\mathbf{B}_{m,v}$, \mathbf{C}_m and $\mathbf{D}_{m,v}$ are real-valued matrices.

4.2 Composite System

The composite system is comprised of the plant and MPC controller, as shown in Figure 5.

4.2.1 Plant

The control input to the plant, \mathbf{u} , is defined as,

$$\mathbf{u}(t) = \mathbf{r}(t) + \Delta \mathbf{u}(t), \quad (14)$$

where $\Delta \mathbf{u}$ is a set-point adjustment computed by the MPC. The output of the plant and corresponding tracking error are represented by \mathbf{y} and \mathbf{e} , accordingly.

4.2.2 MPC Controller

The MPC consists of a linear reference model, state observer and an optimizer. The state-observer accepts the current control input, \mathbf{u} , and plant output, \mathbf{y} , as inputs. The output of the

state observer, \mathbf{x}_m , is used to initialize the reference model. The reference model is used to predict future state estimates for a given set of future input references over a prediction horizon and takes the form of (13). The optimizer solves for a set-point adjustment $\Delta \mathbf{u}$ by minimizing the cost function in (12).

4.3 Error Analysis

Two types of errors are analyzed in this section: tracking error and modeling error. The tracking error corresponds to the difference between the desired input reference, \mathbf{r} , and plant output, \mathbf{y} . The modeling error represented by the symbol \mathbf{e}_m , corresponds to the error in the state estimates that occur as a result of model mismatches propagated over the prediction horizon. The tracking error of the composite system, \mathbf{e}^+ (referred to as the extended tracking error in the previous chapter), is described by the differential equation

$$\dot{\mathbf{e}}^+(t) = \mathbf{A}_e \mathbf{e}^+(t) + \delta_u[k] \cdot \delta(t - kT_s). \quad (15)$$

where \mathbf{A}_e is Hurwitz and $\mathbf{e}^+(t_0) = \mathbf{e}(t_0)$. The following theorem provides the boundaries for the tracking error of the composite system.

Theorem 1 (Tracking Error Boundaries with MPC). *Let the tracking error of the composite system be described by (15). If the set-point adjustment, $\Delta \mathbf{u}$, is uniformly bounded in time by $\Delta \mathbf{u}_{\min} \leq \Delta \mathbf{u}(t) \leq \Delta \mathbf{u}_{\max}$, then the tracking error of the composite system, \mathbf{e}^+ , must also be uniformly bounded in time by, $\mathbf{e}(t_0) + \Delta \mathbf{u}_{\min} \leq \mathbf{e}^+(t) \leq \mathbf{e}(t_0) + \Delta \mathbf{u}_{\max}$.*

Proof. First, the explicit solution of (15) can be found,

$$\mathbf{e}^+(t) = \exp(\mathbf{A}_e(t - t_0)) \mathbf{e}(t_0) + \dots \int_{t_0}^t \exp(\mathbf{A}_e(t - \tau)) (\delta_u[k] \cdot \delta(\tau - kT_s)) d\tau. \quad (16)$$

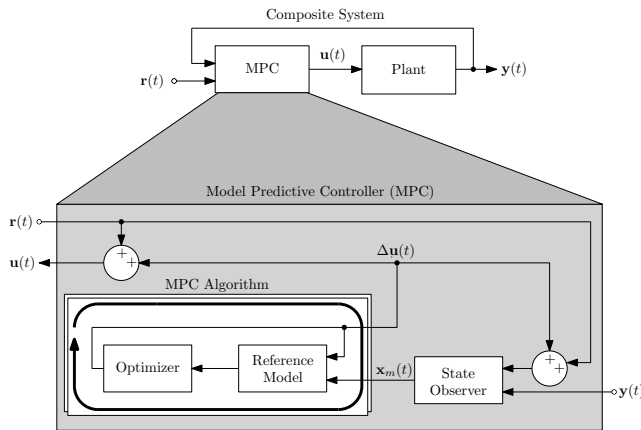


Figure 5. Block diagram of the composite system showing the inner-connections between the state observer, reference model and optimizer within the MPC.

Applying the translation property of the Dirac-delta function gives,

$$\mathbf{e}^+(t) = \exp(\mathbf{A}_e(t - t_0)) \mathbf{e}(t_0) + \dots \sum_{n=0}^k \delta_u[n] \cdot \exp(\mathbf{A}_e(t - nT_s)). \quad (17)$$

Since $\Delta \mathbf{u}$ is uniformly bounded, the cumulative sum of δ_u is bounded by,

$$\Delta \mathbf{u}_{\min} \leq \sum_{n=0}^k \delta_u[n] \leq \Delta \mathbf{u}_{\max}. \quad (18)$$

Now, consider the worst case when $\text{eig}(\mathbf{A}_e) \rightarrow \mathbf{0}^-$. Under this condition, the explicit expression for \mathbf{e}^+ (17) becomes,

$$\mathbf{e}^+(t) = \mathbf{e}(t_0) + \sum_{n=0}^k \delta_u[n]. \quad (19)$$

By applying (18) to (19), the boundary for the case when $\text{eig}(\mathbf{A}_e) \rightarrow \mathbf{0}$ can be given as,

$$\mathbf{e}(t_0) + \Delta \mathbf{u}_{\min} \leq \mathbf{e}^+(t) \leq \mathbf{e}(t_0) + \Delta \mathbf{u}_{\max}. \quad (20)$$

Finally, if \mathbf{A}_e is Hurwitz, then (17) must always be less than or equal to (19) for all $t \geq t_0$. Therefore, by the comparison theorem, (17) must also be bounded by (20). \square

4.4 State-Variable Reconfiguration Analysis

Now that the tracking error of the composite system is shown to be bounded, the effects of set-point adjustment on the future state values can be studied.

4.4.1 Ideal (Matched) Case

First, consider the following definition for the change in the state-variable,

$$\Delta \mathbf{x}(t) \triangleq \mathbf{x}^+(t) - \mathbf{x}(t), \quad (21)$$

where \mathbf{x} is the state of the system if $\Delta \mathbf{u} \equiv \mathbf{0}$ and \mathbf{x}^+ is the reconfigured state of the system if a non-zero set-point adjustment $\Delta \mathbf{u}$ were applied. Now, consider the case where the linear reference model matches the dynamics of the plant. The predicted change in the state-variable after reconfiguration can be found using Theorem 2.

Theorem 2 (State Adjustment (Matched Model)). *Consider a closed-loop system which matches the linear-deterministic reference model described by (13). The estimated change in the state, $\Delta \hat{\mathbf{x}}$, at time t_{k+q} given at time t_k can be computed by,*

$$\Delta \hat{\mathbf{x}}(t_{k+q}|t_k) = e^{\mathbf{A}_m(t_{k+q}-t_k)} \Delta \mathbf{x}_0 + \dots \int_{t_k}^{t_{k+q}} [e^{\mathbf{A}_m(t-\tau)} \cdot \mathbf{B}_{m,\tau} \Delta \mathbf{u}(\tau)] d\tau, \quad (22)$$

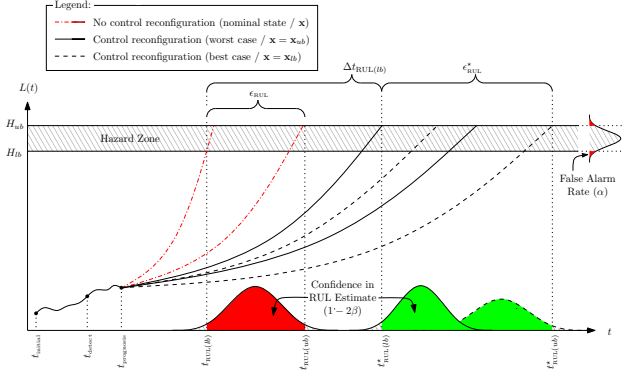


Figure 6. Predicted fault-growth curves, hazard zone and corresponding projection on the time axis for the best-case and worst-case reconfiguration boundaries.

Proof. The dynamics of the reconfigured state can be expressed as,

$$\dot{\mathbf{x}}^+(t) = \mathbf{A}_m \mathbf{x}^+(t) + \mathbf{B}_{m,r} \mathbf{r}(t) + \mathbf{B}_{m,r} \Delta \mathbf{u}(t). \quad (23)$$

Next, taking the time derivative of (21) and substituting the first-order dynamics of \mathbf{x} and \mathbf{x}^+ gives,

$$\Delta \dot{\mathbf{x}}(t) = \mathbf{A}_m \Delta \mathbf{x}(t) + \mathbf{B}_{m,r} \Delta \mathbf{u}(t), \quad (24)$$

The explicit solution to this first-order differential equation is found as,

$$\Delta \mathbf{x}(t) = e^{\mathbf{A}_m(t-t_0)} \Delta \mathbf{x}_0 + \dots \int_{t_0}^t \left[e^{\mathbf{A}_m(t-\tau)} \cdot \mathbf{B}_{m,r} \Delta \mathbf{u}(\tau) \right] d\tau. \quad (25)$$

Finally, since this is assumed to be a perfectly matched model, $\Delta \dot{\mathbf{x}} \equiv \Delta \mathbf{x}$. Therefore, the state estimate at time t_{k+q} given at time t_k can be found by using (25). \square

4.4.2 Non-Ideal (Unmatched) Case

The estimated change in the state at time t_{k+q} given at time t_k when the reference model does not match the closed-loop system dynamics can be found if the structure of the reference and the closed-loop system models are assumed.

Claim 1 (State Adjustment (Unmatched Model)). Consider the case of the unmatched linear reference model in (13). If modeling error over the prognostic horizon, q , is bounded by a constant \bar{e}_m such that,

$$-|\bar{e}_m| \leq \mathbf{e}_m(t) \leq |\bar{e}_m|. \quad (26)$$

for $\forall t \in [t_k, t_{k+q}]$ at time-instant k , then the change in the state, $\Delta \mathbf{x}$, at time t_{k+q} given at time t_k is bounded by,

$$\Delta \mathbf{x}_{lb} \leq \Delta \mathbf{x}(t_{k+q}|t_k) \leq \Delta \mathbf{x}_{ub}, \quad (27)$$

where,

$$\begin{cases} \Delta \mathbf{x}_{lb} &= \Delta \hat{\mathbf{x}}(t_{k+q}|t_k) - |\bar{e}_m| \\ \Delta \mathbf{x}_{ub} &= \Delta \hat{\mathbf{x}}(t_{k+q}|t_k) + |\bar{e}_m| \end{cases} \quad (28)$$

4.5 RUL Analysis

Given uncertainty boundaries for the state-vector \mathbf{x} , the best-case and worst-case prediction boundaries for RUL estimates can be studied in a stochastic manner.

4.5.1 Boundary Conditions

The absolute upper and lower-boundary conditions for each state vector at time t are defined as \mathbf{x}_{ub} and \mathbf{x}_{lb} ,

$$\begin{aligned} \mathbf{x}_{ub}(t_{k+p}|t_k) &= \dots \\ \begin{cases} \mathbf{x}_m + \Delta \mathbf{x}_{ub} : |\mathbf{x}_m + \Delta \mathbf{x}_{ub}| \geq |\mathbf{x}_m + \Delta \mathbf{x}_{lb}| \\ \mathbf{x}_m + \Delta \mathbf{x}_{lb} : |\mathbf{x}_m + \Delta \mathbf{x}_{ub}| < |\mathbf{x}_m + \Delta \mathbf{x}_{lb}| \end{cases}, \end{aligned} \quad (29)$$

$$\begin{aligned} \mathbf{x}_{lb}(t_{k+p}|t_k) &= \\ \begin{cases} \mathbf{x}_m + \Delta \mathbf{x}_{ub} : |\mathbf{x}_m + \Delta \mathbf{x}_{ub}| \leq |\mathbf{x}_m + \Delta \mathbf{x}_{lb}| \\ \mathbf{x}_m + \Delta \mathbf{x}_{lb} : |\mathbf{x}_m + \Delta \mathbf{x}_{ub}| > |\mathbf{x}_m + \Delta \mathbf{x}_{lb}| \end{cases}. \end{aligned} \quad (30)$$

Now, assume that $\frac{\partial L}{\partial \mathbf{x}} \geq 0$. Then, the lower boundary (or worst case conditions) for RUL must occur when $\mathbf{x} = \mathbf{x}_{ub}$,

$$t_{RUL(lb)}^* = t_{RUL(lb)} \Big|_{\mathbf{x}=\mathbf{x}_{ub}}. \quad (31)$$

Similarly, the upper boundary (or best-case condition) for RUL occurs when $\mathbf{x} = \mathbf{x}_{lb}$,

$$t_{RUL(ub)}^* = t_{RUL(ub)} \Big|_{\mathbf{x}=\mathbf{x}_{lb}}. \quad (32)$$

By applying the lower-bound as the most conservative estimate for $t_{RUL(lb)}$, the resulting RUL gained after reconfiguration is defined as,

$$\Delta t_{RUL(lb)} \triangleq t_{RUL(lb)}^* - t_{RUL(lb)}. \quad (33)$$

Additionally, the corresponding confidence interval width of the reconfigured RUL is defined as,

$$\epsilon_{RUL}^* = t_{RUL(ub)}^* - t_{RUL(lb)}^*. \quad (34)$$

An illustration of the predicted fault growth curves for nominal, best-case reconfiguration and worst-case reconfiguration conditions is provided in Figure 6.

4.6 Metrics

Presented are three metrics to evaluate the effectiveness of the reconfiguration routine: remaining life increase (RLI) and prediction uncertainty increase (PUI) and reconfiguration efficiency, represented by the symbol η .

4.6.1 Remaining Life Increase (RLI)

RLI is a standardized measure of the relative net increase in RUL, defined as,

$$RLI \triangleq \frac{t_{RUL(lb)}^* - t_{RUL(lb)}}{t_{RUL(lb)}}. \quad (35)$$

For the case when $RLI < 0$, the RUL decreases ($\Delta t_{RUL} < 0$) thereby leading to an implausible or undesirable reconfiguration action.

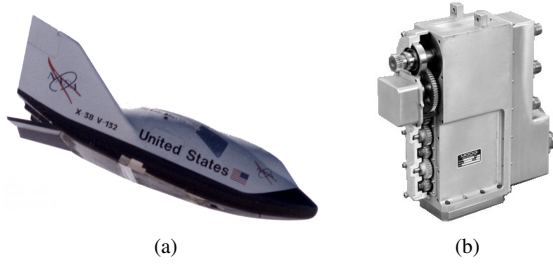


Figure 7. Photo of the (a) X-38 crew re-entry vehicle and its (b) corresponding rudder actuator.

4.6.2 Prediction Uncertainty Increase (PUI)

PUI is a standardized measure of the relative net increase in the width of the RUL confidence interval, defined as,

$$PUI \triangleq \frac{\epsilon_{\text{RUL}}^* - \epsilon_{\text{RUL}}}{\epsilon_{\text{RUL}}}. \quad (36)$$

4.6.3 Reconfiguration Efficiency

Evaluation of RUL feasibility can be difficult to explicitly quantify. A quick estimate of the relative increase in RUL can be made by evaluating the relative change in the cost associated with the plant state before and after reconfiguration. First, define the cost corresponding to the weight ρ as,

$$J(\rho, \mathbf{x}_m, \Delta \mathbf{u}) = \min_{\Delta \mathbf{u} \in \mathcal{U}_s} \left\{ J_x(\rho, \mathbf{x}_m) + J_{\Delta \mathbf{u}}(\Delta \mathbf{u}) \right\}, \quad (37)$$

where,

$$J_x(\rho, \mathbf{x}_m) = \rho (\mathbf{x}^* - \mathbf{x}_m)^\top (\mathbf{Q}\rho[k]) (\mathbf{x}^* - \mathbf{x}_m) \quad (38)$$

and,

$$J_{\Delta \mathbf{u}}(\Delta \mathbf{u}) = \Delta \mathbf{u}^\top \mathbf{R} \Delta \mathbf{u}. \quad (39)$$

Now, the percent change in the cost before and after reconfiguration, η , can be computed,

$$\eta(\rho) = \frac{J_x(0, \mathbf{x}_m) - J_x(\rho, \mathbf{x}_m)}{J_x(0, \mathbf{x}_m)}, \text{ for } \rho \in (0, \infty), \quad (40)$$

where $\eta > 0$ corresponds to a net increase in RUL and $\eta < 0$ corresponds to a net reduction in RUL.

5 EXAMPLE APPLICATION

An EMA is examined as an example for PHM-based control reconfiguration. An EMA was selected in part due to its availability and its emergence as a solution of choice for future flight control actuation systems. More specifically, the rudder of the NASA X-38 crew re-entry vehicle, shown in Figure 7, was selected as the system of interest. A failure modes, effects and criticality analysis (FMECA) of the X-38 rudder actuator was examined to identify the most critical component, degradation of the motor winding insulation.

5.1 Prognostic Model

In the case of the brushless DC motor, the winding temperature is related to the power loss in the copper windings, assuming the copper losses are the primary source of power loss. A first order thermo-electrical model, shown in Figure 8, can be used to describe the relationship between power loss in the copper windings with respect to the winding-to-ambient temperature (Gokdere, Bogdanov, Chiu, Keller, & Vian, 2006; Nestler & Sattler, 1993), represented as T_{wa} and defined as,

$$T_{wa}(t) \triangleq T_w(t) - T_a(t) \quad (41)$$

where the symbols T_w and T_a correspond to the winding temperature and ambient temperature respectively. The symbols R_0 , C_{wa} and R_{wa} refer to the winding resistance, thermal capacitance and thermal resistance of the windings, accordingly. The equivalent state space representation can be written as,

$$\dot{T}_{wa}(t) = \left[-\frac{1}{R_{wa}C_{wa}} \right] T_{wa}(t) + \left[\frac{R_0}{C_{wa}} \right] i_m^2(t) \quad (42)$$

Motor winding insulation degrades at a rate related to the winding temperature, T_w . Let the RUL be represented as, t_{RUL} . The RUL at time t can be related to T_w using Arrhenius' law (Gokdere et al., 2006),

$$t_{\text{RUL}}(t) = c_0 \exp\left(\frac{E_a}{k_B T_w(t)}\right), \quad (43)$$

where the symbols E_a , k_B and c_0 are constants representing activation energy, Boltzmann's constant and an empirical model fit, respectively. Next, let the fault dimension, L , be defined as the accumulated RUL consumed,

$$L(t) = L_0 + \int_{t_0}^t \frac{1}{t_{\text{RUL}}(\tau)} d\tau. \quad (44)$$

where L_0 is the initial fault dimension. Substituting (43) into (44) gives,

$$L(t) = L_0 + \int_{t_0}^t c_0^{-1} \exp\left(-\frac{E_a}{k_B T_w(\tau)}\right) d\tau. \quad (45)$$

By differentiating both sides with respect to time and applying the second fundamental theorem of integral calculus to the right-hand side, an expression for \dot{L} can be found,

$$\dot{L}(t) = c_0^{-1} \exp\left(-\frac{E_a}{k_B T_w(t)}\right), \quad (46)$$

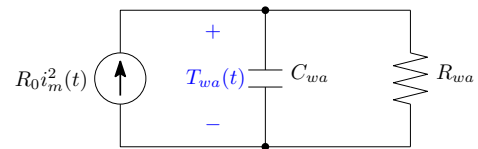


Figure 8. Schematic of the first-order thermal model.

5.2 Model Uncertainty

Consider the effects of model uncertainty for a linear system with an unmatched linear reference model. In this example only the motor current is of interest; therefore, consideration of the entire state estimate is simplified to the scalar quantity, i_m . Values for $\Delta \mathbf{A}_m$ and $\Delta \mathbf{B}_{m,r}$ are obtained by using new values for the modeling parameters after adjusting physical modeling parameters randomly to within their corresponding uncertainties. A sinusoidal input with an amplitude of 60 deg was used as the reference applied to the linear actuator model. The sample-time between predictions was set at $T_s = 0.05$ s. Monte Carlo simulations were conducted for a range of frequencies from 0.1 Hz to 10 Hz. The percent change in motor current vs. reference frequency was estimated from each set of Monte Carlo simulations. According to the results, the standard deviation of the percent change in motor current is approximately less than 0.2014 for 95% of the simulations.

5.3 Long-Term State Predictions with Uncertainty

It can be shown the modeling error corresponding to a 95% confidence interval is approximately,

$$-0.395\hat{i}_m \leq \mathbf{e}_m(t) \leq 0.395\hat{i}_m(t) \quad (47)$$

where \hat{i}_m is the estimated value of the motor current. By applying (29) and (30), a 95% confidence interval for the motor current can be expressed as,

$$0.605\hat{i}_m(t) \leq i_m(t) \leq 1.395\hat{i}_m(t). \quad (48)$$

To demonstrate this boundary, consider the actuator example with the maximum reconfiguration possible, which corresponds to an adaptation parameter $\rho \gg 1$. Let the reference signal be sinusoidal with an amplitude of 60 deg and a fixed frequency of 2 Hz. The corresponding actuator position reference signal before and after reconfiguration is shown in Figure 9. Also shown is the set-point adjustment applied to the reference signal and the corresponding motor current with uncertainty boundaries. The reconfiguration efficiency for large values of ρ was computed as $\eta = 0.19$.

5.4 RUL Estimation & Uncertainty

RUL estimation and uncertainty are examined for a simple reference signal. Recall, the input to the prognostic model is the squared current value, i_m^2 and $J_x(\rho, \mathbf{x}_m)$ is directly proportional to i_m^2 . After reconfiguration, the cost function reduces by a factor of η . Therefore, the quantity $(1 - \eta)i_m^2$ can be used to represent the input to the prognostic model after reconfiguration. This can be demonstrated using a simple example. Let the reference signal be a sinusoidal input with a frequency of 2 Hz and an amplitude of 60 deg. For these conditions i_m was simulated as,

$$i_m(t) = 7.07 \sin(4\pi ft). \quad (49)$$

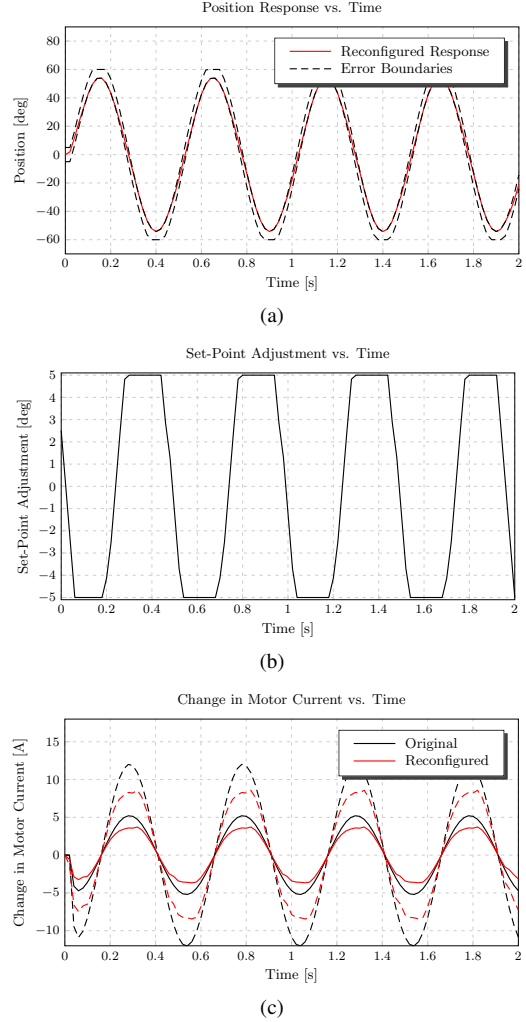


Figure 9. Plots of the (a) actuator position (b) applied set-point adjustment and (c) corresponding motor current with 95% uncertainty boundaries. Dashed / solid lines correspond to the upper / lower 95% confidence boundaries.

For the nominal case when $\rho = 0$, the current is bounded by,

$$4.277 \sin(4\pi t) \leq i_m(t) \leq 9.863 \sin(4\pi t). \quad (50)$$

and the input to the prognostic model is bounded by,

$$18.29 \sin^2(4\pi t) \leq i_m^2(t) \leq 97.27 \sin^2(4\pi t). \quad (51)$$

Similarly, after applying the MPC for $\rho \gg 1$, the reconfiguration efficiency becomes $\eta = 0.19$. The input to the prognostic model is adjusted by a factor of $(1 - \eta)$, which becomes,

$$14.82 \sin^2(4\pi t) \leq i_m^2(t) \leq 78.79 \sin^2(4\pi t). \quad (52)$$

Applying these boundaries to the input of the prognostic model, a plot of the fault dimension (life consumed) versus the prognostic horizon can be obtained for the cases before and after reconfiguration, provided in Figure 10. From the plot values for Δt_{RUL} , ϵ_{RUL} and ϵ_{RUL}^* are computed as

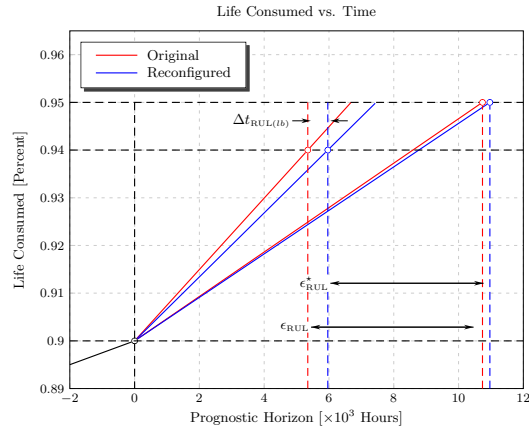


Figure 10. Plot of life consumed versus prognostic horizon before and after reconfiguration for a sinusoidal reference input with a frequency of 2 Hz and an amplitude of 60 deg.

6.178×10^2 hrs, 5.393×10^3 hrs and 5.003×10^3 hrs, accordingly. This allows the metrics RLI and PUI to be computed as 0.116 and -0.080 , respectively.

6 CONCLUSIONS

This body of work constitutes a significant effort regarding the specific role of RUL in control systems. The overall control scheme was defined as a module which adjusts the reference set-points to the local production controller in order to sacrifice a fixed amount of performance to achieve an increase in RUL. The modules of the reconfigurable controller, the MPC and state observer, were defined mathematically and analyzed to demonstrate stability and boundedness. Finally, the reconfigurable control framework was evaluated using an EMA Simulink model. Results acquired from the simulation demonstrated the feasibility of the approach.

ACKNOWLEDGMENT

This research was supported by the U.S. Department of Defense, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a.

REFERENCES

- Bogdanov, A., Chiu, S., Gokdere, L., & Vian, W., J. (2006, December). Stochastic Optimal Control of a Servo Motor with a Lifetime Constraint. In *Proceedings of the 45th IEEE Conference on Decision & Control* (p. 4182-4187).
- Darby, R. (2006, August). Commercial Jet Hull Losses, Fatalities Rose Sharply in 2005. *Datalink*, 51-53.
- Filippetti, F., Franceschini, G., Tassoni, C., & Vas, P. (2000, October). Recent Developments of Induction Motor Drives Fault Diagnosis Using AI Techniques. *IEEE Transactions on Industrial Electronics*, 47(5), 994-1004.
- García, C. E., Prett, D. M., & Morari, M. (1989, May). Model predictive control: Theory and practice – A survey. *Automatica*, 25(3), 335-348.
- Gokdere, L. U., Bogdanov, A., Chiu, S. L., Keller, K. J., & Vian, J. (2006, March). Adaptive control of actuator lifetime. In *IEEE Aerospace Conference*.
- Isermann, R. (1984). Process Fault Detection Based on Modeling and Estimation Methods - a Survey. *Automatica*, 20(4), 387-404.
- Isermann, R. (1997, May). Supervision, fault-detection and fault diagnosis methods – An introduction. *Journal of Control Engineering Practice*, 5(5), 639-652.
- Jiang, J., & Zhao, Q. (1997, August). Should we use parameter estimation or state estimation based methods for FDI. In *Proceedings of the international federation of automatic control on SAFEPROCESS* (p. 474-479). Hull, UK.
- Kleer, J. D., & Williams, B. C. (1987). Diagnosing Multiple Faults. *Artificial Intelligence*, 32(1), 97-130.
- Kwon, W. H., Bruckstein, A. N., & Kailath, T. (1983). Stabilizing state feedback design via the moving horizon method. *International Journal of Control*, 37(3), 631-643.
- Levis, A. H. (1987, April). Challenges to Control: A Collective View. *IEEE Transactions on Automatic Control*, 32(4), 275-285.
- Monaco, J. F., D.G., W., & Bateman, A. J. D. (2004, September 20-22). A Retrofit Architecture for Model-Based Adaptive Flight Control. In *AIAA 1st Intelligent Systems Technical Conference*. Chicago, IL, USA.
- Nestler, H., & Sattler, P. K. (1993, January). On-Line Estimation of Temperatures in Electrical Machines by an Observer. *Electric Power Components and Systems*, 21(1), 39-50.
- Orchard, M. (2007). *A Particle Filtering-based Framework for On-line Fault Diagnosis and Failure Prognosis*. Unpublished doctoral dissertation, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA.
- Orchard, M., Kacprzynski, G., Goebel, K., Saha, B., & Vachresevanos, G. (2009). *Advances in Uncertainty Representation and Management for Particle Filtering Applied to Prognosis* (Vol. 39; K. P. Valavanis, Ed.) (No. 1). Springer Netherlands. (ISBN 978-90-481-3017-7)
- Paris, P., & Erodogan, F. (1963). A critical analysis of crack propagation laws. *ASME Journal of Basic Engineering Transactions*, 85, 528-534.
- Patterson, R. J. (1997, August). Fault-tolerant control: the 1997 situation. In *Proceedings of the International Federation of Automatic Control symposium on SAFEPROCESS* (p. 1033-1055). Hull, UK.
- Rausch, R., Goebel, K., Eklund, N., & Brunell, B. (2007,

October). Integrated Fault Detection and Accommodation: A Model-Based Study. *Journal of Engineering for Gas Turbines and Power*, 129(4), 962-969.

Richalet, J. (1993). Industrial applications of model based predictive control. *Automatica*, 29, 1251-1274.

Richalet, J., Rault, A., Testud, J. L., & Papon, J. (1978). Model predictive heuristic control: Applications to industrial processes. *Automatica*, 14(5), 413-428.

Saberi, A., Stoorvogel, A. A., Sannuti, P., & Niemann, H. (2000, November). Fundamental problems in fault detection and identification. *International Journal of Robust and Nonlinear Control*, 10(14), 1209-1236.

Schwabacher, M. (2005). A Survey of Data-Driven Prognostics. In *Proceedings of the AIAA Infotech@Aerospace Conference*. Reston, VA, USA.

Skormin, V. A., Apone, J., & Dunphy, J. J. (1994, January). On-line Diagnostics of a Self-Contained Flight Actuator. *IEEE Transactions on Aerospace and Electronic Systems*, 30(1), 186-196.

Srivastava, A. N., Mah, R. W., & Meyer, C. (2008, December). *Integrated Vehicle Health Management – Automated detection, diagnosis, prognosis to enable mitigation of adverse events during flight* (Technical Plan No. Version 2.02). National Aeronautics and Space Administration.

Vachtsevanos, G., Lewis, F., Roemer, M., Hess, A., & Wu, B. (2006). *Intelligent Fault Diagnosis and Prognosis for Engineering Systems*. Hoboken, NJ, USA: John Wiley & Sons. (ISBN 987-0-0471-72999-0)

Wald, M. (2007, October 1). Fatal Airplane Crashes Drop 65%. *The New York Times*.

Willsky, A. S. (1976). A Survey of Design Methods for Failure Detection in Dynamic Systems. *Automatica*, 12(6), 601-611.

Wu, B., Abhinav, S., Khawaja, T., & Panagiotis, S. (2004, September 20-23,). An Approach to Fault Diagnosis of Helicopter Planetary Gears. In *IEEE Autotestcon*. San Antonio, TX, USA.

Wu, N. E. (1997, September). Robust Feedback Design with Optimized Diagnostic Performance. *IEEE Transactions on Automatic Control*, 42(9), 1264-1268.

Wu, N. E., Zhang, Y. M., & Zhou, K. (2000). Detection, estimation and accommodation of loss of control effectiveness. *International Journal of Adaptive Control and Signal Processing*, 14(7), 948-956.

Yu, W., & Harris, T. A. (2001). A new stress-based fatigue life model for ball bearings. *Tribology Transactions*, 44(1), 11-18.

Zhang, Y., & Jiang, J. (2003). Bibliographical review on reconfigurable fault-tolerant control systems. In *Proceeding of the SAFEPROCESS 2003: 5th Symposium on Detection and Safety for Technical Processes* (p. 265-276). Washington D.C., USA.

Zhang, Y., & Jiang, J. (2008, March). Bibliographical re-

view on reconfigurable fault-tolerant control systems. *Annual Reviews in Control*, 32(2), 229-252.

Zhang, Y. M., & Jiang, J. (2002). An active fault-tolerant control system against partial actuator failures. In *IEE Proceedings - Control Theory and Applications* (p. 95-104).

Zhenyu, & Hicks, D. L. (2006, December 5-8,). Synthesis of Robust Restructurable/Reconfigurable Control. In *9th International conference on control, automation, robotics and vision* (p. 1-6). Singapore.



Douglas W. Brown received the B.S. degree in electrical engineering from the Rochester Institute of Technology in 2006 and the M.S degree in electrical engineering from the Georgia Institute of Technology in 2008. He is a recipient of the National Defense Science and Engineering Graduate (NDSEG) Fellowship and is currently a Ph.D. candidate in electrical engineering at the Georgia Institute of Technology specializing in control systems. His research interests include incorporation of Prognostics Health Management (PHM) for fault-tolerant control. Prior to joining Georgia Tech, Mr. Brown was employed as a project engineer at Impact Technologies where he worked on incipient fault detection techniques, electronic component test strategies, and diagnostics/prognostic algorithms for power supplies and RF component applications.



George J. Vachtsevanos is a Professor Emeritus of Electrical and Computer Engineering at the Georgia Institute of Technology. He was awarded a B.E.E. degree from the City College of New York in 1962, a M.E.E. degree from New York University in 1963 and the Ph.D. degree in Electrical Engineering from the City University of New York in 1970.

He directs the Intelligent Control Systems laboratory at Georgia Tech where faculty and students are conducting research in intelligent control, neurotechnology and cardiotechnology, fault diagnosis and prognosis of large-scale dynamical systems and control technologies for Unmanned Aerial Vehicles. His work is funded by government agencies and industry. He has published over 240 technical papers and is a senior member of IEEE. Dr. Vachtsevanos was awarded the IEEE Control Systems Magazine Outstanding Paper Award for the years 2002-2003 (with L. Wills and B. Heck). He was also awarded the 2002-2003 Georgia Tech School of Electrical and Computer Engineering Distinguished Professor Award and the 2003-2004 Georgia Institute of Technology Outstanding Interdisciplinary Activities Award.

Damage Identification in Frame Structures, Using Damage Index, Based on H_2 -Norm

Mahdi Saffari¹, Ramin Sedaghati², and Ion Stiharu³

^{1,2,3} *Concordia University, Montreal, Quebec, H3G 1M8, Canada*

m_saffa@encs.concordia.ca

sedagha@encs.concordia.ca

istih@encs.concordia.ca

ABSTRACT

A simulation method to detect and locate damage in frame structures by defining a damage index is proposed. Structural members are Timoshenko beam type. The method defines a damage index which is the reduction percentage of H_2 norm of the structure at certain locations in both healthy and damaged states. Structure modeling is done by finite element method.

1. INTRODUCTION

Defining a damage index (D.I.) has been on focus in many publications. Extensive literature reviews on vibration-based damage detection methods is published by Doebling, Farrar, Prime and Shevitz (1996) and Carden and Fanning (2004). Looking to these various vibration based techniques, particularly those using modal parameters, the D.I. method seems more promising. The basic idea behind defining damage indices is that changes in physical properties of a structure will eventually alter some of the system intrinsic properties such as some of natural frequencies, mode shapes or mode shape curvatures (Choi & Stubbs, 2004). A Damage Index is defined based on the changes of the j th mode curvature at location i (Stubbs, Kim, & Farrar, 1995). Choi and Stubbs (2004) used the strain energy of pre and post damaged structure to define D.I.. Also combination of D.I. and neural network method is used to identify damage in structures (Dackermann, Li, & Samali, 2010). In mode shape curvature based D.I.; changes in the damage index and relating these changes with the potential locations are assessed by statistical methods. Normal distribution of damage indices in different locations is extracted and D.I. values which are two or more standard deviation away from the mean D.I. value are reported to be most probable location

of damage (Stubbs, et al., 1995). An extension to mode shape curvature method is that one can take into account all frequencies in the measurement range and not just the modal frequencies. In other words one may use Frequency Response Function (FRF) instead of mode shape data. It is claimed that this method can detect, localize and assess damage extent. The theory is fostered with some experimental results (Sampaio, Maia, & Silva, 1999).

Nevertheless development of suitable and reliable damage metrics and identification algorithms is still an issue to be investigated. D.I. as a scalar quantity is a damage metric that gives a criterion to judge the extent of damage of a structure (Giurgiutiu, 2008). Although these methods are well applicable in some cases but are not usually applicable to the cases that the sizes of cracks are small relative to the structure, or the crack is somewhere in a wide area of the structure. The main reason is that small cracks do not change the lower modal properties appreciably and thus they are not easily detectable using experimental data. It should be noted that this limitation is not due to lack of sensitivity of the method, but it is due to the practical limitations of exciting higher modes. Excitation of higher modes requires significant amount of energy which may not be viable to large structural systems (Ginsberg, 2001).

2. PROBLEM STATEMENT

A 2D frame type structure as shown in Figure 1 is studied. A D.I. based on H_2 norm, as discussed in next section, is formulated to compare the healthy and damaged state of the structure and localize the damage. The structure is modeled using 16 two-node Timoshenko beam element in which each node has 3 degrees of freedom (DOF). Timoshenko beam theory has proved to give more accurate results when the length of the beam element is relatively short (Reddy, 2004). Damage is modeled by reducing the stiffness in the element confined between nodes 11 and 4 by 80%. The

Mahdi Saffari et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

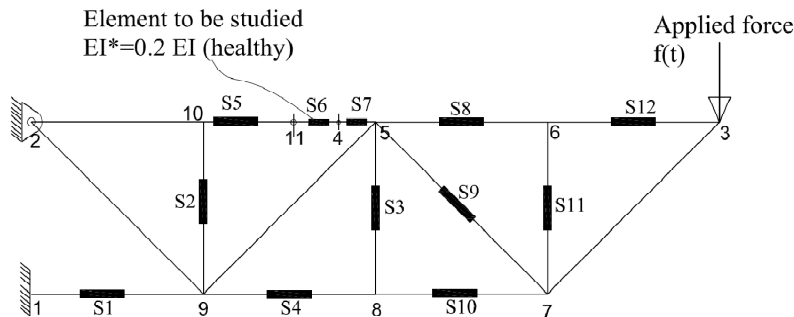


Figure 1. Frame structure configuration and strain gauge sensors placement

material properties of the members are considered as:

$$E = 200 \text{ GPa}, G = 80 \text{ GPa}, \rho = 7800 \frac{\text{kg}}{\text{m}^3} \quad (1)$$

The cross section of members is $3\text{cm} \times 3\text{cm}$ and the length of each horizontal or vertical member is 1 m (Figure 1).

The structure is fixed in all DOF (X, Y, θ) at node 1 and only in translational DOF (X, Y) at node 2. Hence the structure has 28 free DOFs. There are 12 strain gauge sensors placed in different locations of the structure. There are relatively more strain gauges near the damaged link to have more accuracy in finding damage. The input force is applied on node number 3 as shown in the Figure 1. Mass and stiffness matrices of the structure are found after assembling the global stiffness and consistent mass matrix of all elements using finite element technique. The system damping is assumed to be proportional to the system stiffness and mass matrices based on Rayleigh damping as:

$$D = \alpha M + \beta K \quad (2)$$

The parameters α and β are considered here to be 0 and 0.001, respectively.

3. PROBLEM FORMULATION

The governing equations of a linear structure in the finite element form can be described as (Gawronski, 2004)

$$M\ddot{q} + D\dot{q} + Kq = B_o u \quad (3)$$

For the 2D frame structure discussed in previous section, M, D and K are 28×28 mass, damping and stiffness matrices, respectively. B_o is input vector and q is nodal displacement vector and both are 28×1 vectors. u is the input force magnitude.

The desired output is the strains in specified members. This output is a linear combination of system nodal displacements. For example, for the element with strain gauge S4:

$$S4: \epsilon_x = \frac{q_{x8} - q_{x9}}{L_{8-9}} \quad (4)$$

q_{x8}, q_{x9} : Displacement of node 8 and 9 in x- direction

L_{8-9} : Length of member 8-9

ϵ_x : Strain in member 8-9

Thus the output vector y has 12 strain components which can be related to the nodal displacement vector q as

$$y = C_q q \quad (5)$$

where C_q is a 12×28 matrix.

3.1 Modal model

Modal model in structures is a standard modeling procedure in which modal displacement vector (q_m) is related to the original nodal displacement vector q as

$$q = \Phi q_m \quad (6)$$

in which (Φ) is the system modal matrix whose columns are eigenvectors (normal modes) of the system.

Now by substituting Eq. (6) into Eq. (3) and then multiplying the resulting equation from left side by transpose of (Φ), one may write:

$$M_m \ddot{q}_m + D_m \dot{q}_m + K_m q_m = \Phi^T B_o u \quad (7)$$

in which

$$\begin{aligned} M_m &= \Phi^T M \Phi \\ D_m &= \Phi^T D \Phi \\ K_m &= \Phi^T K \Phi \end{aligned} \quad (8)$$

are modal mass, modal damping and modal stiffness matrices which are diagonal due to orthogonality of eigenvectors. (Rao, 2007)

Also the output vector described in Eq. (5) can be written as:

$$y = C_{mq}q_m \quad (9)$$

in which C_{mq} is the modal system output matrix written as:

$$C_{mq} = C_q \Phi \quad (10)$$

Multiplying Eq. (7) by the inverse of the modal mass, M_m^{-1} , from the left side yields:

$$\ddot{q}_m + M_m^{-1}D_m\dot{q}_m + M_m^{-1}K_mq_m = M_m^{-1}\Phi^T B_o u \quad (11)$$

or

$$\ddot{q}_m + 2Z\Omega\dot{q}_m + \Omega^2q_m = B_m u \quad (12)$$

in which $\Omega = M_m^{-1/2}K_m^{1/2}$ is the diagonal matrix of eigenvalues (natural frequencies):

$$\Omega = \begin{bmatrix} \omega_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \omega_n \end{bmatrix} \quad (13)$$

and Z is the diagonal modal damping matrix defined as:

$$Z = \frac{D_m}{2\sqrt{M_m K_m}} \quad (14)$$

Modal system input matrix B_m is also defined by

$$B_m = M_m^{-1}\Phi^T B_o \quad (15)$$

3.2 H₂ norm

Based on modal representation of the linear system, and derived system modal matrices, the H₂ norm of the system is defined. Norms are employed to quantify the intensity of system response to standard excitations, such as unit impulse, or white noise of unit standard deviation. H₂ norm is used to compare two different situations. It should be noted that H₂ norm of a mode with multiple inputs (or outputs) can be broken down into the rms sum of norms of that mode with a single input (or output) (Gawronski, 2004).

Now let us consider a flexible structure with one actuator (or one input) and n modes (n=system DOF), the modal input matrix B is then:

$$B_m = \begin{bmatrix} B_{m1} \\ B_{m2} \\ \vdots \\ B_{mn} \end{bmatrix} \quad (16)$$

For the 2D frame structure discussed before, B_m has 28 rows and one column and B_{mi} corresponds to the actuator effect on i th mode.

Similar to the actuator properties, for r sensors installed on a n DOF structure, the output matrix is as follows:

$$C_m|_{r \times n} = [C_{m1}, C_{m2}, \dots, C_{mn}] \quad (17)$$

For mode number j

$$C_{mj} = \begin{bmatrix} C_{m1j} \\ C_{m2j} \\ \vdots \\ C_{mrj} \end{bmatrix} \quad (18)$$

The H₂ norm of the i th mode of a structure with a set of r sensors is the rms sum of the H₂ norms of the mode with each single sensor from this set. Norm of a structure with one actuator and multiple sensors is defined as (Gawronski, 2004)

$$\|G_{mi}\|_2 \cong \frac{\|B_{mi}\|_2 \|C_{mi}\|_2}{2\sqrt{\zeta_i \omega_i}} \quad (19)$$

The j th sensor H₂ norm of the structure corresponding to each sensor could be derived similar to modal H₂ norm as (Gawronski, 2004):

$$\|G_{sj}\|_2 \cong \frac{\|B_{mj}\|_2 \|C_{msj}\|_2}{2\sqrt{\zeta_j \omega_j}} \quad (20)$$

4. DAMAGE INDEX (D.I.)

To localize damaged elements of a structure, a damage index attributed to the sensor (sensor damage index) is defined (Gawronski, 2004). By denoting the norm of the j th sensor of the healthy structure by $\|G_{sj}^h\|_2$, and the norm of the j th sensor of the damaged structure by $\|G_{sj}^d\|_2$. The j th sensor index of the structural damage is defined as a weighted difference between the j th sensor norms of a healthy and damaged structure as:

$$DI_j^s = \left| \frac{\|G_{sj}^h\|_2^2 - \|G_{sj}^d\|_2^2}{\|G_{sj}^h\|_2^2} \right| \quad (21)$$

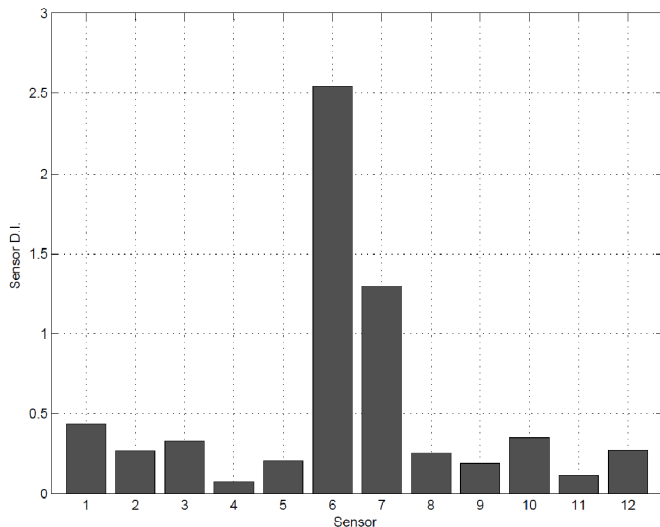


Figure 2. Sensor damage indices

5. RESULTS

The H_2 norm damage index defined in Section 4 has been evaluated for the 2D frame structure as described before in section 2. Using the modal finite element formulation elaborated in Section 3, Figure 2 indicates the sensor D.I. in all 12 sensors.

As it can be seen from Figure 2, the sensor number 6 (S6) has the highest D.I. value indicating that the most probable place to have damage is member between nodes 4 and 11 (member 11-4) which is indeed the location of the defined damage.

The developed algorithm can be easily applied to identify multiple damage locations in the case that structure has more than one damaged spot. Naturally, more sensors should be added to reasonably accurate results and increase the algorithm sensitivity.

In this example it is assumed that the structural member between nodes 5 and 7 (member 5-7) is divided into 4 elements and members 5-6 and 5-8 are also divided into 3 elements and new strain gauges are installed on these new elements as shown in Figure 3. Damage is introduced to element 13-14 (S14) as well as previous member 11-4 (S5).

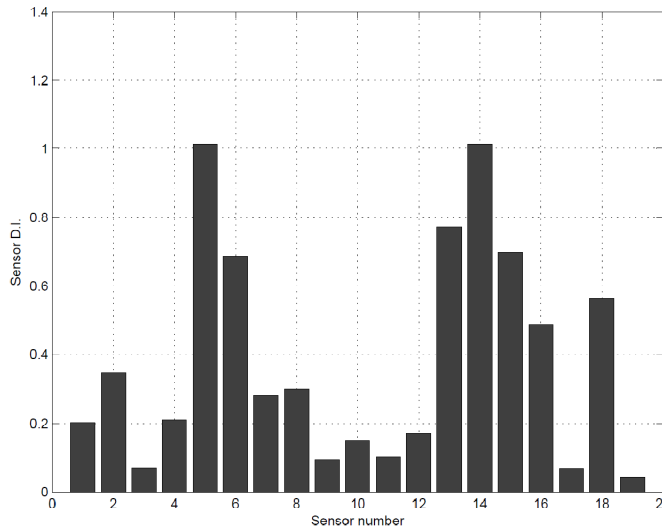


Figure 4. Sensor damage index for structure with two damage spots.

It is assumed that both members have 80% reduction in stiffness EI . Figure 4 indicates the sensor D.I. for this new damage configuration. It could be seen that the algorithm has accurately identified the exact damage locations because the damage index in 5th and 14th locations are the two highest.

6. CONCLUSION

A methodology to detect and locate damage in frame structures by defining a damage index is formulated. Structural members are modeled as Timoshenko beams type. The method defines a damage index which is the reduction percentage of H_2 norm of the structure at certain locations where strain gauges are installed and compares both healthy and damaged states. However to have accurate results one should install enough number of sensors. There is room to extend this work by installing different types of sensors such as accelerometers or to find the minimum number of required sensors to have accurate results as possible.

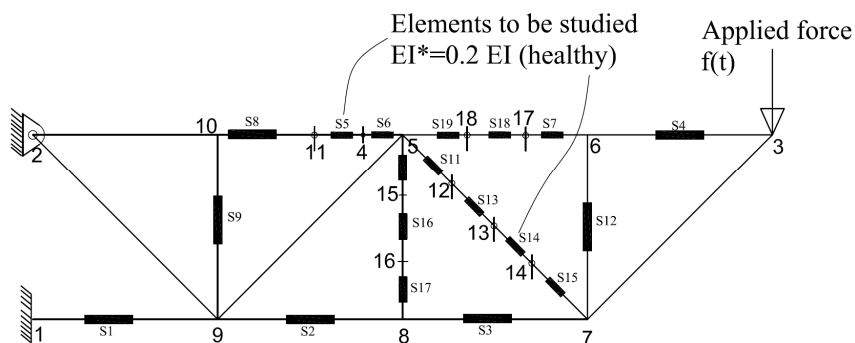


Figure 3. Frame structure with two damage spots

ACKNOWLEDGEMENT

NSERC and Faculty of Engineering and Computer Science at Concordia are acknowledged for their financial support of the project.

REFERENCES

- Carden, E. P., & Fanning, P. (2004). Vibration based condition monitoring: a review. *Structural Health Monitoring*, 3(4), 355.
- Choi, S., & Stubbs, N. (2004). Damage identification in structures using time-domain response. *Journal of Sound and Vibration*, 275(Copyright 2004, IEE), 577-590.
- Dackermann, U., Li, J., & Samali, B. (2010). Technical Papers: Dynamic-Based Damage Identification Using Neural Network Ensembles and Damage Index Method. *Advances in Structural Engineering*, 13(6), 1001-1016.
- Doebling, S. W., Farrar, C. R., Prime, M. B., & Shevitz, D. W. (1996). *Damage identification and health monitoring of structural and mechanical systems from changes in their vibration characteristics: a literature review*: Los Alamos National Lab., NM (United States).
- Gawronski, W. K. (2004). *Advanced structural dynamics and active control of structures*: Springer Verlag.
- Ginsberg, J. H. (2001). *Mechanical and structural vibrations: theory and applications*: John Wiley & Sons.
- Giurgiutiu, V. (2008). *Structural health monitoring with piezoelectric wafer active sensors*: Academic Pr.
- Rao, S. S. (2007). *Vibration of continuous systems*: Wiley Online Library.
- Reddy, J. N. (2004). *Mechanics of laminated composite plates and shells: theory and analysis*: CRC.
- Sampaio, R., Maia, N., & Silva, J. (1999). Damage detection using the frequency-response-function curvature method. *Journal of Sound and Vibration*, 226(5), 1029-1042.
- Stubbs, N., Kim, J. T., & Farrar, C. R. (1995). *Field verification of a nondestructive damage localization and severity estimation algorithm*.

Enhanced Multivariate Based Approach for SHM Using Hilbert Transform

Rafik Hajrya, Nazih Mechbal, Michel Vergé

Process and Engineering in Mechanics and Materials Laboratory (PIMM)

Arts et Métiers ParisTech

Paris, France

Rafik.HAJRYA-7@etudiants.ensam.eu

nazih.mechbal@ensam.eu

michel.verge@ensam.eu

ABSTRACT

In structural health monitoring, features extraction from measured data plays an important role. In order to enhance information about damage, we propose in this paper, a new damage detection methodology, based on the Hilbert transform and multivariate analysis. Using measurements given by distributed sensors of a smart composite structure, we apply the Hilbert transform to calculate an envelope matrix. This matrix is then treated using multivariate analysis. The subspaces associated to the envelope matrix are used to define a damage index (DI). Furthermore, from perturbation theory of matrices, we propose a bound associated to this DI, by inspecting this bound, decision on the health of the structure is generated. Experimentation on an actual composite smart structure will show the effectiveness of the proposed approach.

1. INTRODUCTION

Composite structures have been increasingly adopted by the aviation community to provide high performance, strength, stiffness and weight reduction. One of the major concerns associated with composites is the susceptibility to impact damage, (Staszewski 2002). Impact damage may occur during manufacture, service or maintenance. Low-velocity impacts are often caused by bird strikes, runway stones and tool-drops during maintenance. Impacts can induce serious damage to composites such as delamination, matrix and fiber cracking. Faced with these various damages, a structural health monitoring system (SHM) is needed and if possible in real time.

SHM methods are implemented on structures known as "smart structures", (Giurgiutiu et al. 2002). These structures consist of a network sensors and actuators and offer a monitoring capability for real-time application. Recently emerged piezoceramic patches have the potential to improve significantly

developments of structural health monitoring systems. These patches offer many advantages, among of them: lightweight properties, relative low-cost and can be produced in different shapes. Recently, (Su et al. 2006) have developed a sensor network for SHM using printed circuit to embed piezoceramic patches into a composite structure.

Damage is a structural state which is different from a reference state that is healthy. A damage event is not meaningful without comparisons between two different structural states. The greatest challenge is to ascertain what changes are sought in the signal after the presence of damage. Features extraction is therefore a key step in the processing of signal sensor. In SHM, feature extraction is the process of identifying damage-sensitive properties derived from the measured response data of a smart structure; it serves an indicator to describe the damage and its severity. These extracted features are termed as damage index (DI). Recently, the method of empirical mode decomposition (EMD) and Hilbert transform have been applied in SHM, (Huang et al. 1998). By applying EMD and Hilbert transform in a measured data, (Yang et al. 2004) have developed a method to detect the damage time instant and damage location, in addition they propose in others works the identification of linear structure using the EMD and Hilbert transform, (Yang et al. 2003a; Yang et al. 2003b).

In recent years, techniques based on multivariate statistics have been also applied in SHM. As the name implies, multivariate analysis is concerned with the analysis of multiple measurements from sensors and treats them as a single entity. There are two major multivariate techniques in SHM, principal components analysis (PCA) and independent components analysis (ICA). These techniques serve two purposes, namely order reduction and feature extraction by revealing structure hidden in the measurement, (Kerschen et al. 2005). By applying a PCA on the sensor time responses, (De Boe and Golinval 2003) have developed

a damage index based on angle between subspace to detect and locate damage, in addition (Hajrya et al. 2011) have applied the same principle and they propose a bound based on correlation coefficient that automatically decides if a composite structure is in healthy or damaged state. Using independent component analysis combined with artificial neural network, (Zang et al. 2004) have used a mixing matrix which is extracted from ICA to detect and locate damage.

In this work, we propose an original damage index (DI) based on the calculation of an envelope matrix. This matrix is built using the Hilbert transform of time response matrix measurements. Furthermore, from perturbation theory of matrices, we define a bound that automatically decides if the composite structure is in healthy or damaged status.

The paper is organized as follows: In the next section the experimental test is presented. In section 3, the mathematical formulation of the Hilbert transform and the multivariate analysis are briefly described. In section 4, our methodology for damage detection is presented. In section 5, the proposed damage detection scheme is applied on an experimental laboratory test bench. Finally, conclusions and further directions will be drawn in section 6. Main terms, table and figures are illustrated at the end of the paper before the references.

2. EXPERIMENTAL TEST BENCH

The structure employed consists of a piece of composite fuselage; it was manufactured by INEO DEFENSE which is a partner in the MSIE research program. The structure consists of a carbone-epoxy composite plate with dimensions: $(400 \times 300 \times 2\text{mm})$ and it is made up of 16 layers. The layers sequences are: $[0^\circ_2, +45^\circ_2, -45^\circ_2, +90^\circ_2, -90^\circ_2, -45^\circ_2, +45^\circ_2, 0^\circ_2]$. The properties of the composite plate are detailed in table 1. Using a modal approach, we have performed in a previous work, (Hajrya et al. 2010), an optimal placement of ten piezoceramic patches (figure 2), with dimensions $(30 \times 20 \times 0.2\text{mm})$. The piezoceramic patches are made on lead zirconate titanate (PZT). Figure 1 is a diagram and it shows the positions of the ten PZT in the composite plate. It is to be noted that in our work, only nine PZT are used (PZT 6 is not taken into account in the **damage detection** methodology). Sensor PZT 6 will be used in another work for **sensor fault detection**.

Figure 2 shows the experimental smart composite plate and it was used as baseline for damage detection. In order to develop a damage detection methodology, we have used a second composite plate with the same dimensions and numbers of PZT (at the same location), but, in this plate, impact damage was produced throwing a ball at high velocity: the damage is located

at the middle of the plate. Figure 3 shows the location of this impact damage.

The input excitation generation and the data acquisition were made using a commercial system dSPACE ®. The input excitation consists in a signal pulse with 1ms width. Signals were acquired with sampling frequency $f_s = 100\text{kHz}$, time duration was $T = 0.65\text{s}$ and $N = 2^{16}$ time samples were recorded for each channel: one corresponding to the excitation applied to the PZT actuator and the others concern the measurements collected by the PZT sensors. Figure 4 shows the time responses of sensor PZT 7 in the case of the healthy and damaged plate while we have used PZT 10 as actuator, *i.e.* (Path PZT 10-PZT7): only the 512 first samples are displayed.

3. MATHEMATICAL FORMULATION

3.1 Hilbert transform

The Hilbert transform of an arbitrary signal $y(t)$ is defined as, (Bendat and Piersol 2000):

$$\tilde{y}(t) = \mathcal{H}[y(t)] = \int_{-\infty}^{+\infty} \frac{y(u)}{\pi(t-u)} du \quad (1)$$

Equation (1) is the convolution integral of $y(t)$ and $(1/\pi t)$ and it performs a 90° phase shift or quadrature filter to construct the so-called analytic $z(t)$ expressed by:

$$z(t) = y(t) + j\tilde{y}(t) \quad (2)$$

Equation (2) can also be written as follow:

$$z(t) = e(t) \cdot e^{j\theta(t)} \quad (3)$$

where

$e(t)$ is called the envelope signal of $y(t)$ and $\theta(t)$ is called the instantaneous phase signal of $y(t)$, we have the relations:

$$\begin{aligned} e(t) &= \sqrt{y^2(t) + \tilde{y}^2(t)} \\ \theta(t) &= \tan^{-1} \left[\frac{\tilde{y}(t)}{y(t)} \right] \end{aligned} \quad (4)$$

The envelope $e(t)$ depicts the energy distribution of $y(t)$ in the time domain.

In practice, the data are discretized in time, let:

$\underline{y}(k)$ be a discretized measurement vector at instant k from n_y PZT sensors, that are instrumented in the composite smart structure:

$$\underline{y}(k) = [y_1(k) \cdots y_i(k) \cdots y_{n_y}(k)]^T \quad (5)$$

The data matrix of measurements $\mathbf{Y} \in \mathbb{R}^{n_y \times N}$ gathering N samples $\underline{y}(k) (k = 1, \dots, N)$ is defined as follows:

$$\mathbf{Y} = [\underline{y}(1) \cdots \underline{y}(k) \cdots \underline{y}(N)] \quad (6)$$

In our case of study, we have $n_y = 8, N = 2^{16}, n_y \ll N$.

The matrix \mathbf{Y} has been autoscaled by subtracting the mean and dividing each line by its standard deviation.

For sensor i and instant k , the analytic signal $z_i(k)$, the envelope signal $e_i(k)$ and the instantaneous phase $\theta_i(k)$ are given by:

$$z_i(k) = y_i(k) + j\tilde{y}_i(k) \quad (7)$$

$$e_i(k) = \sqrt{y_i^2(k) + \tilde{y}_i^2(k)} \quad (8)$$

$$\theta_i(k) = \tan^{-1} \left[\frac{\tilde{y}_i(k)}{y_i(k)} \right] \quad (9)$$

Using Eq. (8), we define the envelope vector $\underline{e}(k)$ at instant k for the n_y sensor:

$$\underline{e}(k) = [e_1(k) \cdots e_i(k) \cdots e_{n_y}(k)]^T \quad (10)$$

For example, the corresponding envelope signal of sensor PZT 7 in the case of healthy and damaged structures are depicted in figure 5, only the 512 first samples of the envelope signals are displayed.

According to Eq. (10), we define the envelope matrix $\mathbf{E} \in \mathbb{R}^{n_y \times N}$ of the matrix measurements $\mathbf{Y} \in \mathbb{R}^{n_y \times N}$ by:

$$\mathbf{E} = [\underline{e}(1) \cdots \underline{e}(k) \cdots \underline{e}(N)] \quad (11)$$

This envelope matrix \mathbf{E} gathers N samples $\underline{e}(k)$, ($k = 1, \dots, N$):

3.2 Multivariate analysis

As stated in section 1, multivariate analysis concerns the analysis of multiple measurements from sensors and treats them as a single entity. In our work, the single entity concerns the envelope matrix $\mathbf{E} \in \mathbb{R}^{n_y \times N}$. One way to study the matrix \mathbf{E} is to use the singular value decomposition (SVD), (Golub 1983):

The matrix $\mathbf{E} \in \mathbb{R}^{n_y \times N}$ admits two orthogonal matrices:

$$\mathbf{U} = [\underline{u}_1, \dots, \underline{u}_{n_y}] \in \mathbb{R}^{n_y \times n_y} \quad (12)$$

$$\mathbf{V} = [\underline{v}_1, \dots, \underline{v}_{n_y}] \in \mathbb{R}^{n_y \times n_y}$$

such that

$$\begin{aligned} \mathbf{\Gamma} &= \mathbf{U}^T \cdot \mathbf{Y} \cdot \mathbf{V} = \text{diag}(\sigma_1, \dots, \sigma_p) \\ p &= \min\{n_y, N\} = n_y \\ \mathbf{U}^T \cdot \mathbf{U} &= \mathbf{I}_{n_y}, \mathbf{V}^T \cdot \mathbf{V} = \mathbf{I}_{n_y} \end{aligned} \quad (13)$$

where $\mathbf{\Gamma} \in \mathbb{R}^{n_y \times n_y}$ is the matrix of singular values, the columns of the matrix $\mathbf{U} \in \mathbb{R}^{n_y \times n_y}$ contain the left singular vectors and the columns of the matrix $\mathbf{V} \in \mathbb{R}^{n_y \times n_y}$ contain the right singular vectors.

The SVD of the matrix \mathbf{E} provides important insight about the orientation of this set of vectors, and determines how much the dimension of \mathbf{E} can be reduced, (Kerschen et al. 2005). One way to reduce the dimension of \mathbf{E} is to take the sum of all singular values then to delete those singular values that fall below some percentage of that sum, (De Boe and Golinval 2003). In our work, we have decided to fix a percentage sum of 98%.

According to this, the SVD of matrix \mathbf{E} take the following form:

$$\begin{aligned} \mathbf{E} &= [\mathbf{U}_1 \quad \mathbf{U}_2] \cdot \begin{bmatrix} \mathbf{\Gamma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Gamma}_2 \end{bmatrix} \cdot [\mathbf{V}_1 \quad \mathbf{V}_2]^T \\ &= \mathbf{E}_1 + \mathbf{E}_2 \end{aligned} \quad (14)$$

where:

$$\begin{aligned} \mathbf{U}_1 &\in \mathbb{R}^{n_y \times n_{POM}}, \mathbf{\Gamma}_1 \in \mathbb{R}^{n_{POM} \times n_{POM}}, \mathbf{V}_1 \in \mathbb{R}^{N \times n_{POM}}, \\ \mathbf{U}_2 &\in \mathbb{R}^{n_y \times (n_y - n_{POM})}, \mathbf{\Gamma}_2 \in \mathbb{R}^{(n_y - n_{POM}) \times (n_y - n_{POM})}, \\ \mathbf{V}_2 &\in \mathbb{R}^{N \times (n_y - n_{POM})}, \end{aligned}$$

n_{POM} is the retained dimension after reduction.

The columns of the matrix \mathbf{U}_1 are called the principal left singular vectors and the columns of the matrix \mathbf{V}_1 are called the principal right singular vectors. Analogously, the columns of the matrix \mathbf{U}_2 are called the residual left singular vectors and the columns of the matrix \mathbf{V}_2 are called the residual right singular vectors.

4. DAMAGE DETECTION METHODOLOGY

The presence of damage in the structure cause change in the stiffness and mass matrices. Consequently, damage will introduce change in the response of the measurement sensor and the matrix measurements \mathbf{Y} , see (Hajrya et al. 2011) for the demonstration. Hence, the envelope matrix \mathbf{E} is also modified. Figure 5 depicts the corresponding envelope signal of sensor PZT 7 and one can see that there is a significant difference in the envelope signal of the healthy and damaged structures.

4.1 Damage index

Let $\mathbf{E}^s, \mathbf{E}^u \in \mathbb{R}^{n_y \times N}$ be respectively the envelope matrices of the healthy and unknown structures. According to section 3.2, there SVD is defined as follow:

$$\begin{aligned} \mathbf{E}^s &= [\mathbf{U}_1^s \quad \mathbf{U}_2^s] \cdot \begin{bmatrix} \mathbf{\Gamma}_1^s & \mathbf{0} \\ \mathbf{0} & \mathbf{\Gamma}_2^s \end{bmatrix} \cdot [\mathbf{V}_1^s \quad \mathbf{V}_2^s]^T \\ &= \mathbf{E}_1^s + \mathbf{E}_2^s \end{aligned} \quad (15)$$

$$\begin{aligned} \mathbf{E}^u &= [\mathbf{U}_1^u \quad \mathbf{U}_2^u] \cdot \begin{bmatrix} \mathbf{\Gamma}_1^u & \mathbf{0} \\ \mathbf{0} & \mathbf{\Gamma}_2^u \end{bmatrix} \cdot [\mathbf{V}_1^u \quad \mathbf{V}_2^u]^T \\ &= \mathbf{E}_1^u + \mathbf{E}_2^u \end{aligned} \quad (16)$$

We suppose that the dimensions of all components in Eq. (15) and (16) are equals to those in Eq. (14).

In our methodology, we are interested in studying the principal left and right singular vectors.

Let:

$\mathbf{U}_1^s = [\mathbf{u}_{11}^s \cdots \mathbf{u}_{1i}^s \cdots \mathbf{u}_{1n_{POM}}^s] \in \mathbb{R}^{n_y \times n_{POM}}$, be the principal left singular vectors of the healthy smart structure,

$\mathbf{V}_1^s = [\mathbf{v}_{11}^s \cdots \mathbf{v}_{1i}^s \cdots \mathbf{v}_{1n_{POM}}^s] \in \mathbb{R}^{N \times n_{POM}}$, be the principal right singular vectors of the healthy smart structure,

$\mathbf{U}_1^u = [\mathbf{u}_{11}^u \cdots \mathbf{u}_{1i}^u \cdots \mathbf{u}_{1n_{POM}}^u] \in \mathbb{R}^{n_y \times n_{POM}}$, be the principal left singular vectors of the unknown smart structure,

$\mathbf{V}_1^u = [\mathbf{v}_{11}^u \cdots \mathbf{v}_{1i}^u \cdots \mathbf{v}_{1n_{POM}}^u] \in \mathbb{R}^{N \times n_{POM}}$, be the principal right singular vectors of the unknown smart structure.

We define the angle between \mathbf{u}_{1i}^s and \mathbf{u}_{1i}^u and the angle between \mathbf{v}_{1i}^s and \mathbf{v}_{1i}^u as, (De Boe and Golinval 2003):

$$\begin{aligned} |\cos \psi_i| &= |\langle \mathbf{u}_{1i}^s | \mathbf{u}_{1i}^u \rangle| = |(\mathbf{u}_{1i}^s)^T \cdot \mathbf{u}_{1i}^u| \\ \psi_i &= \cos^{-1} |\cos \psi_i|, \quad \psi_i \in \left[0, \frac{\pi}{2}\right] \\ |\cos \varphi_i| &= |\langle \mathbf{v}_{1i}^s | \mathbf{v}_{1i}^u \rangle| = |(\mathbf{v}_{1i}^s)^T \cdot \mathbf{v}_{1i}^u| \\ \varphi_i &= \cos^{-1} |\cos \varphi_i|, \quad \varphi_i \in \left[0, \frac{\pi}{2}\right] \end{aligned} \quad (17)$$

According to this, we define two angle vectors $\underline{\psi}$ and $\underline{\phi}$ by :

$$\underline{\psi} = [\psi_1 \cdots \psi_i \cdots \psi_{n_{POM}}]^T, \quad \underline{\phi} = [\varphi_1 \cdots \varphi_i \cdots \varphi_{n_{POM}}]^T$$

We propose the following new damage index DI:

$$DI = \sqrt{\|\sin \underline{\psi}\|_2^2 + \|\sin \underline{\phi}\|_2^2} \quad (18)$$

Theoretically, when the current state is healthy, then the damage index DI is null, but if the current state is damaged, then the damage index is different from zero.

In order to improve the damage detection methodology under experimental conditions, we define in the next subsection a bound associated to the DI and it is based on the work of Wedin, (Wedin 1972).

4.2 Definition of a bound for the damage index

Wedin have studied the perturbation of matrices using subspaces. Our contribution in this subsection is to extend the theoretical work developed by Wedin in the case of experimental SHM system.

Define first a new envelope matrix $\tilde{\mathbf{E}}^s \in \mathbb{R}^{n_y \times N}$ of the healthy smart structure:

$$\tilde{\mathbf{E}}^s = \mathbf{E}^s + \delta \mathbf{E}^s \quad (19)$$

where

$\delta \mathbf{E}^s \in \mathbb{R}^{n_y \times N}$ is a matrix which reflects the effect of noise in an experiment.

According to subsection 3.2, the SVD of \mathbf{E}^s and $\tilde{\mathbf{E}}^s$ are defined as follow:

$$\mathbf{E}^s = [\mathbf{U}_1^s \quad \mathbf{U}_2^s] \cdot \begin{bmatrix} \Gamma_1^s & \mathbf{0} \\ \mathbf{0} & \Gamma_2^s \end{bmatrix} \cdot [\mathbf{V}_1^s \quad \mathbf{V}_2^s]^T \quad (20)$$

$$= \mathbf{E}_1^s + \mathbf{E}_2^s$$

$$\tilde{\mathbf{E}}^s = [\tilde{\mathbf{U}}_1^s \quad \tilde{\mathbf{U}}_2^s] \cdot \begin{bmatrix} \tilde{\Gamma}_1^s & \mathbf{0} \\ \mathbf{0} & \tilde{\Gamma}_2^s \end{bmatrix} \cdot [\tilde{\mathbf{V}}_1^s \quad \tilde{\mathbf{V}}_2^s]^T \quad (21)$$

$$= \tilde{\mathbf{E}}_1^s + \tilde{\mathbf{E}}_2^s$$

Let $\underline{\psi}^s$ and $\underline{\phi}^s$ the two angle vectors, respectively between the left singular vectors of \mathbf{E}^s and $\tilde{\mathbf{E}}^s$ and the right singular vectors \mathbf{E}^s and $\tilde{\mathbf{E}}^s$, these angle vectors are calculated using Eq. (17).

According to (Wedin 1972), we define two residual matrices $\mathbf{R}_{11}, \mathbf{R}_{21}$ as:

$$\mathbf{R}_{11} = \mathbf{E}^s \cdot \tilde{\mathbf{V}}_1^s - \tilde{\mathbf{U}}_1^s \cdot \tilde{\Gamma}_1^s = (\mathbf{E}^s - \tilde{\mathbf{E}}^s) \cdot \tilde{\mathbf{V}}_1^s \quad (22)$$

$$= -\delta \mathbf{E}^s \cdot \tilde{\mathbf{V}}_1^s$$

$$\mathbf{R}_{21} = (\mathbf{E}^s)^T \cdot \tilde{\mathbf{U}}_1^s - \tilde{\mathbf{V}}_1^s \cdot \tilde{\Gamma}_1^s \quad (23)$$

$$= ((\mathbf{E}^s)^T - (\tilde{\mathbf{E}}^s)^T) \cdot \tilde{\mathbf{U}}_1^s$$

$$= -(\delta \mathbf{E}^s)^T \cdot \tilde{\mathbf{U}}_1^s$$

Given, the aforementioned definitions, Wedin's theorem states:

Theorem

If $\exists \alpha \geq 0$ and $\eta > 0$ such that

$$\min \sigma(\tilde{\mathbf{E}}_1^s) \geq \alpha + \eta \text{ and } \max \sigma(\tilde{\mathbf{E}}_2^s) \leq \alpha$$

And let $\mu = \max \sqrt{\|\mathbf{R}_{11}\|_2 + \|\mathbf{R}_{21}\|_2}$, then

$$\begin{cases} \|\sin \underline{\psi}^s\|_2 \leq \frac{\mu}{\eta} \\ \|\sin \underline{\phi}^s\|_2 \leq \frac{\mu}{\eta} \end{cases}$$

According to this theorem, we define a bound \mathcal{B} as:

$$\mathcal{B} = \sqrt{2} \frac{\mu}{\eta} \quad (24)$$

In order to improve the bound \mathcal{B} , we make n experimental tests of the healthy smart structure and we calculate the mean of the bound:

$$\mu_{\mathcal{B}} = \frac{1}{n} \sum_{j=1}^n \mathcal{B}_j \quad (25)$$

The detection procedure is as follow

If $DI < \mu_{\mathcal{B}}$ then the unknown smart structure is in healthy state,

Else the unknown smart structure is in damaged state.

To summarize the damage detection methodology, we use the following steps:

Damage detection methodology

1. Measure acquisition of the healthy smart structure \mathbf{Y}^s ,
2. Repeat n times the experiment for the healthy smart structure: $\tilde{\mathbf{Y}}_j^s, j = 1 \dots n$,
3. Center the data matrices $\mathbf{Y}^s, \tilde{\mathbf{Y}}_j^s$ and normalize them using the standard deviation,
4. Using Eq. (8) and (11), calculate the envelope matrix \mathbf{E}^s and $\tilde{\mathbf{E}}_j^s$,
5. Using Eq. (13), applied the SVD for matrices \mathbf{E}^s and $\tilde{\mathbf{E}}_j^s$,
6. Reduce the dimension if possible,
7. Using the Wedin' theorem and Eq.(24), calculate the bound $\mathcal{B}_j, j = 1 \dots n$,
8. Calculate the mean bound $\mu_B = \frac{1}{n} \sum_{j=1}^n \mathcal{B}_j$,
9. Measure acquisition of the unknown smart structure \mathbf{Y}^u ,
10. Center the data matrix \mathbf{Y}^u and normalize it using the standard deviation,
11. Using Eq. (8) and (11), calculate the envelope matrix \mathbf{E}^u ,
12. Using Eq. (13), applied the SVD for the matrix \mathbf{E}^u ,
13. Reduce the dimension if possible,
14. Using Eq. (17), calculate $\cos \underline{\psi}$ and $\cos \underline{\phi}$
15. Calculate $\sin \underline{\psi}$ and $\sin \underline{\phi}$,
16. Using Eq. (18), calculate the damage index DI between the healthy envelope matrix \mathbf{E}^s and the unknown envelope matrix \mathbf{E}^u ,
17. If :
 $DI < \mu_B$: Then the unknown smart structure is in healthy state,
 Else the unknown smart structure is in damaged state.

5. APPLICATION TO THE COMPOSITE SMART STRUCTURE

The damage detection methodology described previously is applied to detect the impact damage of the composite plate presented in section 2. In the first step of our application, we were interested by using PZT 10 as an actuator while the others PZT are sensors (PZT 6 is not taken into account in the damage detection). Following the methodology developed, we have performed six measurements for the healthy composite plate and one measurement for the damaged composite plate. Using these measurement matrices, the envelope matrix for each healthy and damaged state was

calculated. Before the calculation of the damage index DI and its associated bound \mathcal{B} , we have search for each state of the composite plate to reduce the dimension of the envelope matrices. According to the 98% percentage sum of singular value fixed in subsection 3.2, we see using figures 6 and 7 that the dimension of the envelope matrices cannot be reduced, those the dimension remain: $\mathbf{E}^s, \mathbf{E}^u \in \mathbb{R}^{8 \times 2^{16}}$. Using the six experiments of the healthy composite state, the mean value of the bound was first calculated: $\mu_B = 0.40$. The damage index between the healthy and damaged composite plates defined in Eq. (18) is: $DI = 3.37$. One can we see that the DI is upper than the mean value of the bound, then damage is detected. In order to illustrate the efficiency of the damage detection methodology in term of false alarms, we have done another experiment of the healthy structure which is strictly independent from the others done previously, in this case, $DI = 0.26$ and it is lower than $\mu_B = 0.4042$.

In second step of our application, we have used PZT 7 as actuator, according to the same methodology, we have obtained the result depicted in table 3 a damage index $DI = 3.30$, one can we see that the DI is upper than the mean value of the bound $\mu_B = 0.54$. No false alarms were detected.

6. CONCLUSION

In this paper, a damage detection methodology was developed to enhance feature information about damage. This methodology is based on the calculation of a damage index which consists on comparing subspaces of the healthy and damaged state of envelope matrix. This DI was associated with a bound.

The efficiency of the proposed approach was successively applied to detect experimentally impact damage in the composite smart plate. The proposed method presents a cheap computational cost and seems to be well adapted for structural health monitoring in real time application.

For the work under progress, we are investigating the localization of the impact damage in damaged composite plate.

ACKNOWLEDGMENT

This work was supported by the MSIE-ASTech Paris Région of the French governmental research program and the authors gratefully acknowledge them for providing the composite structures.

MAIN TERMS

- $\tilde{y}(t)$ Hilbert transform of signal $y(t)$
- $z(t)$ Analytic signal
- $e(t)$ Envelope signal
- $\theta(t)$ Instantaneous phase signal

- n_y Number of sensors in the composite smart structure
- N Number of samples
- $\underline{y}(k)$ Measurements vector at instant k
- \mathbf{Y} Matrix measurements
- $\underline{e}(k)$ Envelope vector
- \mathbf{E} Envelope matrix
- \mathbf{E}^s Envelope matrix of the healthy structure
- $\tilde{\mathbf{E}}^s$ Envelope matrix of a second experiment of the healthy structure
- \mathbf{E}^u Envelope matrix of the unknown structure
- \mathbf{U} Matrix of left singular vectors
- \mathbf{V} Matrix of right singular vectors
- $\underline{\psi}$ Angle vector between the left singular vectors of the healthy matrix \mathbf{E}^s and unknown matrix \mathbf{E}^u ,
- $\underline{\Phi}$ Angle vector between the right singular vectors of the healthy matrix \mathbf{E}^s and unknown matrix \mathbf{E}^u ,
- $\underline{\psi}^s$ Angle vector between the left singular vectors of the two healthy matrices \mathbf{E}^s $\tilde{\mathbf{E}}^s$
- $\underline{\Phi}^s$ Angle vector between the right singular vectors of the two healthy matrices \mathbf{E}^s $\tilde{\mathbf{E}}^s$
- DI Damage index
- B Bound of the damage index
- μ_B Mean value of the damage index
- $\delta\mathbf{Y}^s$ Matrix of noise
- \mathbf{Y}^T Transpose of matrix \mathbf{Y}
- j Imaginary number
- \mathbb{R} Set of real number



Figure 2 Healthy composite plate bonded with ten PZT patches

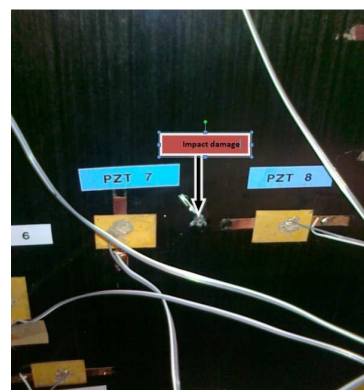


Figure 3: Impact damage in the composite structure

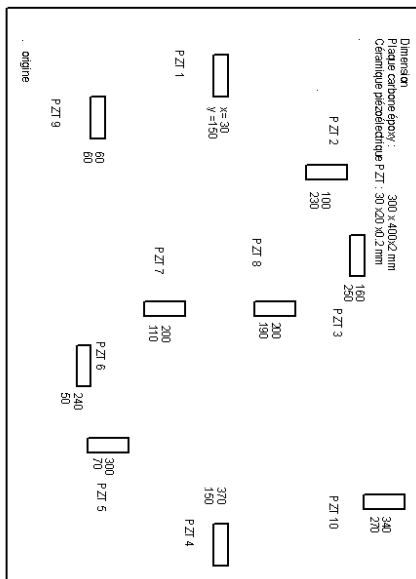


Figure 1: Placement of the PZT in the composite plate

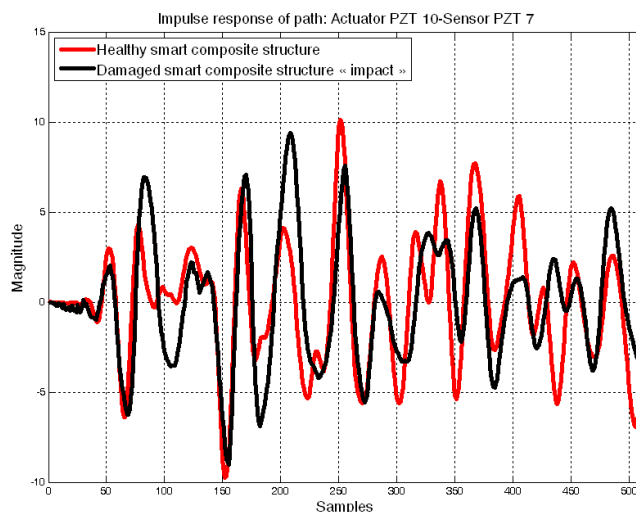


Figure 4: Impulse response of the healthy and damaged smart structures path: actuator PZT 10-sensor PZT7

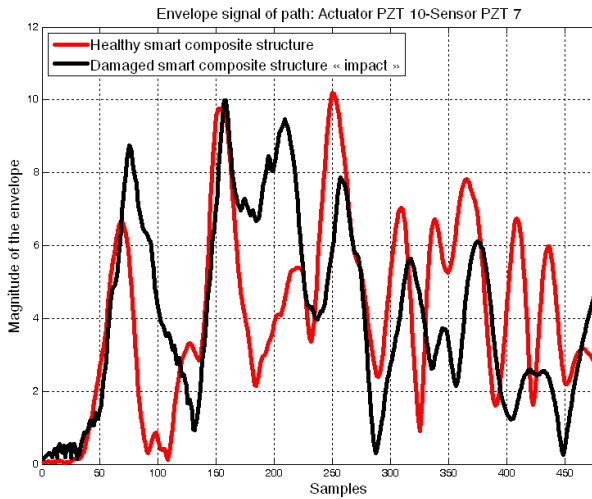


Figure 5: Envelope signal of the healthy and damaged structures: path: Actuator PZT 10-Sensor PZT7

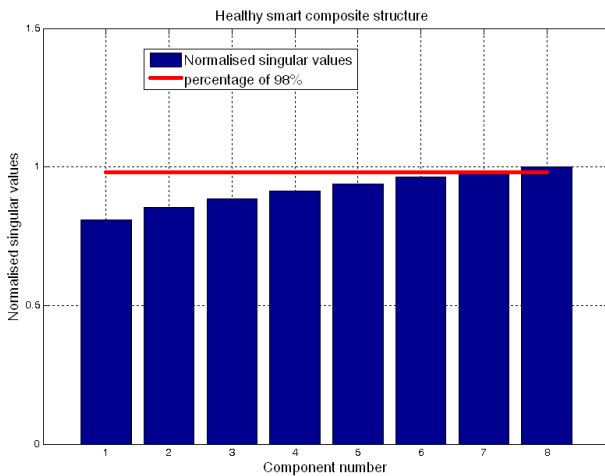


Figure 6: Order reduction of the healthy smart structure

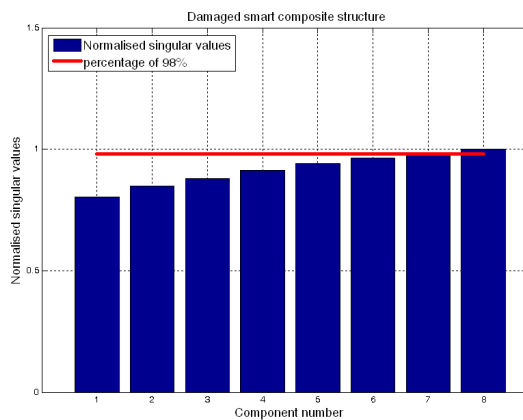


Figure 7: Order reduction of the damaged smart structure

Table 1 Mechanical property of the carbone-epoxy composite plate

Property	E_1	$E_2 = E_3$	$G_{12} = G_{13}$	G_{23}	$\nu_{12} = \nu_{13}$	ν_{23}	ρ
Unit	Gpa	Gpa	Gpa	Gpa	-	-	Kg/m ³
Value	127.7	7.217	5.712	2.614	0.318	0.38	1546

Table 2 Result of the damage detection in the case of the use of actuator PZT 10

	DI_{POD}	B
Damage plate	3.37	0.4042
Safe plate	0.2602	0.4042

Table 3 Result of the damage detection in the case of the use of actuator PZT 7

	DI_{POD}	B
Damage plate	3.3056	0.5374
Safe plate	0.2190	0.5374

REFERENCES

Bendat, J. S., and Piersol, A. G. (2000). "Random data: analysis and measurement procedures." New York, NY: Wiley-Interscience.

De Boe, P., and Golinval, J. C. (2003). "Principal component analysis of a piezosensor array for damage localization." *Structural Health Monitoring*, 2(2), 137-144.

Giurgiutiu, V., Zagari, A., and Bao, J. J. (2002). "Piezoelectric wafer embedded active sensors for aging aircraft structural health monitoring." *Structural Health Monitoring*, 1(1), 41-61.

Golub, G. H. V. L. C. F. (1983). "Matrix computation." Johns Hopkins University Press, Baltimore.

Hajrya, R., Mechbal, N., and Vergé, M. (2010). "Active Damage Detection and Localization Applied to a Composite Structure Using Piezoceramic Patches." *Conference on Control and Fault-Tolerant Systems*. Nice, France.

Hajrya, R., Mechbal, N., and Vergé, M. (2011). "Proper Orthogonal Decomposition Applied to Structural Health Monitoring." *IEEE International Conference on Communications, Computing and Control Applications*. Hammamet, Tunisia.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Snin, H. H., Zheng, Q., Yen, N. C., Tung, C. C., and Liu, H. H. (1998). "The empirical mode decomposition and the Hubert spectrum for nonlinear and non-stationary time series analysis." *Proceedings of the*

- Royal Society A: Mathematical, Physical and Engineering Sciences, 454(1971), 903-995.
- Kerschen, G., Golinval, J. C., Vakakis, A. F., and Bergman, L. A. (2005). "The method of proper orthogonal decomposition for dynamical characterization and order reduction of mechanical systems: An overview." *Nonlinear Dynamics*, 41(1-3), 147-169.
- Staszewski, W. J. (2002). "Intelligent signal processing for damage detection in composite materials." *Composites Science and Technology*, 62(7-8), 941-950.
- Su, Z., Wang, X., Chen, Z., Ye, L., and Wang, D. (2006). "A built-in active sensor network for health monitoring of composite structures." *Smart Materials and Structures*, 15(6), 1939-1949.
- Wedin, P. (1972). "Perturbation Bounds in Connection with Singular value Decomposition." *Numerical Mathematics*, 12(1), 99-111.
- Yang, J. N., Lei, Y., Lin, S., and Huang, N. (2004). "Hilbert-Huang based approach for structural damage detection." *Journal of Engineering Mechanics*, 130(1), 85-95.
- Yang, J. N., Lei, Y., Pan, S., and Huang, N. (2003a). "System identification of linear structures based on Hilbert-Huang spectral analysis. Part 1: Normal modes." *Earthquake Engineering and Structural Dynamics*, 32(9), 1443-1467.
- Yang, J. N., Lei, Y., Pan, S., and Huang, N. (2003b). "System identification of linear structures based on Hilbert-Huang spectral analysis. Part 2: Complex modes." *Earthquake Engineering and Structural Dynamics*, 32(10), 1533-1554.
- Zang, C., Friswell, M. I., and Imregun, M. (2004). "Structural damage detection using independent component analysis." *Structural Health Monitoring*, 3(1), 69-83.

AUTHORS BIOGRAPHIES

Rafik Hajrya was born in Algeria, on 07 November 1984; actually, he is a PhD student at the laboratory of Processes and Engineering in Mechanics and Materials (PIMM-UMR CNRS) of Arts et Métiers ParisTech (Paris, France). He obtained: a Diploma in electronic-control engineering at the University of USTHB (Algiers, Algeria), a Master degree in Robotic and intelligent system at the University of Paris VI (Paris-France). His research interests focus on structural health monitoring of composite smart structure in particular damage detection using advanced signal processing.

Nazih Mechbal was born in Morocco, on 18 March 1971, he is an associate professor at the laboratory of Processes and Engineering in Mechanics and Materials (PIMM-UMR CNRS) at the engineering school Arts et Métiers ParisTech (ENSAM) of Paris, where he is member of the control and supervising team. He received his PhD degree in robotics from the ENSAM Paris in 1999. His research interests include structural health monitoring, robust fault detection and diagnosis, active control and robotics.

Michel Vergé was born in France, on 09 July 1950; he is a professor on control and supervising at the laboratory of Processes and Engineering in Mechanics and Materials (PIMM-UMR CNRS) of Arts et Métiers ParisTech (Paris, France). He obtained HDR at Nancy University (France) in 1991. His research interests focus on the fault detection methods and structural health monitoring.

Fault Detection in Non Gaussian Problems Using Statistical Analysis and Variable Selection

João P. P. Gomes¹, Bruno P. Leão¹, Roberto K. H. Galvão² and Takashi Yoneyama²

¹ *EMBRAER, São Jose dos Campos, São Paulo, 12227-901, Brazil*

joao.pordeus@embraer.com.br

bruno.leao@embraer.com.br

² *ITA – Instituto Tecnológico de Aeronáutica, São José dos Campos, São Paulo, 12228-900, Brazil*

kawakami@ita.br

takashi@ita.br

ABSTRACT

This work concerns the problem of fault detection using data-driven methods without the assumption of gaussianity. The main idea is extend the Runger's U^2 statistical distance measures to the case where the monitored variables are not gaussian. The proposed extension is based on Gaussian Mixture Models and Parzen windows classifiers to estimate the required conditional probability distributions. The proposed methodology was applied to an APU dynamic model and showed better results when compared to classical fault detection techniques using Multivariate Statistical Process control with Hotelling's T^2 metrics.*

1. INTRODUCTION

Data-driven methods comprise a powerful set of tools for performing failure prognosis and diagnosis. Such group of methods includes clustering and classification techniques, where the data is divided into groups on the basis of some specific distance measure (Duda et al., 2001). Statistical measures are a usual choice for such methods. The origins of clustering and classification methods based on statistical measures may be linked to the works of Mahalanobis (1936) and Hotelling (1933), which are related, respectively, to the Mahalanobis distance (MD) (De Maesschalck et al., 2000) and the T^2 statistic (Kourti and MacGregor, 1995). Such statistical distance measures are the basis of Multivariate Statistical Process Control (MSPC), which consists of a group of multivariate analysis techniques that can be used in health monitoring and diagnosis in industrial

environment, such as chemical plants (Kourti and MacGregor, 1995) and mining enterprises (Yacher and Orchard, 2003). Statistical measures are also employed in other fields of knowledge such as image processing and pattern recognition (Webb, 2002). In these fields, similar concepts are used for the definition of the Gaussian. Both in the Gaussian classifier and in the MSPC techniques, an usual assumption is to consider the underlying joint distributions of the monitored variables as Gaussian (or at least can be approximated as). Literature presents various examples of the use of such types of methods for Prognostics and Health Management (PHM): Kumar et al. (2008) present the use of MD for monitoring electronic systems; Mimmagh et al. (2000) present the use of Hotelling's T^2 statistic for the diagnostics of a helicopter drive system; Leão et al. (2009) show the application of MSPC for monitoring the health of electro-mechanical systems.

The abovementioned methods may provide poor performance when the gaussianity assumption is not verified. Since many practical problems do not satisfy such gaussianity assumption, extensions of these methods have been proposed to address non-Gaussian problems (Webb, 2002). One solution of this type is the use of Gaussian Mixture Models (GMM) to approximate the joint probability density of the variables of interest. Another possibility is the use a composition of Gaussian kernels for approximating the joint density in a non-parametric way. Such method is commonly referred as Parzen windows.

This work presents extensions to GMM or Parzen windows classifiers, which can provide better results for PHM solutions. Such extensions are inspired by the U^2 statistical distance (Runger, 1996), which was introduced by Runger in the context of MSPC (on the assumption of Gaussian joint distributions). Runger's

* Gomes, J. P. P. et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

U^2 statistic is based on the division of the set of monitored variables (\mathbf{x}) into two subsets: the first one (\mathbf{y}) includes variables which are affected by the failure under consideration; the second one (\mathbf{z}) encompasses variables which are not affected by the failure but are correlated to the variables of the first subset. Examples of the latter include operational and environmental conditions. The proposed methods are extensions of Runger's work to non-Gaussian problems, which are based on the calculation of the conditional likelihood $p(\mathbf{y}|\mathbf{z})$ using the densities estimated through GMM or Parzen windows.

In order to illustrate the use of the proposed methods, a simulation model of an aircraft Auxiliary Power Unit (APU) is employed. Different failure modes are simulated using such model and the proposed methods are used for failure diagnosis. Their performance is compared to that of the aforementioned traditional methods described in literature.

The remaining sections are organized as follows: section 2 describes the theoretical background associated to MSPC, Gaussian, GMM and Parzen windows classifiers; section 3 presents the novel methods proposed in this work; section 4 presents the simulated tests and results and section 5 is the conclusion.

2. THEORETICAL BACKGROUND

2.1 Statistical Distances, MSPC and the Gaussian Classifier

The application of statistical theory for fault detection relies on the assumption that the characteristics of the data variations are relatively unchanged unless a fault occurs in the system. This is a reasonable assumption under the definition of a fault as an abnormal process condition. It implies that the statistical properties of the data are repeatable for the same operating conditions, although the actual values may not be predictable (Chiang et al, 2001). The repeatability of the statistical properties allows the use of statistical measures, based on statistical distances, for the detection of abnormal behaviors on a process.

Eq. (1) presents the well known Mahalanobis Distance (MD):

$$M(\mathbf{x}) = (\mathbf{x} - \mu) \Sigma^{-1} (\mathbf{x} - \mu)^T \quad (1)$$

where \mathbf{x} is the feature vector associated to an observation and μ and Σ are respectively the mean values and the covariance matrix of a given dataset. These statistical properties can be estimated as

$$\mu = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \quad (2)$$

and

$$\Sigma = \frac{1}{n-1} \sum_{i=1}^n [(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T] \quad (3)$$

where $\{\mathbf{x}_i, i = 1, 2, \dots, n\}$ is a given set of observations.

The MD is used to define one of the most popular MSPC methods called Hotelling's T^2 statistic. In Hotelling's T^2 statistic, a statistical model is built using Eq. (2) and Eq. (3) given a dataset \mathbf{X} containing n instances of a feature vector \mathbf{x} . Each instance is composed by k monitored variables. All feature vectors in \mathbf{X} are obtained for a healthy system (without faults).

After this training stage, the MD is calculated for each new instance \mathbf{x}_{new} and the result is compared with a threshold in order to detect anomalies. Hotelling's T^2 statistic is defined as:

$$T^2 = (\mathbf{x}_{new} - \mu) \Sigma^{-1} (\mathbf{x}_{new} - \mu)^T \quad (4)$$

where μ and Σ are estimated using Eq. (2) and Eq. (3) for the dataset \mathbf{X} and \mathbf{x}_{new} is a new instance of the feature vector \mathbf{x} that needs to be classified as healthy or faulty.

The same principles are involved in a popular classification method, mainly employed in the pattern recognition literature, the Quadratic Gaussian Classifier (QGC, or simply Gaussian Classifier). For fault detection, the QGC can be formulated to solve the problem of one class classification, that is, to classify the operation of a system as healthy or not.

Using Bayes' theorem one could obtain the probability of a system being healthy ($H=1$) given a feature vector \mathbf{x} according to Eq. (5).

$$P(H = 1 | \mathbf{x}) = \frac{p(\mathbf{x} | H = 1)P(H = 1)}{p(\mathbf{x})} \quad (5)$$

Since the unconditional probability density $p(\mathbf{x})$ is not related to the health state of the system, it is not useful to decide if the system operation is healthy or faulty ($H=0$). Therefore, it can be ignored in the statistical measure.

Considering the prior probabilities of the system being healthy or faulty ($P(H=1)$ or $P(H=0)$) are not affected by the \mathbf{x} , these can be also be ignored, resulting on:

$$P(H = 1 | \mathbf{x}) \propto p(\mathbf{x} | H = 1) \quad (6)$$

Assuming $p(\mathbf{x}|H=1)$ to be a Gaussian distribution one can use the following statistical measure to detect anomalies based on \mathbf{x}_{new} .

$$p(H = 1 | \mathbf{x}_{new}) \propto \frac{1}{(2\pi)^k |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_{new} - \mu) \Sigma^{-1} (\mathbf{x}_{new} - \mu)^T\right] \quad (7)$$

It is possible to notice in Eq. 7 the presence of a term identical to that of expression Eq. 1, which is the

MD. It is the only term that depends on \mathbf{x}_{new} , so that the result is similar to that presented by Hotelling's T^2 statistic.

An MSPC method proposed by (Runger, 1996) provides relevant improvements in Hotelling's T^2 statistic and QGC. In many real applications it is possible that only a subset of the monitored variables are affected by a failure. The main idea in this method is to restrict the analysis to these fault-sensitive variables but without excluding the influences of other non fault-sensitive variables in system behavior.

Consider an observation \mathbf{x}_i of the feature vector:

$$\mathbf{x}_i = [\mathbf{x}_{i1} \quad \dots \quad \mathbf{x}_{ij} \quad \dots \quad \mathbf{x}_{ik}] \quad (8)$$

Assuming that a fault only affects variables \mathbf{x}_{i1} up to \mathbf{x}_{ij} , one can divide \mathbf{x}_i into two sets:

$$\mathbf{y}_i^T = [\mathbf{x}_{i1} \quad \dots \quad \mathbf{x}_{ij}] \quad (9)$$

and

$$\mathbf{z}_i^T = [\mathbf{x}_{i(j+1)} \quad \dots \quad \mathbf{x}_{ik}] \quad (10)$$

where \mathbf{y}_i contains the features that are sensitive to an incipient failure and \mathbf{z}_i contains those that are not sensitive to the failures.

The idea of Runger's U^2 statistic is to calculate Hotelling's T^2 statistic and subtract the influence of \mathbf{z}_i in the final calculated distance while keeping \mathbf{z}_i influence in \mathbf{y}_i behavior.

Runger's U^2 statistic can be defined as:

$$U^2 = T^2 - (\mathbf{z}_{new} - \mu_z)^T \sum_z^{-1} (\mathbf{z}_{new} - \mu_z) \quad (11)$$

It can be noticed that the MD is employed to compare \mathbf{z}_{new} with a statistical model built using a subset of \mathbf{X} comprising only the variables not affected by faults. The result is subtracted from Hotelling's T^2 statistic.

2.2 GMM and Parzen Windows Classifiers

All methods presented in section 2.1 have the assumption that the healthy data follows a Gaussian distribution. However, that assumption may be invalid in some real applications.

In order to overcome this problem, many authors have proposed methodologies mostly based on the usage of nonparametric estimators for the distribution of the healthy data (Webb, 2002), (Duda et al., 2001). With that estimation, it is possible to approximate $p(\mathbf{x}_{new}|H=1)$. In these cases, the statistical measure can be defined by Eq. (6) with no need for assumptions on the particular distribution for the data.

Two of the most commonly used nonparametric estimation methods are the Parzen windows and the GMM.

Parzen windows is a non parametric estimator based on the idea of approximating the distribution to be estimated by a superposition of kernel functions centered on each of the \mathbf{x}_i samples in \mathbf{X} . Based on that, and using the formulation presented in Eq. (6), it is possible to estimate $p(\mathbf{x}_{new}|H=1)$ according to Eq. (12).

$$p(\mathbf{x}_{new} | H=1) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{\mathbf{x}_{new} - \mathbf{x}_i}{h}\right) \quad (12)$$

where h is the smoothing parameter and $K(\cdot)$ is the kernel function, chosen *a priori*.

One important drawback when applying Parzen windows is the curse of dimensionality occurring when dealing with high dimension data. In these cases, a limited number of data vectors can result in a sparse dataset which could difficult the task of distribution estimation. One way to overcome this problem is the application of so called semi-parametric estimators such as GMM.

The GMM approach models the distribution to be estimated as a composition of a set of weighted Gaussian distributions. The general expression for $p(\mathbf{x}_{new}|H=1)$ can be written as:

$$p(\mathbf{x}_{new} | H=1) = \sum_{l=1}^m \pi_l G(\mathbf{x}_{new}, \theta_l) \quad (13)$$

where π_l are the weights of each of the l Gaussian models whose parameters μ_l and Σ_l are expressed in θ_l .

The values for parameters π_l and θ_l can be obtained according to the Expectation-Maximization algorithm as presented originally in (Dempster et al, 1977).

Implementation details and information about other nonparametric estimators based classifiers can be found in many references such as (Webb, 2002) and (Duda et al., 2001).

3. PROPOSED METHOD

Although some authors proposed methodologies to monitor systems that provide non-Gaussian data, no previous work exploited the differentiation of variables which are affected by failure from those that are not. This feature was the main contribution in the development of Runger's U^2 when compared to Hotelling's T^2 in the Gaussian case.

Using the same definitions presented in Eq. (8), Eq. (9) and Eq. (10), one can rewrite Eq. (5) as:

$$P(H=1 | \mathbf{y}, \mathbf{z}) = \frac{p(\mathbf{y}, \mathbf{z} | H=1)P(H=1)}{p(\mathbf{y}, \mathbf{z})} \quad (14)$$

The joint distribution of \mathbf{y} and \mathbf{z} can be rewritten in terms of the conditional probability. This substitution leads to.

$$P(H=1 | \mathbf{y}, \mathbf{z}) = \frac{p(\mathbf{y} | \mathbf{z}, H=1)p(\mathbf{z} | H=1)P(H=1)}{p(\mathbf{y}, \mathbf{z})} \quad (15)$$

The hypothesis assumed for Runger U^2 is that the fault only affects the subset of the feature vector defined by \mathbf{y} . This hypothesis can be reformulated by saying that the distribution of \mathbf{z} does not change whenever the system is healthy or faulty. In this case we have:

$$p(\mathbf{z} | H = 1) = p(\mathbf{z}) \quad (16)$$

With Eq. (15) and Eq. (16) we have:

$$P(H = 1 | \mathbf{y}, \mathbf{z}) = \frac{p(\mathbf{y} | \mathbf{z}, H = 1) p(\mathbf{z}) P(H = 1)}{p(\mathbf{y}, \mathbf{z})} \quad (17)$$

The joint probability $p(\mathbf{y}, \mathbf{z})$ can be expressed by:

$$p(\mathbf{y}, \mathbf{z}) = p(\mathbf{y} | \mathbf{z}) p(\mathbf{z}) \quad (18)$$

Resulting in:

$$P(H = 1 | \mathbf{y}, \mathbf{z}) = \frac{p(\mathbf{y} | \mathbf{z}, H = 1) p(\mathbf{z}) P(H = 1)}{p(\mathbf{y} | \mathbf{z}) p(\mathbf{z})} \quad (19)$$

That leads to:

$$P(H = 1 | \mathbf{y}, \mathbf{z}) = \frac{p(\mathbf{y} | \mathbf{z}, H = 1) P(H = 1)}{p(\mathbf{y} | \mathbf{z})} \quad (20)$$

Using the same simplification procedures described for the Gaussian case, it yields:

$$P(H = 1 | \mathbf{y}, \mathbf{z}) \propto p(\mathbf{y} | \mathbf{z}, H = 1) \quad (21)$$

The conditional probability of \mathbf{y} given \mathbf{z} in a healthy system can be obtained by:

$$P(H = 1 | \mathbf{y}, \mathbf{z}) \propto \frac{p(\mathbf{y}, \mathbf{z} | H = 1)}{p(\mathbf{z})} \quad (22)$$

For the classification of \mathbf{x}_{new} as healthy or faulty we have:

$$P(H = 1 | \mathbf{x}_{new}) \propto \frac{p(\mathbf{y}_{new}, \mathbf{z}_{new} | H = 1)}{p(\mathbf{z}_{new})} \quad (23)$$

where $p(\mathbf{y}_{new}, \mathbf{z}_{new} | H = 1)$ and $p(\mathbf{z}_{new})$ can be estimated using any nonparametric estimation method as the ones presented in section 2.2.

Analyzing the result obtained in Eq. (23) it may be noticed that the basis of the proposed method is to estimate the conditional probability of \mathbf{y} given \mathbf{z} instead of the joint probability of \mathbf{y} and \mathbf{z} as presented in Eq. (6). The new method is expected to provide greater sensitivity and therefore better performance for fault diagnosis and health monitoring applications.

4. SAMPLE APPLICATION

To demonstrate how the proposed method could be applied in a real system and to compare the results against some classical methods, a sample application will be presented. The application consists of the detection of faults in an Auxiliary Power Unit (APU).

The Auxiliary Power Unit is a gas turbine device on a vehicle with the purpose of providing power to other systems when main engines are turned off. This power can be either pneumatic, obtained through the bleeding of compressed air, or electrical, obtained by coupling a generator to the APU shaft. They are commonly found on medium and large aircraft, as well as some large land vehicles. Its primary purpose is usually to provide bleed air to start the main engines. It is also used to run accessories such as air conditioning units and hydraulic pumps. A simplified APU representation is illustrated in Figure 1.

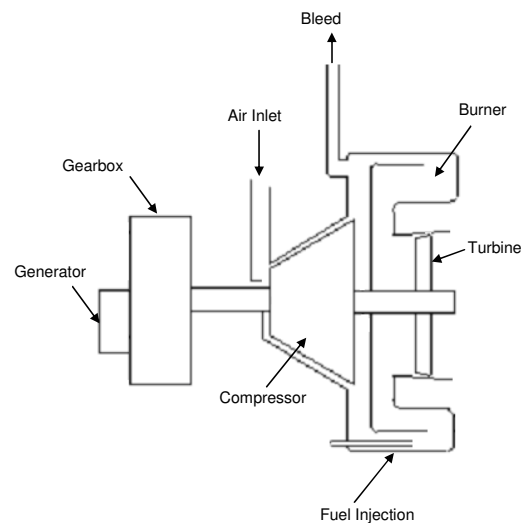


Figure 1: A simplified representation of an APU

In order to provide data for the APU fault detection, a mathematical model was developed using Matlab/Simulink. Figure 1 shows a schematic view of the mathematical model developed. The main modeled blocks are represented in Figure 2.

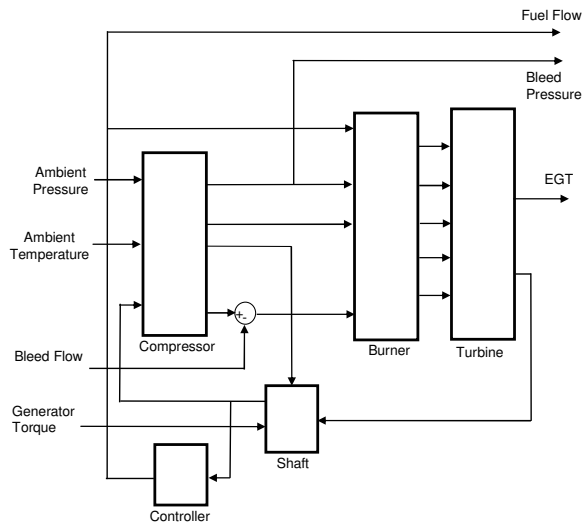


Figure 2: Simulation blocks and their relations in the developed APU model.

In the model, the compressor, burner and turbine model blocks were designed according to thermodynamic principles and information provided by nonlinear maps that describe the behavior of a real APU. The controller block comprises the control of shaft speed using fuel flow. The shaft block receives torque values from mechanical loads coupled to the APU shaft and calculates shafts rotation.

The main variables present in the model and used as measurements in the sample applications are the Exhaust Gas Temperature (EGT), Bleed Pressure (BP) and Fuel Flow (FF). The influences of ambient pressure (P_{amb}) and ambient temperature (T_{amb}) were also modeled.

Six different failure modes were seeded into the model, one at a time. These failure modes are:

- Bleed pressure sensor bias
- Fuel flow sensor bias
- Shaft speed sensor bias
- Exhaust gas temperature sensor bias
- Loss of compressor efficiency
- Loss of turbine efficiency

For the present study, four different fault detection methodologies were applied. Hotelling’s T^2 , Runger U^2 , the GMM classifier (GMMC) and the GMM classifier with selection of variables of interest (GMMC-SV). The GMMC-SV classifier is the proposed method described in section 3 using GMM to estimate $p(y_{new}, z_{new} | H=1)$ and $p(z_{new})$.

The feature vectors comprised the steady state values for EGT, BP, FF, T_{amb} and P_{amb} during a simulation of APU startup. Considering Eq. (23), values of EGT, BP and FF were selected to form y and T_{amb} and P_{amb} composed vector z . All signals were corrupted with gaussian noise.

To characterize the behavior of the APU without faults, 1,000 simulations of APU startups were performed for different condition of pressure and ambient temperature. Both ambient conditions were simulated as following Gaussian distributions.

In Hotelling’s T^2 and Runger’s U^2 the generated dataset was used to estimate the mean vectors and covariance matrices presented in Eq. (4) and Eq. (11) respectively. In GMMC and GMMC-SV the distributions were estimated using a composition of five Gaussian distributions. The weights and parameters of each Gaussian were estimated using the EM algorithm.

For the generation of the test dataset, 12,000 simulations were performed, being 6,000 simulations of a healthy system and 6,000 simulations of the system with fault in different levels of severity (1,000 simulations for each failure mode).

To verify the performance of each method the Receiver Operating Characteristic (ROC) curve was used. The ROC curve was generated by varying the fault detection threshold and collecting false alarm rate and correct detection rate for each of the methods. In a ROC curve it is possible to classify the performance of the methods by evaluating the area under the curve. Better methods yield greater areas under the curve.

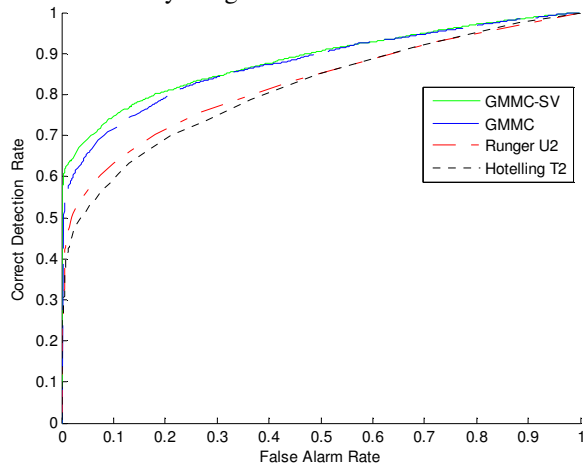


Figure 3: ROC curves for the implemented fault detection methods

The area under the ROC curve for each method is presented in Table 1.

Table 1: Area under ROC curve

Method	Hotelling’s T^2	Runger U^2	GMMC	GMMC-SV
Area	0.838	0.860	0.875	0.884

It is possible to notice a significant difference in the performance when comparing methods that assume gaussianity in the data and methods that do not rely on that assumption. Methods that can deal with non-

Gaussian distributions achieve better scores. In the present case, such a finding can be ascribed to the nonlinearity of the simulation model employed to generate the data.

Analyzing Gaussian and non Gaussian methods separately, one can notice a superior performance of the methods where the subset of the monitored variables are selected. This result was already mentioned in (Runger, 1996) for Gaussian data and was extended for non Gaussian data in this work. The proposed GMMC-SV presented the better performance overall.

5. CONCLUSION

This work presented a novel data-driven methodology for fault detection. The concept of anomaly detection in a subset of the monitored variables proposed by (Runger, 1996) was extended to the case where the monitored variables do not followed a Gaussian distribution.

The method was tested using an APU dynamic model and showed better results when compared to classical fault detection methods.

REFERENCES

- Chiang L. H., Russel E. L. and Braatz R. D. (2001) *Fault Detection and Diagnosis in Industrial Systems*. 1st ed. Springer-Verlag London.
- Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. 2nd ed. New York: Wiley.
- De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D.L. (2000). The Mahalanobis Distance, *Chemometrics and Intelligent Laboratory Systems*, 50, 1–18, 2000.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B*, 39, 1–38.
- Hottelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components, *Journal of Educational Psychology*, 24, 498–520.
- Kourti, T. and MacGregor, J. F. (1995). Process Analysis, Monitoring and Diagnosis, Using Multivariate Projection Methods”, *Chemometrics and Intelligent Laboratory Systems* 28, 3–21.
- Kumar, S., Sotiris, V., and Pecht, M. (2008), Mahalanobis Distance and Projection Pursuit Analysis for Health Assessment of Electronic Systems, in *Proceedings IEEE Aerospace Conference*, Big Sky, MO.
- Leão, B. P., Gomes, J. P. P., Galvão, R. K. H., and Yoneyama, T. (2009). Aircraft Flap and Slat Systems Health Monitoring Using Statistical Process Control Techniques, in *Proceedings of IEEE Aerospace Conference*, Big Sky, MO.
- Mahalanobis, P. C. (1936). On the Generalized Distance in Statistics, *Proceedings of the National Institute of Science of India*, 12, 49–55.
- Minnagh, M. L., Hardman, W., and Sheaffer, J. (2000), Helicopter Drive System Diagnostics Through Multivariate Statistical Process Control, in *Proceedings IEEE Aerospace Conference*, Big Sky, MO.
- Runger, G. C. (1996). Projections and the U2 Multivariate Control Chart”, *Journal of Quality Technology*, 28, 313–319.
- Webb, A (2002), *Statistical Pattern Recognition*. 2nd ed. West Sussex: John Wiley and Sons Ltd.
- Yacher, L., and Orchard, M. (2003), Statistical Multivariate Analysis and Dynamics Monitoring for Plant Supervision Improvement, in *Proceedings Copper International Conference*.
- João Paulo Pordeus Gomes** holds a bachelor’s degree on Electrical Engineering (2004) from Universidade Federal do Ceará (UFC), Brazil, and Master Degree on Aeronautical Engineering (2006) from Instituto Tecnológico de Aeronáutica (ITA), Brazil. He is currently pursuing his Ph.D. from ITA. He is with Empresa Brasileira de Aeronáutica S.A (EMBRAER), São José dos Campos, SP, Brazil, since 2006. He works as a Development Engineer of a R&T group at EMBRAER focused on PHM technology applications in aeronautical systems
- Bruno P. Leão** holds a bachelor’s degree on Control and Automation Engineering (2004) from Universidade Federal de Minas Gerais (UFMG), Brazil, and a master’s degree on Aeronautical Engineering (2007) from Instituto Tecnológico de Aeronáutica (ITA), Brazil. He is currently pursuing his Ph.D. on the theme of failure prognosis from ITA. He is with EMBRAER S.A. in Brazil since 2005. He has worked as a Systems Engineer on the areas of Flight Controls and Automatic Flight Controls. Since 2007 he is with the PHM research group at EMBRAER developing diagnosis, prognosis and health monitoring solutions for aircraft systems. He has published over 10 PHM related papers in peer-reviewed conferences.
- Roberto Kawakami Harrop Galvão** holds a bachelor’s degree on Electronic Engineering (Summa cum Laude, 1995) from Instituto Tecnológico de Aeronáutica (ITA), Brazil. He also obtained the master’s (1997) and doctorate (1999) degrees in Systems and Control from the same institution. Since 1998 he has been with the Electronic Engineering Department of ITA as a full-time academic. Dr. Galvão is a Senior Member of the IEEE and an Associate

Member of the Brazilian Academy of Sciences. He has published more than 100 papers in peer-reviewed journals and conferences. His main areas of interest are fault diagnosis and prognosis, wavelet theory and applications, and model predictive control.

Takashi Yoneyama is a Professor of Control Theory with the Electronic Engineering Department of ITA. He received the bachelor's degree in electronic engineering from Instituto Tecnológico de Aeronáutica (ITA), Brazil, the M.D. degree in medicine from Universidade de Taubaté, Brazil, and the Ph.D. degree in electrical engineering from the University of London, U.K. (1983). He has more than 250 published papers, has written four books, and has supervised more than 50 theses. His research is concerned mainly with stochastic optimal control theory. Prof. Yoneyama served as the President of the Brazilian Automatics Society in the period 2004-2006.

Fleet-wide health management architecture

Maxime Monnin^{1,*,\dagger}, Alexandre Voisin^{2,*}, Jean-Baptiste Leger^{1,*,\dagger}
and Benoit Iung^{2,*,\dagger}

¹*PREDICT 19, Avenue de la Forêt de Haye, CS 10508, 54519 Vandoeuvre-Lès-Nancy, FRANCE*
maxime.monnin@predict.fr
jean-baptiste.leger@predict.fr

²*Centre de Recherche en Automatique de Nancy (CRAN), Nancy Université, UMR 7039 CNRS-UHP-INPL, Faculté des Sciences-1er Cycle - BP239, 54506 Vandoeuvre-Les-Nancy Cedex - France*
alexandre.voisin@cran.uhp-nancy.fr
benoit.iung@cran.uhp-nancy.fr

ABSTRACT

Large complex systems, such as power plants, ships and aircraft, are composed of multiple systems, subsystems and components. When they are considered as embedded in system operating as a fleet, it raises mission readiness and maintenance management issues. PHM (Prognostics and Health Management) plays a key role for controlling the performance level of such systems, at least on the basis of adapted PHM strategies and system developments. However, considering a fleet implies to provide managers and engineers with a relevant synthesis of information and keep it updated regarding both the global health of the fleet and the current status of their maintenance efforts. For achieving PHM at a fleet level, it is thus necessary to manage relevant corresponding knowledge arising both from modeling and monitoring of the fleet. In that way, this paper presents a knowledge structuring scheme for fleet PHM management applied to marine domain.

1. INTRODUCTION

1.1 Context

Large complex systems, such as power plants, ships and aircraft, are composed of multiple systems, subsystems and components built on different technologies (mechanical, electrical, electronic or software natures). These components follow different rates and modes of failures (Verma et al., 2010), for which behaviour can vary all along the different phases of their lifecycle (Bonissone and Varma, 2005), and maintenance actions strongly depends on this context (e.g. failure modes that occur, Cochetoux et al., 2009).

*The authors are all members of the French DIAG 21 Association (www.diag21.com). Dr Leger is a member of the management board and Pr Iung is co-chairing the prognostic working group

\daggerThe authors are member of the PHM Society

When they are considered as embedded in system operating as a fleet, it raises mission readiness and maintenance management issues.

In many cases, a fleet or plant operation is optimized (in terms of production or mission planning), making system availability a primary day to day concern. Thus, PHM plays a key role to ensure system performance and required, most of the time, to move from “fail and fix” maintenance practices to “predict and prevent” strategies (Iung et al., 2003), as promoted by Condition Based Maintenance (CBM)/PHM strategy mainly based on Condition-Monitoring capacities. Nevertheless, even if a condition monitoring program is in operation, failures still occur, defeating the objective for which the investment was made in condition monitoring (Campos, 2009). Moreover, the huge amount of condition monitoring activity, coupled with limitations in setting alarm levels (Emmannouilidis et al., 2010), has led to a problem for maintenance crew coping with the quantity of alarms on a daily basis (Moore and Starr, 2006).

From a practical point of view, predictive diagnosis aims at providing, to maintenance crew, key information about component current state and/or helping to decide the adapted maintenance action to be done, in order to anticipate/avoid failure. However, when considering a fleet of systems in the way to enhance maintenance efforts and facilitate the decision-making process, it is necessary, at the fleet level, to provide managers and engineers with a relevant synthesis of information and keep it updated regarding both the global health of the fleet and the current status of their maintenance efforts on components (Hwang et al., 2007).

Such an issue, at the fleet level, has to be tackled considering an information system enabling to gather/share information from individuals for synthesis,

case retrieval, engineering purposes. It enables to reuse particular data, such as maintenance history, reliability analysis, failure analysis, data analysis at a fleet level in order to provide knowledge. The reuse of such data requires turning them into information by adding semantic aspect while considered at the fleet level (Umiliacchi et al., 2011).

The semantic perspective at the fleet level allows:

- to unambiguously understand the data,
- to use them for reasoning as far as the reasoning knowledge has been modeled
- to put them in situation in order to enable comparison.

1.2 From collection of PHM systems to fleet integrated PHM system

PHM systems involve the use of multiple methods for acquiring and gathering data, monitoring and assessing the health, diagnosis and prognosis. Numerous approaches have been developed both for the diagnostic and prognostics purpose within system health monitoring. Such approaches are mainly data-driven methods, model-based and even hybrid. Moreover, dealing with systems requires, on the one hand, to consolidate data with for instance data fusion strategies (Roemer et al., 2010, Niu et al., 2010), and on the other hand, to take into account the system environment (Peysson et al., 2008), in order to provide relevant information for supporting diagnosis, prognostics, expertise or reporting processes.

However, most of these approaches cannot be applied in a straight-forward manner because they insufficiently support the multitude of different equipment, sub-system at system/plant-wide and provide only limited automation for failure prediction (Krause et al., 2010).

Hence, a main concern today in single and, even more, in multiple PHM systems design lies in the limitation due to the use of proprietary/closed information system leading to harden the integration of multiple applications. Hence, for instance, the Department of Defense policy community requires the use of open information systems to enable information sharing (Williams et al., 2008). Main standards used in the PHM systems are CBM+, Integrated Vehicle Health Management (IVHM) architecture (Williams et al., 2008), MIMOSA*... The two main parts of the later are dedicated to Open System Architecture for Enterprise Application Integration (OSA-EAI) and Open System Architecture for Condition Based Maintenance (OSA-CBM) (Thurston and Lebold, 2001). OSA-CBM improves CBM application by dividing a standard

CBM system into seven different layers, with technical modules solution as shown in figure 1. According to the OSA-CBM architecture, the health assessment is based on consumed data issued from different condition monitoring systems or from other health assessment modules. In that way, health assessment can be seen as the first step to manage global health state of complex systems (Gu et al., 2009). It allows to define if the health in the monitored component, sub-system or system has been degraded.

Although the use of standard brings syntaxes to warehouse data collection (Umiliacchi et al., 2011), it lacks semantics to benefit from information/event/decision made upon a component for its reuse on another component at the fleet level. Gebraeel (2010) proposes to consider a fleet of identical systems where each system consists of the same critical equipment. Such an approach is context dependent and provides a low level of reusability but allows, to some extent, comparison.

In a general case, where several different systems are considered as a fleet, several PHM systems and data warehouse coexist. Hence, a straightforward way to bring semantic at a fleet level is to develop and use ontology.

1.3 Fleet integrated PHM review

A fleet generally refers to a gathering of group of ships and by extension the term is also used for any kind of vehicle (e.g. trains, aircrafts, or cars). For industrial systems, the term fleet designs a set of assets or production lines. In general, a fleet refers to the whole of an owner's systems. In operational context, it refers to a subset of the owner fleet, e.g. a set of ships managed by a superintendant, or assets of a production site. Hence, the fleet here is only an abstraction point of view to consider a set of objects for a specific purpose (e.g. a unit maintenance planning), for a given time (e.g. before the end of the current mission). Indeed, the fleet can be viewed as a population consisting of a finite set of objects (individuals) on which a study is ongoing. In this context, a fleet is generally a subset of the real fleet under consideration, i.e. a sub fleet related to the aim of the study. Individuals making up the fleet/sub fleet may be, as needed, the systems themselves (Bonissone and Varma, 2005), (Patrick et al., 2010). When specific subsystems are under investigation, a fleet of all similar subsystems or installations is considered. Finally, a set of equipment may be also considered when a fleet is fitted (Umiliacchi et al., 2011). In the following, systems, sub-systems or equipments constituting the fleet, according to the study purpose, will be referred to as units.

* www.mimosa.org

In fact, fleet's units must share some characteristics that enable to group them together according to a specific purpose. These common characteristics may be of technical, operational or contextual nature. They allow to put data or information related to all the fleet units on the same benchmark in order to bring out pertinent results for monitoring, diagnostics or maintenance decision making.

Both fleet assignment and fleet maintenance scheduling problems have been studied mainly focusing on an optimization purpose (e.g. (Charles-Owaba et al., 2008), (Patrick et al., 2010)). Fleet management aims at maximizing adaptability, availability and mission success while minimizing costs and resources usage. When considering maintenance operator's point of view, fleet management aims at making decisions that affect asset life extension and performance, operational costs and future planning (Wheeler et al., 2009), (Bonissone and Varma, 2005), (Williams et al., 2008).

Nevertheless, fleet's predictive maintenance, i.e the fact of monitoring units' behaviors regarding the comparable behavior within the fleet, has rarely been addressed as a whole in the literature. (Umiliacchi et al., 2011) show the importance of having a standard format for the diagnostic data in order to facilitate their understanding across several subsystems and trains within a railway fleet. In (Patrick et al., 2010), the authors notice that thresholds indicative of condition indicators limits could be derived from statistical studies of fleet wide behaviors and known cases of faults. A more direct and less expensive maintenance technique is mentioned in (Reymonet et al., 2009). It consists in applying to the failed system the technical solution corresponding to a similar incident already solved with a comparable asset. Nevertheless, knowledge derived from the fleet in (Patrick et al., 2010) and (Reymonet et al., 2009) which arises from the same kind of units, in a domain where customized units are common, may give poor results.

1.4 Industrial Challenge

Behind the need of fleet PHM management stand an industrial demand. On one hand, the users of PHM system are fleet owners as well as fleet maintainers. Fleet owners aim at operating their fleet using indicators regarding not only single system but (sub) sets of systems as well. It requires being able to handle several indicators coming from several PHM systems in a common way in order to make easier data fusion/aggregation/synthesis, Human-Machine Interface (HMI) and their interpretation. Fleet maintainers would like to take benefit from event/decision already made in order to facilitate, enhance and/or confirm them. On the other hand, PHM system developers would like to decrease their

development time and cost. All the previous requirements could be done through the reuse of parts of PHM system already existing on similar systems.

From the operational point of view, efficient maintenance decision needs to analyze complex and numerous interrelated symptoms in order to identify the real (health) problem. The diagnostic process requires comparison between information coming from several subsystems. Moreover, diagnostics tasks are today still under the supervision of human experts, who can take advantage of their wide and long-term experience allowing appropriate actions to be taken (Umiliacchi et al., 2011). Such practical consideration raises limitations due to time consuming, repeatability of results, storage and transfer of knowledge.

For achieving PHM at a fleet level, it is necessary to manage relevant corresponding knowledge arising both from modeling and monitoring of the fleet. That leads to increasingly consider environment and condition of usage within the PHM main processes (Patrick et al., 2010) in order to allow monitored data and corresponding health to be analyzed by means of comparison from different points of view (for instance regarding the level considered or the operating condition). Indeed, monitored data and elaborated Health indicators strongly depends on the usage of the component. For instance engine cylinder temperatures are related to both the required power output and the cooling system for which inlet air or water depends on the external temperature. It is thus necessary to manage these criteria in order to compare for instance cylinder temperature within similar condition in terms of both power and external temperature in the available fleet-wide data.

The paper focuses on a knowledge structuring scheme for fleet PHM management in the marine domain. The goal of the proposed approach is to allow fleet units to benefit from the predictive maintenance features within a fleet scale. This could be possible by looking at the fleet level for further and complementary knowledge to the unit level. Such knowledge may emerge from similar situations already encountered among fleet units historical data/information. Next section introduces Fleet-wide Knowledge-based model development starting with the issue raised, and then presenting the basis of knowledge domain modeling and finally the fleet-wide expertise retrieval. The last section is dedicated to an illustrative industrial example dealing with fleet of diesel engines.

2. Fleet-wide Knowledge-based model

2.1 Issues

PHM development is a knowledge-intensive process, requiring a processing of expert knowledge together

with heterogeneous sources of data (Emmannouilidis et al., 2010). Such issue is strengthened at the fleet level. To support the main PHM processes development and to achieve a better understanding of monitored data, especially for diagnostic and maintenance decision making purposes, the underlying domain knowledge needs to be structured. Such system should enable to:

- Manage condition monitoring activities
- Associate monitored data with component operating condition
- Support diagnostic process with fleet-wide comparison facilities (i.e. benefits in a repeatable way of the fleet-wide expertise)
- Pro-actively anticipate failure (i.e. provide targeted maintenance actions recommendation).

It will ensure consistent information to be used throughout, from raw data acquisition to fleet-wide comparison (Figure 1). The key factor to turn data into such information is to enhance data with semantic context by means of ontology.

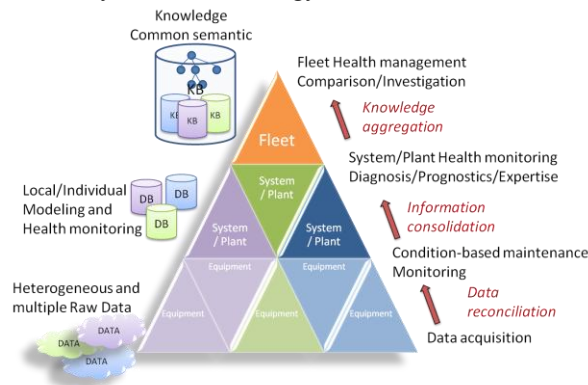


Figure 1: Proactive fleet management hierarchy, (Monnin et al., 2011a)

2.2 Basis of Knowledge modeling

Knowledge domain modeling relies on formal language that allows concepts to be described as well as the relationships that hold between these concepts. Starting from basic concepts, complex concepts can therefore be built up in definitions out of simpler concepts. Recent developments in the semantic modeling, based on information used and its context, have led to techniques using ontology to model complex systems. The ontology stores the relationships between physical components in a system, as well as more abstract concepts about the components and their usage (Figure 2). The key benefit over simple databases is that reasoning can take place to infer the consequences of actions or changes in the ontology instances (Umiliacchi et al. 2011).

Thus, information about the system can be inferred from the contextual information provided by the

ontology. For instance, consider a fleet of ships each of them having one or more diesel engines for propulsion and/or electric power generation. With an ontology-based system, both propulsion engine and generator engine can be considered as diesel engine. Thus, the system can handle a generic request for the state of the diesel engine and the corresponding data.

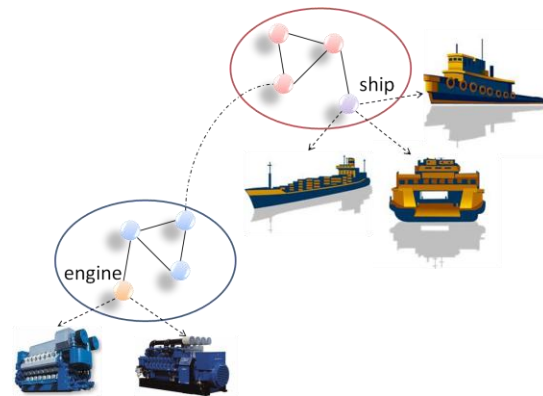


Figure 2: Scheme of concepts relationships

2.3 Fleet-wide expertise retrieval

For both diagnostic comparison and expertise sharing purposes, contextual information from the ontology enables to group component together given a particular context (e.g. component with the same usage). Four levels of context are defined in order to provide comparison facilities:

- Technical context
- Service context
- Operational context
- Performance context

These contexts defined within the ontology allow both to group instance sharing similar properties and to infer information about the system such as health indicators.

The technical context can be seen as the first and obvious level of comparison. It allows the technical features of the components to be described in the ontology. By means of taxonomy of components (Figure 3), it enables to conceptually describe components of a fleet. As a consequence, for instance, two different components (e.g. a propulsion engine and power generator engine) can be considered of the same type if a particular feature is considered (e.g. aspiration system).

However, from a practical point of view, the operating context influences the component behavior. The operating context can be split in service context and operational context.

The service context deals with sub-system for which component, even if similar, undergoes different solicitations. For instance, diesel engines can be both used for propulsion and electric power generation. Both

engines are diesel engines and can be compared from technical points of view. However, even if the components belong to the same type, their functioning (i.e. service context) is quite different (e.g. load changes, redundancy). On the other hand, components that belong to different types can be compared in a way, since they operate in the same service context

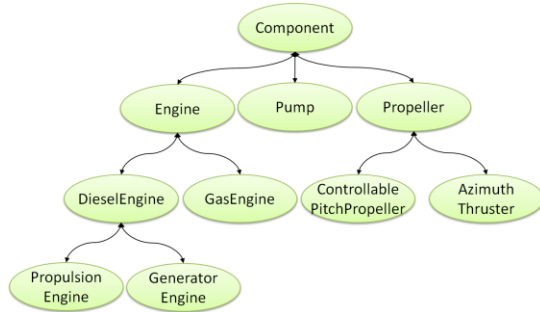


Figure 3: Part of the component ontology

The operational context defines the operating condition of a system (e.g. environment, threats). It provides contextual information according to the system operation. The definition of system taxonomy within the ontology enables to distinguish the operational contexts (e.g. Figure 4). This level describes higher operational requirements that can help the diagnostic process. For instance, abnormal behavior can be caused by the system environment. In that case the contextual information do not only concern technical or service context level.

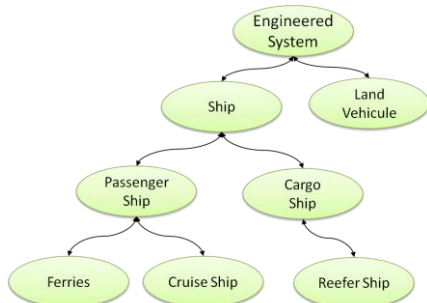


Figure 4: Part of system taxonomy

Finally, the performance context is linked to the key purpose of the fleet and defines, to some extent, the needs of optimization. For instance, a commercial fleet will focus on costs whereas a military application will be focused on availability. From a fleet-wide comparison point of view, the performance context enables large and global consideration to comparatively assess the global health of the fleet.

By means of taxonomies, each context can be described and both similarities and heterogeneities can be considered within the diagnostic process.

Therefore, the contextual information provided by the ontology allows better identification of component

operating condition - i.e. component health. It enables to provide the data of the monitored component with the corresponding context defined in the ontology. The significant health indicator can be defined according to the corresponding component and context.

In that way, health condition situation of component can be gathered according to different criteria (i.e. context description). From the diagnosis point view, abnormal behaviors, which are depicted through the health condition, can be defined by symptom indicators. The relationship between symptoms and faults is also considered in order to make available a certain understanding (i.e. diagnosis) of the corresponding health condition (Figure 5).

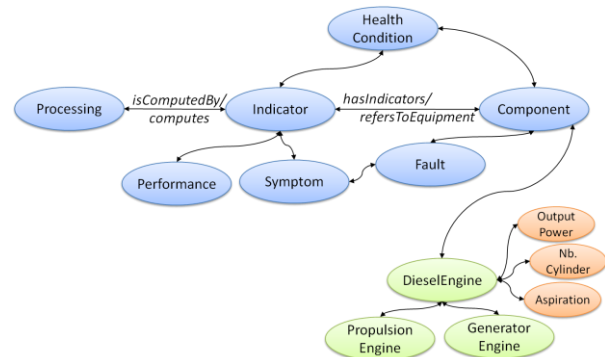


Figure 5: Part of the PHM ontology

Coupling with the data of monitored component, the abnormal behavior can be early detected. The corresponding indicators (performance, symptom...) allow early diagnostic and enable failure anticipation leading to plan adapted maintenance actions. The fleet-wide knowledge-based model, supported by means of ontology enables efficient predictive diagnosis and failure anticipation. The contextual information structured and stored within the ontology makes fleet-wide comparison easier. The fleet-wide expertise can be gathered, analyzed and reused, in a repeatable way.

The next section provides a case study of the fleet-wide knowledge-based model within an industrial PHM platform.

3. Industrial application

The industrial application demonstrates how the preceding concepts are embedded in a commercial application (Leger, 2004, Monnin, 2011b) developed by PREDICT. The example presents abnormal situation analysis helping using similar case retrieval within the fleet. The aim of the analysis is to anticipate failure, i.e. to perform predictive diagnosis. First we present the case under consideration, second the fleet wide knowledge platform, and finally situation monitoring and analysis.

3.1 Case Description

Diesel engines are critical onboard component of ship. In many cases they provide both propulsion of the ship and electrical power within many possible configurations. Avoiding blackout is of primary concerns and marine diesel engine monitoring and maintenance tend to benefit from advanced technology. Indeed, because embedded maintenance facilities are limited, a better knowledge of the engine health condition will allow to better drive maintenance actions needed when ships are in port.

For the purpose of this example, the fleet is limited to diesel engines. Seven engines are considered and briefly presented in Table 1. In this table an extract of the technical features of the engines are given as well as their use (i.e. propulsion, electric power generation and auxiliary).

Engine Ref	Output power (kW)	Nb. of Cylinder	...	Use
Wärtsilä 12V38	8 700	12V		ElectricPower
Wärtsilä 12V38	8 700	12V		ElectricPower
Baudouin6M26SRP1	331	6L		Auxiliary
Man V8-1200	883	8V		ElectricPower
Man V8-1200	883	8V		Propulsion
Wärtsilä 16V38	11600	16V		ElectricPower
Wärtsilä 12V38	8 700	12V		Propulsion

Table 1: Extract of engine fleet technical features

3.2 Fleet-wide knowledge-based platform

The ontology model is coded in OWL (Ontology Web Language) which is a formal ontology language, using the [†]Protégé ontology editor. The Protégé platform supports the modeling of ontologies. The ontologies can be exported into several formats including Resource Description Framework (RDF) and OWL.

For the purpose of the underlying software application, the ontology model is integrated by means of an SQL-backed storage and the java framework JENA[‡] is used for ontology exploitation through the KASEM platform. It provides the user with a web portal that allows benefiting of the fleet-wide expertise. The JENA inference engine allows semantic queries and inference rules to be solved within the platform. Relevant contextual information can be retrieved and gathered for the purpose of, for instance, failure anticipation, investigation or expertise sharing.

The underlying monitoring data are collected by means of a data warehouse (MIMOSA compliant). The

[†] <http://protege.stanford.edu/>

[‡] <http://jena.sourceforge.net/index.html>

platform integrates the ontology model on top of the warehouse data collection. Given an application, the data can be made available on-line, off-line or even on-demand. A typical architecture is given Figure 6.

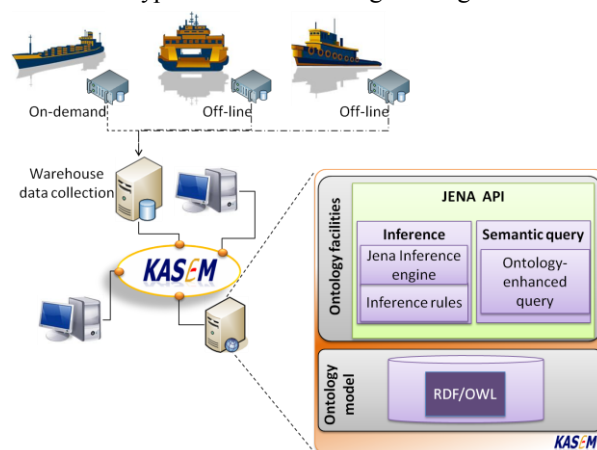


Figure 6: Typical architecture of Fleet-wide PHM system

3.3 Abnormal behavior Monitoring and Predictive Diagnosis

The diesel engine under consideration within the fleet includes regulatory sensor measurement as well as alarm monitoring system for the purpose of certification. Moreover further sensor measurements are also available for the engine operation. Some of commonly used sensor measurement are Cylinder temperature, Oil temperature, Oil pressure, SeaWater Temperature, SeaWater Pressure, FreshWater Temperature, FreshWater Pressure, Turbocharger temperature, Speed, Power output.

From a predictive diagnosis point of view existing alarm monitoring systems are not sufficient since they do not allow failure to be anticipated. Once the alarm occurs, the remaining time to failure is too short for preventing it. Moreover, the cause identification of such alarms must be analyzed subsequently.

Abnormal behavior can be monitored by means of specific indicators such as symptoms and analyzed within their contexts (i.e. technical, service, operational and performance). For the sake of illustration, we consider cylinder temperatures for diesel engines. In normal conditions the cylinders temperatures are changing in a similar way. Thus, a health indicator of abnormal behavior shall be built by detecting any evolution of one of the temperatures disconnected from the rest of the set of temperatures. Figure 7 illustrates temperatures measurement evolution of a diesel engine. Two behaviors are highlighted on the graph. The first behavior, labeled A, shows a normal situation where the temperatures are correlated despite one of them is a

couple of degrees below. The second behavior, labeled B, shows a decorrelation of the lowest signal.

Such data trend analysis, even if coupled with a detection process, will not allow to anticipate failure. Whereas the abnormal behavior is highlighted, contextual information that enable the understanding (i.e. diagnostic) of the behavior are missing. Retrieving similar situation and comparing it is almost not possible.

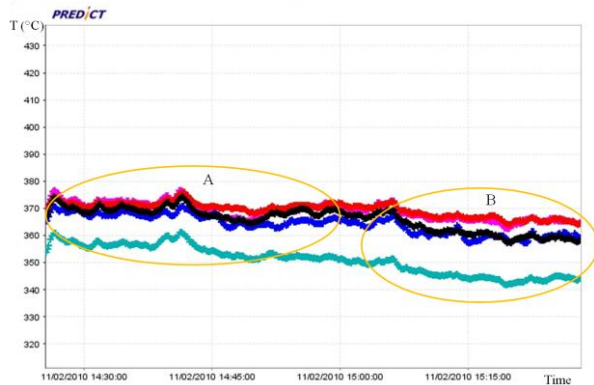


Figure 7: Zoom over a one-hour period of cylinder temperature measurement, zone A shows a normal behavior, while zone B an abnormal situation.

The knowledge-based model proposed allows providing such monitoring data with the corresponding context at different levels. Thus, fleet-wide comparison of the cylinder temperature evolution is enabled according to criteria such as technical context (e.g. same number of cylinders), service context (e.g. propulsion vs. electric power generation). If the corresponding fault has been identified and linked to the health condition situation (Figure 5), the underlying expertise can be retrieved.

Figure 8 presents an example of fleet-wide expertise retrieval results. For the given engines of the fleet (Table 1), some diagnostic results are proposed and summarized. With such a system, the expert, in face with a particular situation, can make any association to find out the closest cases with the case to solve and shall concentrate on the most frequent degradation modes already observed. From the different contextual information available, the system helps understanding the behavior without hiding its complexity with too simplistic rules.

4. CONCLUSION

Fleet-wide PHM requires knowledge-based system that is able to handle contextual information. Diagnosis and maintenance decision making processes are improved by means of semantic modeling that deals with concepts definition and description. In this paper, a knowledge model is proposed. Contextual information is structured by means of specific contexts. These

contexts allow considering fleet component similarities and heterogeneities. Data of the monitored component are considered within their context and enhance the identification of the corresponding health condition.

From a diagnosis point of view, the analysis of abnormal health condition leads to link the description of such behavior with the corresponding diagnosis and maintenance decision. Thus, the expertise becomes available within the fleet.

The fleet knowledge model has been done according to a marine application. The resulting ontology has been integrated in the KASEM industrial PHM platform and an example of use and results have been shown.

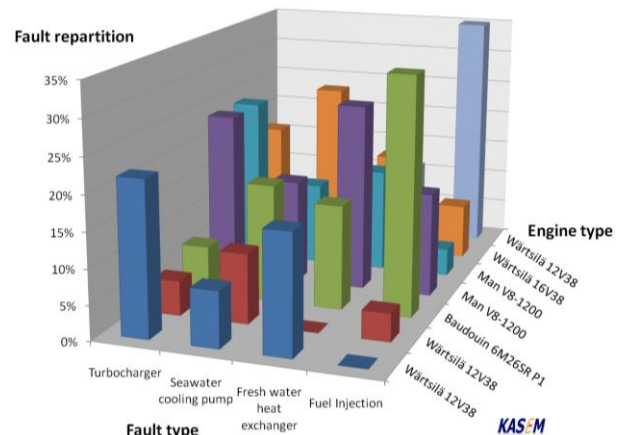


Figure 8: Sample of results for a fleet-wide cases retrieval visualization

ACKNOWLEDGEMENT

This work is based upon work supported by the BMCI project funded by the DGCIS and territorial collectivity.

REFERENCES

- Bonissone, P.P., Varma, A. (2005). Predicting the Best Unit within a Fleet: Prognostic Capabilities Enabled by Peer Learning, Fuzzy Similarity, and Evolutionary Design Process. *In Proceedings of the 14th IEEE International Conference on Fuzzy Systems*, IEEE, pp. 312-318.
- Campos J. (2009). Development in the application of ICT in condition monitoring and maintenance, *Computers in Industry*, vol. 60, pp. 1–20.
- Charles-Owaba O.E., Oluleye A.E., Oyawale F.A., Oke S.A. (2008). An opportunity cost maintenance scheduling framework for a fleet of ships: a case study, *Journal of Industrial Engineering International*, vol. 4, pp. 64-77.
- Emmannouilidis, C., Fumagalli, F., Jantunen, E., Pistofidis, P., Macchi, M., Garetti, M. (2010). Condition monitoring based on incremental learning and domain ontology for condition-based

- Maintenance, in *Proceedings of 11th international Conference on advances in Production Management Systems*, October 11-13, Cernobbio, Como, Italy.
- Cocheteux, P., Voisin, A., Levrat, E., Iung, B. (2009). Prognostic Design: Requirements and Tools, in *Proceedings of 11th International Conference on The Modern Information, Technology in the Innovation Processes of the Industrial Enterprises*, Bergamo, Italy.
- Gebraeel, N. (2010). Prognostics-Based Identification of the Top-k Units in a Fleet, *IEEE transactions on automation science and engineering*, vol. 7, pp. 37-48.
- Gu J., Lau D., Pecht M. (2009). Health assessment and prognostics of electronic products, in *Proceedings of 8th International Conference on Reliability, Maintainability and Safety*, July 21-25, Chengdu, China, pp. 912-919.
- Hwang, W.T., Tien S.W. and Shu, C.M. (2007). Building an Executive Information System for Maintenance Efficiency in Petrochemical Plants—An Evaluation, *Process Safety and Environmental Protection*, vol. 85, pp 139-146.
- Iung, B., Morel, G. and Léger, J.B. (2003). Proactive maintenance strategy for harbour crane operation improvement. n: *H. Erbe, Editor, Robotica. Special Issue on Cost Effective Automation*, vol. 21.
- Krause, J., Cech, S., Rosenthal, F., Gössling, A., Groba, C. and Vasyutynskyy, V. (2010). Factory-wide predictive maintenance in heterogeneous environments, in *Proceedings of 8th IEEE International Workshop on Factory Communication Systems*, May 18-21, Nancy, France, pp. 153-156.
- Léger J-B. (2004). A case study of remote diagnosis and e-maintenance information system, *Keynote speech of IMS'2004, International Conference on Intelligent Maintenance Systems*, Arles, France.
- Monnin, M, Leger, J-B., Morel, D. (2011a). Proactive facility fleet/plant monitoring and management, in *Proceedings of 24th International Congress on Condition Monitoring and Diagnostics Engineering Management*, 29th May – 1st June, Stavanger, Norway.
- Monnin, M, Leger, J-B., Morel, D. (2011b). KASEM®: e-Maintenance SOA Platform, in *Proceedings of 24th International Congress on Condition Monitoring and Diagnostics Engineering Management*, 29th May – 1st June, Stavanger, Norway.
- Moore, W.J., Starr, A.G. (2006). An intelligent maintenance system for continuous cost-based prioritisation of maintenance activities, *Computers in Industry*, vol. 57, pp. 595–606.
- Niu G, Yang B, Pecht M. (2010). Development of an optimized condition-based maintenance system by data fusion and reliability-centered maintenance, *Reliability Engineering and System Safety*, vol. 95, pp. 786–796.
- Patrick, R., Smith, M J., Byington, C S., Vachtsevanos, G J., Tom, K., Ly, C. (2010). Integrated Software Platform for Fleet Data Analysis, Enhanced Diagnostics, and Safe Transition to Prognostics for Helicopter Component CBM, in *Proceedings of Annual Conference of the Prognostics and Health Management Society*, October 10-16, Portland, Oregon.
- Peysson, F., Ouladsine, M., Outbib, R., Leger, J-B., Myx, O., Allemand, C. (2008). Damage Trajectory Analysis based Prognostic, in *Proceedings of IEEE International Conference on Prognostics and Health Management*, October 6-9, Denver, CO, pp. 1-8.
- Reymonet, A., Thomas, J., Aussenac-Gilles, N. (2009). Ontology Based Information Retrieval: an application to automotive diagnosis, in *Proceedings of International Workshop on Principles of Diagnosis*, June 14-17, Stockholm, Sweden, pp. 9-14.
- Roemer, M.J., Kacprzyński, G.J. and Orsagh, R.F. (2001). Assessment of data and knowledge fusion strategies for prognostics and health management, in *Proceedings of IEEE Aerospace Conference Proceedings*, Big Sky, MT, USA, pp. 2979–2988.
- Thurston, M., Lebold, M. (2001). Open standards for condition-based maintenance and prognostic systems, in *Proceedings of MARCON2001*, <http://www.osacbm.org>.
- Umiliacchi, P., Lane, D., Romano, F. (2011). Predictive Maintenance of railway subsystems using an Ontology based modelling approach, in *Proceedings of 9th world Conference on Railway Research*, May 22-26, Lille, France.
- Verma, A. K. and Srividya, A. and Ramesh, P. (2010). A systemic approach to integrated E-maintenance of large engineering plants, *International Journal of Automation and Computing*, vol. 7, pp. 173-179.
- Wheeler, K., Kurtoglu, T., Poll, S.D. (2009). A survey of health management user objectives related to diagnostic and prognostic metrics, in *Proceedings of International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, August 30–September 2, San Diego, California, USA.
- Williams, Z., Gilbertson, D. & Sheffield, G., (2008). Fleet analysis and planning using CBM+ open architecture, in *Proceedings of IEEE International Conference on Prognostics and Health Management*, Denver, CO

Improving data-driven prognostics by assessing predictability of features

Kamran Javed¹, Rafael Gouriveau¹, Ryad Zemouri², Noureddine Zerhouni¹

¹ *Femto-st Institute, AS2M Department, 25000 Besançon, France*

kamran.javed@femto-st.fr

rgourive@ens2m.fr

noureddine.zerhouni@ens2m.fr

² *Laboratoire d'Automatique du CNAM, 75003 Paris, France*

ryad.zemouri@cnam.fr

ABSTRACT

Within condition based maintenance (CBM), the whole aspect of prognostics is composed of various tasks from multi-dimensional data to remaining useful life (RUL) of the equipment. Apart from data acquisition phase, data-driven prognostics is achieved in three main steps: features extraction and selection, features prediction, and health-state classification. The main aim of this paper is to propose a way of improving existing data-driven procedure by assessing the predictability of features when selecting them. The underlying idea is that prognostics should take into account the ability of a practitioner (or its models) to perform long term predictions. A predictability measure is thereby defined and applied to temporal predictions during the learning phase, in order to reduce the set of selected features. The proposed methodology is tested on a real data set of bearings to analyze the effectiveness of the scheme. For illustration purpose, an adaptive neuro-fuzzy inference system is used as a prediction model, and classification aspect is met by the well known Fuzzy C-means algorithm. Both enable to perform RUL estimation and results appear to be improved by applying the proposed strategy.

1. INTRODUCTION

Due to rapid growth in industrial standards, effective maintenance support systems are main area of focus nowadays. Different strategies have been adapted to assess machinery condition in real time and to avoid costly maintenance procedures. In this context, Condition Based Maintenance (CBM) strategy facilitates the competitive needs of industry by preventing costly maintenance activities, and thus, improving availability, reliability and security of machinery (Tobon-Mejia et al.,

2011). In CBM, researchers show keen interest in less developed phase of prognostics that determines or predicts the remaining useful life (RUL) of a system (machinery) under certain operational conditions (Jardine et al., 2006). However, accurate prognostic systems are still scarce in the industry and need for an improvement is inevitable.

Prognostics can be categorized mainly into three approaches: experience based, model based and data driven methods (Heng & Zhang, 2009; Lebold & Thurston, 2001b; Ramasso & Gouriveau, 2010). Among these approaches data driven methods are considered to be a trade-off between experience based and model based approaches. They are increasingly applied to machine prognostics due to their effectiveness and ability to overcome limitations of latter categories (El-Koujok et al., 2008).

Mainly, the degradation process of a system (component) is reflected by features that are extracted from a sensor signal. These features are main source of information for prognostics model to estimate RUL. So, most importantly, in existing data-driven procedure of prognostics, critical phase of prediction should be met in appropriate manner for further classification and RUL estimation. However, from afore said procedure two issues can be pointed out. Firstly, there is no unique way to select most relevant features that are predictable and contribute for better RUL estimation. Secondly, the predictability should be assessed according to prediction model as well as horizon of prediction. This paper contributes to extend the existing approach by proposing a slight modification of features selection phase on the basis of predictability.

This paper is organized as follows. Section 2. discusses data-driven prognostics approach and points out the importance of the prediction accuracy. Following that, section 3. presents an improved framework for feature selection, based on the predictability assessment of features. Section 4. aims at defining the whole prognostics model that is employed in this paper. Both aspects of multi-steps ahead prediction and of health

Javed et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

state classification are considered. Section 5. deals with simulation and results discussion. Finally, section 6. concludes this research work.

2. DATA-DRIVEN PROGNOSTICS

2.1 Prognostics process flow

In maintenance field, prognostics considered as a key task within CBM that predicts RUL of machinery under certain operational modes and facilitates decision making. Thereby, the main objective of prognostics is to estimate RUL of system (component) before occurrence of failure state. Therefore, within CBM concept, the whole aspect prediction and failure can be viewed as set of certain activities that must be performed in order to accomplish predictive maintenance procedures (Lebold & Thurston, 2001a).

Mainly, data-driven methods alter raw (unprocessed) data into useful information and forecast global performance of the system. In order to deduce RUL, prognostic task is applied by performing forecasts in time and further analyzing them by classification module to approximate most probable states of the system (Fig. 1 and 2). More precisely, in a first stage, data acquisition from sensor sources is performed, and further pre-processed before feeding prediction model. The second stage of data-preprocessing is composed of two distinct phases i.e., feature extraction module, that is accomplished by signal processing techniques and feature selection module that depends on data mining approaches. Finally, in third stage of prognostics, prediction module forecasts observations in time, that are further analyzed by the classifier module to determine most probable states of the system. Lastly, RUL is derived by the estimated time to attain the failure state.

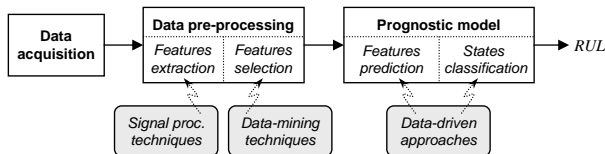


Figure 1. Prognostics process flow

2.2 Underlying predictability problem

From data-driven approaches, artificial intelligence (AI) based tools like artificial neural networks and neuro-fuzzy (NFs) have successfully been employed to perform non-linear modeling of prognostics (W.Q. Wang et al., 2004; Lebold & Thurston, 2001a). The standard of AI approaches is divided into two phases, i.e., learning phase and testing phase. As, monitored input/output data is the main source of information for prediction model, therefore, firstly the behavior is learned by monitored data and secondly, the test phase uses learned model to predict current and future states of degrading equipment.

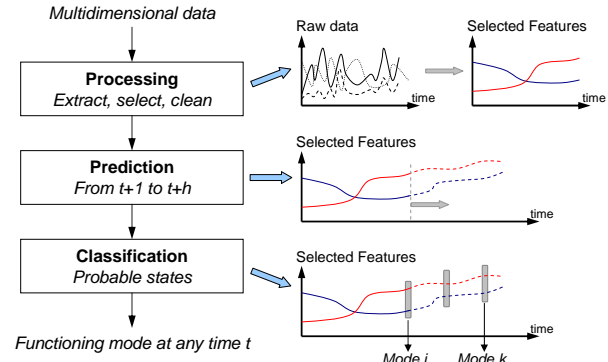


Figure 2. From data to RUL

In classical way prognostics model is learned by set of features that are acquired by sensor signal. Thereby, the model must be retrained upon these features, until significant performance level is achieved. This approach can be time expensive because some of the features can be very hard to be predicted. In other words, there is no use in retaining such features that are not predictable. So, the learning phase of prognostics modeling should consider the important steps of “feature selection” and “prediction modeling” in a simultaneous manner in order to retain or reject features on the basis of predictability. Thereby, this implies predictability to be defined (next section).

3. SELECTION OF PREDICTABLE FEATURES

3.1 Accuracy vs predictability

Predictability attributes to the significance in making predictions of future occurrence on the basis of past information. It is important to understand the prediction quality in a framework that is dependent on the considered time series predictability. As, predictability in terms of given time series is not a well defined terminology for real-world processes, few works focus on the predictability aspect (Kaboudan, 1999; W. Wang et al., 2008; Diebold & Kilian, 2001). Assuming that, in order to determine prediction quality, predictability can be measured on the basis of forecast error based approach. Various measures have been reported in literature to judge the quality of prediction or selecting a prediction model (Saxena et al., 2008, 2009, 2010; Monnet & Berger, 2010). See Eq. (1) for a set of potential metrics that can be used to assess predictability:

$$\begin{aligned}
 \text{MSE} &= \frac{1}{N} \times \sum_{i=1}^N \left(y_{pred}^i - y_{act}^i \right)^2 \\
 \text{MAPE} &= \frac{100}{N} \times \sum_{i=1}^N \left| \left(y_{pred}^i - y_{act}^i \right) / y_{act}^i \right| \\
 \text{RMSE} &= \sqrt{\text{MSE}} \\
 \text{CVRMSE} &= \text{RMSE} / \mu_y \\
 \text{MFE} &= \frac{1}{N} \times \sum_{i=1}^N \left(y_{pred}^i - y_{act}^i \right)
 \end{aligned} \tag{1}$$

From these measures MSE, MAPE and RMSE are most common accuracy measures for prediction, whereas CVRMSE and MFE can be employed to model selection. However, there is no general measure that can be explicitly employed to predictability factor of prognostics.

Generally, any type of signal will not be predicted with the same accuracy at different horizons of prediction. So, assuming that, the critical prediction phase in prognostics must be met accurately in order to provide efficient information. Therefore, predictability in prognostics not only is closely related to prediction model but also to the horizon of prediction that is judged as useful. On this basis, a new measure is proposed in this paper to assess predictability in prognostics.

3.2 Defining the predictability concept

Assessing the prognostics model requires the user to be able to define a suitable limit to prediction for the desired performance. According to author’s knowledge, the predictability concept is not well described. So, it can be defined as: “The ability of a given time series TS to be predicted with an appropriate modeling tool M , that facilitates future outcomes over a specific horizon H , and with desired performance limit L ”. Formally we propose it as:

$$Pred(TS/M, H, L) = exp^{-\left| \ln\left(\frac{1}{2}\right) \cdot \frac{MFE_{TS/M, H}}{L} \right|} \quad (2)$$

where, Eq. (2) shows the empirical formulation in which $MFE_{TS/M, H}$ represents the mean forecast error Eq. (1), that measures average deviation of predicted values from actuals. The ideal value for this criteria is 0, if the value of $MFE > 0$ then prediction model tends to underforecast, else if the value of $MFE < 0$ then prediction model tends to overforecast. Moreover, the fixed limit of accuracy is denoted by L (chosen by the user). The exponential form of predictability can attain maximum value “1” as MFE is minimizes, and a given TS is considered predictable, if the coefficient of predictability ranges between $[0.5, 1]$ (Fig. 3).

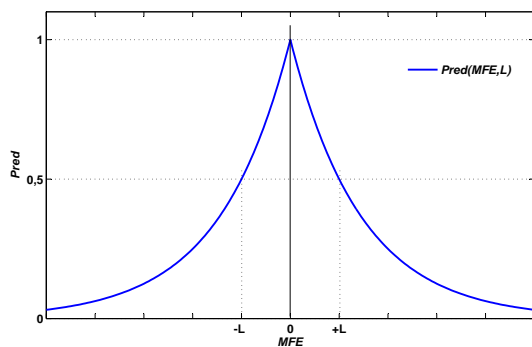


Figure 3. Illustration of predictability measure

4. PROGNOSTICS MODELING

4.1 Multi-steps ahead prediction

In prognostics, forecasting the global health state of a system is difficult task to achieve due to inherent uncertainty. However, from the category of data driven prognostics, AI based approaches like ANN and NFs can be quiet easily applied to such complex and non-linear environment.

Such connexionist systems have good capability to learn and adapt from environment and capture complex relationship among data. They are increasingly applied to prediction problems in maintenance field (Yam et al., 2001; Chinnam & Baruah, 2004; El-Koujok et al., 2011). They appear to be potential tools, in order to predict degrading behavior, and thus forecast the global state of the system.

Multi-step ahead (MSP) modeling can be achieved different ways by using connexionist tools. However, in this case, the most common MSP model can be achieved via iterative approach. MSPs are obtained using a single connexionist tool that is tuned for single-step ahead prediction \hat{x}_{t+1} . The predicted value is further utilized as one of the regressors of prediction model, and this process is followed in an iterative way until estimation \hat{x}_{t+H} , as shown in Fig. 4. Formally:

$$\hat{x}_{t+h} = \begin{cases} * \text{ if } h = 1, \\ f^1(x_t, \dots, x_{t+1-p}, [\theta^1]) \\ * \text{ elseif } h \in \{2, \dots, p\}, \\ f^1(\hat{x}_{t+h-1}, \dots, \hat{x}_{t+1}, x_t, \dots, x_{t+h-p}, [\theta^1]) \\ * \text{ elseif } h \in \{p+1, \dots, H\}, \\ f^1(\hat{x}_{t+h-1}, \dots, \hat{x}_{t+h-p}, [\theta^1]) \end{cases} \quad (3)$$

where, t denotes temporal index variable, p is for number of regressors used and H states the horizon of prediction. Whereas, $\{f^1, [\theta^1]\}$ states for single-step ahead prediction model, with its parameter calculation performed during learning phase.

In this paper the Adaptive Neuro-Fuzzy Inference System (ANFIS) is used as a the one step-ahead prediction model. A detailed description of this tool can not be given in the paper. One can cite to (Jang, 1993; Li & Cheng, 2007) for theoretical background.

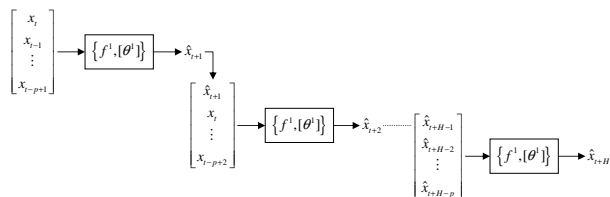


Figure 4. Multi-steps ahead predictions with iterative model

4.2 Classification step

The main aim of the classification phase is to determine most probable states of the degrading system, and thus providing a snapshot of time from projected degradations. In this phase, the temporal predictions made by the prediction module are analyzed by classifier module to determine most probable functioning modes of system (component). Most importantly, reliable and effective classification results better RUL estimation (Fig. 5). However in this case, due to the absence of ground-truth information the classification phase is met by well known Fuzzy C-Means (FCM) approach to illustrate our concept.

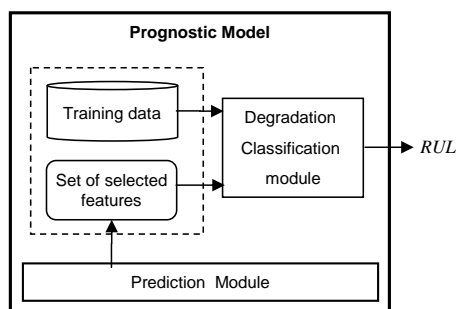


Figure 5. Classification Module

FCM is used as an unsupervised clustering approach that assigns temporal predictions to different classes based on fuzzy partitioning. In other words, a data point with a membership grade between $[0, 1]$, can belong to various groups (Bezdek, 1981). Formally, the FCM clustering is attained by assigning membership to every data point that corresponds to each cluster center that is based on the measured distance between a data point and center of the cluster. Mainly, if a data point is closer to particular cluster center, therefore, a greater membership value is assigned. Moreover, the summation of membership grades from all data points correspond to a membership equal to '1'. Mainly, FCM aims to operate in an iterative manner to determine cluster centers that reduces following objective function:

$$J = \sum_{i=1}^n \sum_{j=1}^c (u_{ij})^m \cdot \|x_i - v_j\|^2 \quad (4)$$

where, $\|x_i - v_j\|^2$ represents the euclidean distance between the i^{th} data point and the j^{th} cluster center, u_{ij} describes the membership of the i^{th} data point to the j^{th} centroid, and $m > 1$ is a weighting exponent.

5. EXPERIMENTS AND DISCUSSION

5.1 Experimental setup

The proposed methodology for feature selection is illustrated by real data set of bearings form NASA data Repository. The

data set consisted of multiple time series (variables) from different instances and contaminated with measurement noise (Fig. 6) i.e., representing history of fault degradation process. Moreover there is no information about the bearing condition and manufacturing variations. The simulation process is composed of three stages i.e., data-preprocessing, feature prediction and selection and health state classification to estimate RUL.

For experimental purpose, in the first stage only 8 variables (features F1-F8) are utilized from bearing data set, and filtered for noise removal.

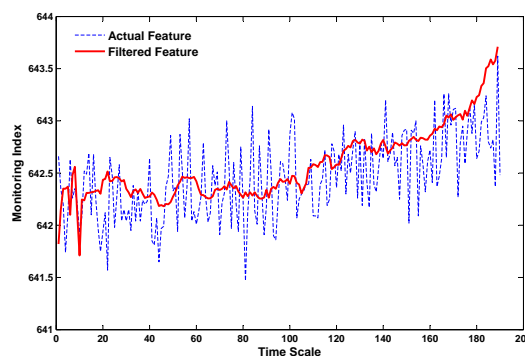


Figure 6. Filtered feature from bearing data set

The second phase corresponds to proposed feature selection methodology based on predictability. So, for illustration purpose ANFIS is used as potential connexionist tools to perform MSP. Each prediction model is tuned according to settings shown in Table 1. The training and testing data sets were composed of 40 bearings data each. However, to achieve MSP over different horizons, model training is met by a data set of 40 bearings, whereas, 5 test cases are employed for analysis purpose. All the predictions are analyzed by potential measures of accuracy (Eq. 1). In order to perform feature selection, proposed predictability measure is employed to validate our concept (Eq. 2).

ANFIS-Parameters	Settings
Input / Output layer neurons	3 / 1
Number of input membership functions	3
Type of input membership functions	Pi-shaped
Number of rules	27
Fuzzy Inference System	First order Sugeno
Defuzzification method	Weighted Average
Output Membership function	Linear
Learning Algorithm	Hybrid Method
Number of epochs	100
Training performance	MSE

Table 1. ANFIS model settings

Finally, classification phase partitions the temporal predictions into four modes of degradation, i.e., each mode repre-

sents fault progression toward end of life.

To show the concept of predictability for better classification and RUL estimation, simulations are performed on all features (F1-F8) and also with selected features that are predictable (excluding F2 and F3). Therefore, the obtained results from both cases give better perception of estimated RUL.

5.2 Prediction results

In the test phase, predictions are performed over different horizons (Fig 7). The horizon length for short term, mid-term and long term based on 35/80/140 steps ahead. The obtained outputs from each prediction tool are analyzed in a comprehensive manner using different performance metrics. To exemplify this scenario, a test case of bearing is presented in Table 2.

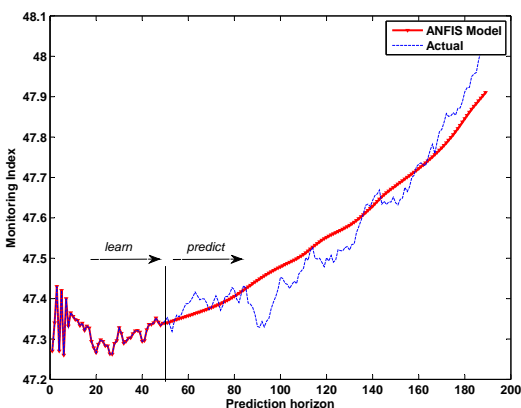


Figure 7. Example of predicted feature

In the selection phase, the outputs from selected models are assessed by MFE criteria and also with the proposed measure of predictability.

Among all features the MFE values for features F2 and F3 were not within bounds of desired performance criteria. Similar findings were achieved with the proposed measure of predictability Eq. (2). The validity of proposed measure can be clearly demonstrated by results from bearing test 1, as shown in Fig. 8. By these results it is well understood that F2 and F3 are not predictable according to defined predictability criteria. Therefore, better predictable features are F1, F4, F5, F6, F7 and F8, which can be selected for further classification to determine probable functioning modes of degrading asset.

5.3 Classification results

For illustration the temporal predictions from bearing test 1 are used for classification and RUL estimation. Therefore, the results are organized in two different cases for in an explicit manner for better perception and understanding (Fig. 9 and 10). In the first case classification is achieved with all features (F1- F8), whereas in the second case the classification is performed on predictable features only i.e.,excluding F2 and F3.

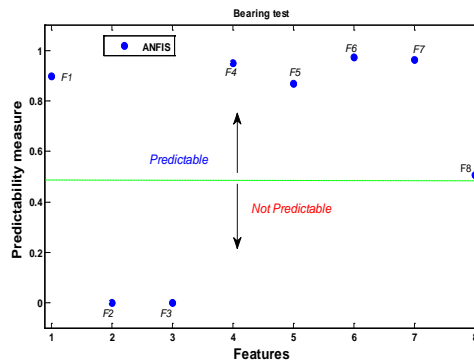


Figure 8. Predictable and not predictable feature set

It can be clearly judged from the results below that the first case shows inferior classification as compared to the classification performed by features that are selected on the basis of predictability. Moreover, the RUL deduced from second case of classifications is closer to the actual RUL, thus, validating better prognostics accuracy and improvements achieved from proposed methodology.

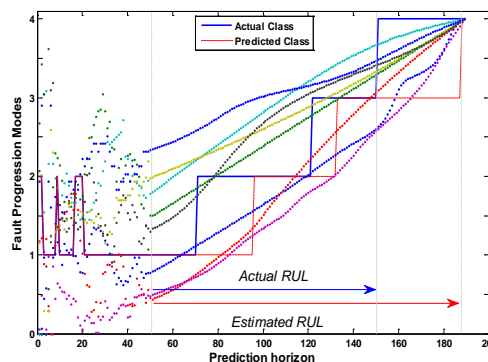


Figure 9. Classification with all features

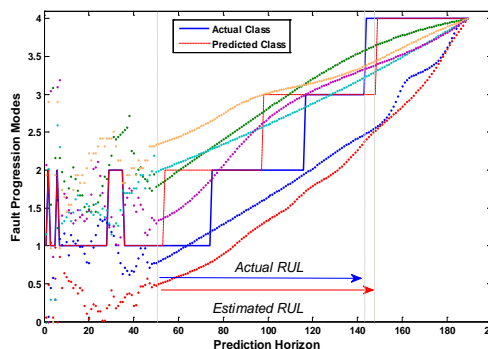


Figure 10. Classification with predictable features

	F1	F2	F3	F4	F5	F6	F7	F8
RMSE	0,111	2,3911	4,431681	0,0742	0,0495	0,0382	0,0334	0,507
MAPE	0,015	0,1166	0,219175	0,0024	0,0829	0,001	0,2909	0,103
CVRMSE	0,017	0,1503	0,314495	0,0031	0,1041	0,0016	0,3959	0,129
MFE	-0,083	1,5625	3,041406	0,0576	-0,008	0,0192	0,0237	0,013
<i>Pred</i>	0,682	0,0007	7,88E-07	0,7662	0,9648	0,9149	0,8961	0,944

Table 2. Predictability of bearing test(1) over long-term horizon

6. CONCLUSION

In this paper an improvement to existing data-driven prognostics approach has been presented. The proposition is based on the assessment of the predictability of features that impacts the accuracy of prognostics. The proposed methodology was met in three phases: 1) learning the prognostics model, 2) assessing temporal predictions on the basis of predictability, and 3) selecting those features that are better to be predictable. Mainly, multi-step ahead predictions were performed by ANFIS predictor. Lastly, set of predictable features were classified to determine possible fault modes, thanks to Fuzzy C-means clustering approach. The comparative analysis of classifications of test cases, show the efficiency of proposed methodology of “predictability based feature selection”.

REFERENCES

- Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithm*. Plenum New York.
- Chinnam R.B., & Baruah, P. (2004). A neuro-fuzzy approach for estimating mean residual life in condition-based maintenance systems. *Int. Jour. of Materials & Product Technology*, 20, 166-179.
- Diebold F.X., & Kilian. (2001). Measuring Predictability: Theory and Macroeconomic Applications. *Jour. of Applied Econometrics*, 16, 657-669.
- El-Koujok, M., Gouriveau, R., & Zerhouni, N. (2011). Reducing arbitrary choices in model building for prognostics: An approach by applying parsimony principle on an evolving neuro-fuzzy system. *Microelectronics Reliability*, 51, 310-320.
- El-Koujok, M., Gouriveau, R., & Zerhouni, N. (2008). Towards a neuro-fuzzy system for time series forecasting in maintenance applications. In *IFAC World Congress, Korea*.
- Heng, A., & Zhang, S. (2009). Rotating machinery prognostic: State of the art, challenges and opportunities. *Mech. systems & signal processing*, 23, 724-739.
- Jang J.S.R. (1993). ANFIS: Adaptive-Network-Based Fuzzy Inference System. *IEEE Trans. Systems, Man, Cybernetics*, 23, 665-685.
- Jardine, A., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mech. systems & signal processing*, 20, 1483-1510.
- Kaboudan, M. (1999). A Measure of Time-Series Predictability Using Genetic Programming Applied to Stock Returns. *Jour. of Forecasting*, 18, 345-357.
- Lebold, M., & Thurston, M. (2001a). Open standards for condition-based maintenance and prognostics systems. In *5th Annual Maint. and Reliability Conf.*
- Lebold, M., & Thurston, M. (2001b). Prognostic Enhancements to diagnostic Systems for Improved Condition-based maintenance. In *Maint. and Reliability Conf. MARCON*.
- Li, C., & Cheng K.H. (2007). Recurrent neuron-fuzzy hybrid-learning approach to accurate system modeling. *Fuzzy Sets and Systems*, 158, 194-212.
- Monnet J.-M., & Berger, F. (2010). Support vector machines regression for estimation of forest parameters from airborne laser scanning data. In *IEEE IGARSS USA*.
- Ramasso, E., & Gouriveau, R. (2010). Prognostics in Switching Systems: Evidential Markovian Classification of Real-Time Neuro-Fuzzy Predictions. In *IEEE Int. Conf. PHM, Hong-Kong*.
- Saxena, A., Celaya, J., Balaban, E., & Saha, B. (2008). Metrics for Evaluating Performance of Prognostic Techniques. In *Int. Conf. PHM*.
- Saxena, A., Celaya, J., & Saha, B. (2009). On Applying the Prognostics Performance Metrics. In *Annual Conf. of the PHM*.
- Saxena, A., Celaya, J., & Saha, B. (2010). Metrics for Offline Evaluation of Prognostic Performance. *Int. Jour. of PHM*, 001, 2153-2648.
- Tobon-Mejia, D., Medjaher, K., & Zerhouni, N. (2011). Estimation of the Remaining Useful Life by using Wavelet Packet Decomposition and HMMs. In *IEEE Int. Aerospace Conf., USA* (Vol. 6, p. 163-171).
- Wang W.Q., Goldnaraghi M.F., & Ismail, F. (2004). Prognosis of machine health condition using neuro-fuzzy systems. *Mech. systems & signal processing*, 18, 813-831.
- Wang, W., Van Gelder, P. H., & Vrijling J. K. (2008). Measuring predictability of Daily Streamflow Processes Based on Univariate Time Series Model. In *iEMSS* (Vol. 16, p. 474-3478).
- Yam R.C.M., Tse P.W., Li, L., & Tu, P. (2001). Intelligent predictive decision support system for condition-based maintenance. *Int. Jour. of Adv. Manufacturing Tech.*, 17, 383-391.

Integrated Robust Fault Detection, Diagnosis and Reconfigurable Control System with Actuator Saturation

Jinhua Fan¹, Youmin Zhang^{2*}, and Zhiqiang Zheng³

^{1,3}*National University of Defense Technology, Changsha, Hunan, 410073, China*

fjhcom@gmail.com
zqzheng@nudt.edu.cn

^{1,2}*Concordia University, Montreal, Quebec, H3G 1M8, Canada*

** ymzhang@encs.concordia.ca*

ABSTRACT

An integrated fault detection, diagnosis and reconfigurable control design method is studied in this paper with explicit consideration of control input constraints. The actuator fault to be treated is modeled as a control effectiveness loss, which is diagnosed by an adaptive algorithm. For fault detection, an observer is designed to generate the output residual and a minimum threshold is set by an H_∞ index. To design the reconfigurable controller, an auxiliary matrix is introduced and a linear parameter varying (LPV) system is constructed by convex combination. Linear matrix inequality (LMI) conditions are presented to compute the design parameters of controllers and related performance index. The system performances are measured by the ellipsoidal sets regarding the domain of attraction and disturbance rejection respectively. For illustration, the proposed design techniques are applied to the flight control of a flying wing aircraft under large effectiveness loss of actuators.

1. INTRODUCTION

The reconfigurable fault-tolerant control design methods have been studied widely in the literature to meet increased requirements for reliability and safety in modern control systems (Zhang & Jiang, 2008). One key component in fault-tolerant control systems is the fault detection and diagnosis (FDD) module, which has been studied extensively in the past decades (Isermann, 2006). With information provided by FDD, the controller is adjusted according to some reconfiguration mechanism to maintain desirable performances. One challenging problem in designing reconfigurable fault-tolerant control system is how to integrate the FDD with the controller effectively to

guarantee the system performance, such as stability, etc. Another practical consideration is to take the control input constraints into control system design procedure, since almost all practical applications involve actuators constrained by limited power, for example, the deflection of control surfaces in aircraft is constrained by amplitude and rate limitation. Hence, it is very significant to provide some design methods for the reconfigurable control problem with explicit consideration of control input constraints.

Currently, the constrained control systems are widely studied in the literature (Tarbouriech & Turner, 2009). Although there are still many open problems remained to be investigated, many useful results have been obtained due to efforts of past decades. Based on the fact that system performance can be improved if the controller can be designed to allow actuator saturation compared with that obtained within control limits. Along with this idea, many researchers have made their efforts in this direction of research. For example, a saturated system is represented by a polytopic model to solve the output tracking problem (Tarbouriech, Pittet & Burgat, 2000). An improved set invariance condition is given in (Hu, Lin & Chen, 2002) to obtain a less conservative estimation of domain of attraction. As will be shown in this paper, these results provide a tool to solve the reconfigurable control problem.

The reconfigurable control problem with actuator saturation is still not well addressed in the literature, and only a few results available in recent years. Generally speaking, there are two types of approaches to deal with such issues: one using the command management techniques (Bodson & Pohlchuck, 1998; Zhang & Jiang, 2003; Zhang, Jiang & Theilliol, 2008), and the other relating to controller design (Pachter, Chandler & Mears, 1995; Guan & Yang,

¹Visiting Ph.D. Student/Visiting Researcher, Department of Mechanical and Industrial Engineering

²Engineering, Corresponding Author

³Professor, College of Mechatronic Engineering and Automation

2009). For example, the actuator rate saturation problem is solved by a linear programming algorithm in (Pachter et al., 1995). An adaptive output-feedback controller is designed with online estimation of actuator faults by Guan et al. (2009). However, only the stability problem is studied in that paper.

In this paper, we aim to solve the reconfigurable output tracking control problem of linear systems subject to both actuator saturation and disturbances. The actuator saturation is tackled with by using the set invariance condition given by Hu et al. (2002). The controller can be adjusted automatically with estimated fault amplitude provided by an adaptive diagnostic algorithm, after a fault occurs and is detected by an observer-based detector.

The paper is organized as follows: The problem to be treated is stated in Section 2. An integrated design of the reconfigurable controller with fault diagnosis is presented in Section 3. To detect a fault, an observer is designed in Section 4. Then, a nonlinear model of an aircraft is used to test the proposed design techniques in Section 5. Finally, some concluding remarks are given in Section 6.

2. PRELIMINARIES AND PROBLEM STATEMENT

To illustrate the basic ideas in this paper, a scalar control system with state-feedback controller is taken as an example:

$$\begin{aligned} \dot{x}(t) &= ax(t) + bu(t) \\ u(t) &= f(x) \end{aligned} \quad (1)$$

The fault under consideration is the loss of control effectiveness such that

$$u_f(t) = \lambda(t)u(t) \quad (2)$$

where $u_f(t)$ represents the output of the impaired actuator, and $\lambda(t) \in [0,1]$ is the control effectiveness factor. $\lambda(t) = 0$ means the total outage of the actuator, while 1 denotes a healthy actuator. Partial loss of control effectiveness is given by a value between 0 and 1. It is assumed that $\lambda(t) \neq 0$ in this paper.

To compensate the control effectiveness loss, the following control law can be adopted:

$$u(t) = \tilde{\lambda}^{-1}(t)f(x) \quad (3)$$

From Eqs. (2) and (3), it follows that

$$u_f(t) = \lambda(t)\tilde{\lambda}^{-1}(t)f(x) = f(x) \quad (4)$$

Obviously, the system performance is not impaired in the presence of actuator fault while the control law shown in Eq. (3) is in action. However, the fault cannot be known a priori, and only its estimation is available. In this case, Eq. (3) should be replaced by

$$u(t) = \tilde{\lambda}^{-1}(t)f(x) \quad (5)$$

where $\tilde{\lambda}(t)$ is an estimation of $\lambda(t)$.

If the estimation process can be carried out accurately and quickly enough, then the performance loss can be reduced to its minimum. For a constant fault $\lambda(t \geq t_f) = \lambda_0$ occurring at t_f , the performance can be recovered completely when $\tilde{\lambda}(t)$ converges to λ_0 . The controller structure for compensation of effectiveness loss is shown in Fig. 1.

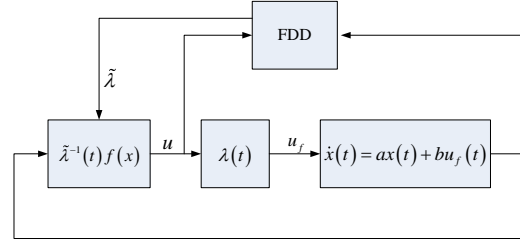


Fig. 1 Compensation principle for effectiveness loss

Above discussions can be extended readily to the multivariable systems. From practical point of view, since the control power is limited and the disturbance exists, then the plant to be controlled in this paper is given by:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + BM(t)\text{sat}[u(t)] + E\omega(t) \\ y(t) &= Cx(t) \\ e(t) &= r(t) - y(t) \end{aligned} \quad (6)$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $y(t) \in \mathbb{R}^p$ are the state, input, and output vectors respectively. $\omega(t) \in \mathbb{R}^q$ is an immeasurable disturbance vector bounded by $\|\omega(t)\| \leq \omega_0$. $r(t) \in \mathbb{R}^p$ is the reference signal vector bounded by $\|r(t)\| \leq r_0$. $e(t)$ is the tracking error vector. A , B , C and E are known parameter matrices of appropriate dimensions. It is assumed that (A, C) is detectable. $\text{sat}(\cdot)$ is a standard vector-valued saturation function with its elements given by:

$$\text{sat}(u_i) = \text{sign}(u_i) \cdot \min\{1, |u_i|\}, \quad i = 1, 2, \dots, m \quad (7)$$

where $\text{sign}(\cdot)$ represents the signum function.

$M(t) \in \mathbb{R}^{m \times m}$ is a diagonal matrix representing the effectiveness factors of actuators, and denoted by:

$$\begin{aligned} M(t) &= \text{diag}\{\lambda_1(t), \lambda_2(t), \dots, \lambda_m(t)\} \\ \lambda_i(t) &\in [\underline{\lambda}_i, \bar{\lambda}_i], \quad 0 < \underline{\lambda}_i \leq 1, \quad \bar{\lambda}_i \geq 1, \quad i = 1, 2, \dots, m \end{aligned} \quad (8)$$

where $\text{diag}\{\cdot\}$ represents a diagonal matrix. $\lambda_i(t)$, $i=1,2,\dots,m$ are unknown stepwise fault signals. $\underline{\lambda}_i$ and $\bar{\lambda}_i$ represent the known lower and upper bound of $\lambda_i(t)$ respectively.

The control objective in this paper is to realize stable tracking of a reference signal in the presence of faults and amplitude constraints of actuators. The overall control system configuration is shown in Fig. 2. The fault detection and diagnosis (FDD) module is used to detect a fault and provide an estimation of fault amplitude denoted by $\tilde{M}(t) = \text{diag}\{\tilde{\lambda}_1(t), \tilde{\lambda}_2(t), \dots, \tilde{\lambda}_m(t)\}$.

With estimated control effectiveness factors from FDD, the reconfigurable controller adjusts automatically its parameter to recover the performance of the closed-loop system. In this paper, an observer is used to detect a fault, and an adaptive algorithm is designed to estimate the fault amplitude. After a fault is detected by the observer, the adaptive diagnostic algorithm is activated automatically. Otherwise, a unitary matrix is passed to $\tilde{M}(t)$. In summary, the observer is used to determine when a fault occurs, and the adaptive diagnostic algorithm is used to estimate its amplitude.

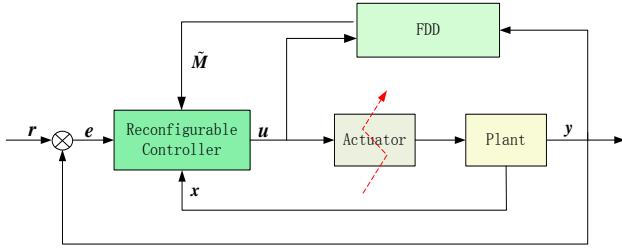


Fig. 2 System configuration

It is well known that the tracking error integral action of a controller can effectively eliminate the steady-state tracking error (Zhang & Jiang, 2001). Denote $\eta(t) = \int_0^t e(\tau) d\tau$, $\zeta(t) = \begin{bmatrix} x(t)^T & \eta(t)^T \end{bmatrix}^T$, then the following augmented system can be obtained from Eq. (6) such that

$$\dot{\zeta}(t) = \bar{A}\zeta(t) + \bar{B}M(t)\text{sat}[u(t)] + \bar{E}d(t) \quad (9)$$

where $\bar{A} = \begin{bmatrix} A & \mathbf{0}_{n \times p} \\ -C & \mathbf{0}_{p \times p} \end{bmatrix}$, $\bar{B} = \begin{bmatrix} B \\ \mathbf{0}_{p \times m} \end{bmatrix}$, $\bar{E} = \begin{bmatrix} E & \mathbf{0}_{n \times p} \\ \mathbf{0}_{p \times q} & I_p \end{bmatrix}$, $d(t) = \begin{bmatrix} \omega(t) \\ r(t) \end{bmatrix}$.

To eliminate the fault effect, the reconfigurable controller is realized by

$$u(t) = \tilde{M}^{-1}(t)K\zeta(t) \quad (10)$$

Substituting Eq. (10) into Eq. (9), it is obtained that

$$\dot{\zeta}(t) = \bar{A}\zeta(t) + \bar{B}M(t)\text{sat}[\tilde{M}^{-1}(t)K\zeta(t)] + \bar{E}d(t) \quad (11)$$

3. INTEGRATION OF RECONFIGURABLE CONTROLLER WITH FAULT DIAGNOSIS

LMI conditions will be presented in this section to design the controller gain K , while the estimated fault amplitude $\tilde{M}(t)$ is obtained by an adaptive algorithm.

Lemma 1 (Hu & Lin, 2001) Let $u, v \in \mathbb{R}^m$ and suppose that $|v_i| \leq 1$, $i=1,2,\dots,m$, then

$$\text{sat}(u) \in \text{Co}\{\Delta_j u + \Delta_j^- v, j=1,2,\dots,2^m\} \quad (12)$$

where $\text{Co}\{\cdot\}$ denotes the convex hull. $\Delta_j \in \mathbb{R}^{m \times m}$ is a diagonal matrix whose elements are either 0 or 1, and $\Delta_j^- = I_m - \Delta_j$. For brevity, $|v_i| \leq 1$, $i=1,2,\dots,m$ is written as $|v| \leq 1$ in the following.

From Lemma 1, if there exists an auxiliary matrix H satisfying

$$|\tilde{M}^{-1}(t)H\zeta(t)| \leq 1 \quad (13)$$

then there always exist $\alpha_j \geq 0$, $\sum_{j=1}^{2^m} \alpha_j = 1$ such that

$$\dot{\zeta}(t) = \sum_{j=1}^{2^m} \alpha_j (\bar{A} + \bar{B}M(t)\tilde{M}^{-1}(t)[\Delta_j K + \Delta_j^- H])\zeta(t) + \bar{E}d(t) \quad (14)$$

If α_j , $j=1,2,\dots,2^m$ are taken as the scheduled parameters and can be obtained online, then Eq. (14) is actually an LPV system. Define

$$F(t) = \tilde{M}^{-1}(t)H \quad (15)$$

then it is not difficult to find out that (13) imposes a polyhedral set constraint on system states of Eq. (14) as follows:

$$\mathcal{L}(F(t)) = \{\zeta(t) \mid |F_i(t)\zeta(t)| \leq 1, i=1,2,\dots,m\} \quad (16)$$

For estimation of domain of attraction, an ellipsoidal set is defined as follows:

$$\Omega(P) = \{\zeta(t) \mid \zeta^T(t)P\zeta(t) \leq 1, P > 0\} \quad (17)$$

Theorem 1 If there exist matrices $Y_k \in \mathbb{R}^{m \times (n+p)}$, $Y_h \in \mathbb{R}^{m \times (n+p)}$, a positive definite matrix $Q \in \mathbb{R}^{(n+p) \times (n+p)}$, and a positive scalar μ such that

$$\begin{bmatrix} 1 & \phi_i k_i^T Y_h \\ * & \mathbf{Q} \end{bmatrix} \geq 0, \quad i=1,2,\dots,m \quad (18)$$

$$\bar{\mathbf{A}}\mathbf{Q} + \mathbf{Q}\bar{\mathbf{A}}^T + \frac{1}{\mu}\bar{\mathbf{E}}\bar{\mathbf{E}}^T + \mu(r_0^2 + \omega_0^2)\mathbf{Q} + 2\bar{\mathbf{B}}(\Delta_j \mathbf{Y}_k + \Delta_j^- \mathbf{Y}_h) < 0 \quad (19)$$

$j=1,2,\dots,2^m$

then $\Omega(\mathbf{P})$ is an invariant set with $\mathbf{P}=\mathbf{Q}^{-1}$, $\mathbf{K}=\mathbf{Y}_k \mathbf{P}$, $\mathbf{H}=\mathbf{Y}_h \mathbf{P}$, and with the fault diagnostic algorithm being realized by:

$$\begin{aligned} \dot{\lambda}_i(0 \leq t < t_f) &= 0, \quad i=1,2,\dots,m \\ \dot{\lambda}_i(t \geq t_f) &= \text{Proj}_{[\underline{\lambda}_i, \bar{\lambda}_i]} \left\{ \gamma_i \tilde{\lambda}_i^{-1}(t) \zeta^T(t) \mathbf{P} \bar{\mathbf{b}}_i k_i^T \sum_{j=1}^{2^m} \alpha_j [\Delta_j \mathbf{K}_r + \Delta_j^- \mathbf{H}_r] \zeta(t) \right\} \end{aligned} \quad (20)$$

where $\phi_i \in \{\underline{\lambda}_i, \bar{\lambda}_i\}$. $\gamma_i > 0$ is pre-specified positive scalar. $\bar{\mathbf{b}}_i$ is the i -th column of $\bar{\mathbf{B}}$. k_i^T is a row vector with its i -th element being 1 and the other elements being 0. $\text{Proj}_{[\underline{\lambda}_i, \bar{\lambda}_i]} \{\cdot\}$ is a projection operator defined as follows:

$$\text{Proj}_{[\underline{\lambda}_i, \bar{\lambda}_i]} \{X\} = \begin{cases} 0, & \hat{\lambda}_i \geq \bar{\lambda}_i \text{ and } X > 0 \\ & \text{or} \\ & \hat{\lambda}_i \leq \underline{\lambda}_i \text{ and } X < 0 \\ X, & \text{else} \end{cases} \quad (21)$$

Proof: Denote

$$\mathbf{E}_\lambda(t) = \tilde{\mathbf{M}}(t) - \mathbf{M}(t) = \text{diag}\{e_{\lambda 1}(t), e_{\lambda 2}(t), \dots, e_{\lambda m}(t)\} \quad (22)$$

Define a Lyapunov function

$$V(t) = \zeta^T(t) \mathbf{P} \zeta(t) + \sum_{i=1}^m \gamma_i^{-1} e_{\lambda i}^2(t) \quad (23)$$

Its derivative with respect to time is given by:

$$\begin{aligned} \dot{V}(t) &= 2\zeta^T(t) \mathbf{P} \sum_{j=1}^{2^m} \alpha_j (\bar{\mathbf{A}} + \bar{\mathbf{B}} \mathbf{M} \tilde{\mathbf{M}}^{-1}(t) [\Delta_j \mathbf{K} + \Delta_j^- \mathbf{H}]) \zeta(t) \\ &\quad + 2\zeta^T(t) \mathbf{P} \bar{\mathbf{E}} \mathbf{d}(t) + 2 \sum_{i=1}^m \gamma_i^{-1} e_{\lambda i}(t) \dot{e}_{\lambda i}(t) \end{aligned} \quad (24)$$

Since

$$\begin{aligned} 2\zeta^T(t) \mathbf{P} \bar{\mathbf{E}} \mathbf{d}(t) &\leq \frac{1}{\mu} \zeta^T(t) \mathbf{P} \bar{\mathbf{E}} \bar{\mathbf{E}}^T \mathbf{P} \zeta(t) + \mu \mathbf{d}^T(t) \mathbf{d}(t) \\ &\leq \frac{1}{\mu} \zeta^T(t) \mathbf{P} \bar{\mathbf{E}} \bar{\mathbf{E}}^T \mathbf{P} \zeta(t) + \mu(r_0^2 + \omega_0^2) \end{aligned} \quad (25)$$

then it follows that:

$$\dot{V}(t) \leq \zeta^T(t) \mathbf{M} \zeta(t) + 2 \sum_{i=1}^m \gamma_i^{-1} e_{\lambda i}(t) \dot{e}_{\lambda i}(t) + \mu(r_0^2 + \omega_0^2) \quad (26)$$

where

$$\mathbf{M} \dagger 2\mathbf{P} \sum_{j=1}^{2^m} \alpha_j (\bar{\mathbf{A}} + \bar{\mathbf{B}} \mathbf{M} \tilde{\mathbf{M}}^{-1}(t) [\Delta_j \mathbf{K} + \Delta_j^- \mathbf{H}]) + \frac{1}{\mu} \mathbf{P} \bar{\mathbf{E}} \bar{\mathbf{E}}^T \mathbf{P} \quad (27)$$

Since

$$\mathbf{M} \tilde{\mathbf{M}}^{-1}(t) = \mathbf{I}_m - \mathbf{E}_\lambda(t) \tilde{\mathbf{M}}^{-1}(t) \quad (28)$$

it follows that:

$$\mathbf{M} = \bar{\mathbf{M}} - 2\mathbf{P} \bar{\mathbf{B}} \mathbf{E}_\lambda(t) \tilde{\mathbf{M}}^{-1}(t) \sum_{j=1}^{2^m} \alpha_j [\Delta_j \mathbf{K} + \Delta_j^- \mathbf{H}] \quad (29)$$

where

$$\bar{\mathbf{M}} = \mathbf{P} \bar{\mathbf{A}} + \bar{\mathbf{A}}^T \mathbf{P} + \frac{1}{\mu} \mathbf{P} \bar{\mathbf{E}} \bar{\mathbf{E}}^T \mathbf{P} + 2\mathbf{P} \bar{\mathbf{B}} \sum_{j=1}^{2^m} \alpha_j [\Delta_j \mathbf{K} + \Delta_j^- \mathbf{H}] \quad (30)$$

Since

$$\bar{\mathbf{B}} \mathbf{E}_\lambda(t) \tilde{\mathbf{M}}^{-1}(t) = \sum_{i=1}^m \bar{\mathbf{b}}_i e_{\lambda i}(t) k_i^T \tilde{\mathbf{M}}^{-1}(t) = \sum_{i=1}^m e_{\lambda i}(t) \tilde{\lambda}_i^{-1}(t) \bar{\mathbf{b}}_i k_i^T \quad (31)$$

then it can be obtained from Eqs. (20) and (26) that

$$\dot{V}(t) \leq \zeta^T(t) [\bar{\mathbf{M}} + \mu(r_0^2 + \omega_0^2) \mathbf{P}] \zeta(t) + \mu(r_0^2 + \omega_0^2) [1 - \zeta^T(t) \mathbf{P} \zeta(t)] \quad (32)$$

With Eqs. (23) and (32), it is not difficult to verify that $\Omega(\mathbf{P})$ is an invariant set by satisfying

$$\bar{\mathbf{M}} + \mu(r_0^2 + \omega_0^2) \mathbf{P} < 0 \quad (33)$$

which is equivalent to Eq. (19).

To complete the proof, it is still needed to guarantee that $\Omega(\mathbf{P}) \subset \mathcal{L}(\mathbf{F}(t))$, of which an equivalent condition can be stated as follows:

$$\Theta_i = \left\{ \begin{array}{l} \max_{\zeta(t)} |F_i(t) \zeta(t)| \\ \text{s. t. } \zeta^T(t) \mathbf{P} \zeta(t) = 1 \end{array} \right\} \leq 1 \quad (34)$$

By using the method of Lagrange multipliers, it is not difficult to obtain that

$$\Theta_i = \sqrt{F_i(t) \mathbf{P}^{-1} F_i^T(t)} \quad (35)$$

Since

$$F_i(t) \mathbf{P}^{-1} F_i^T(t) = [\tilde{\mathbf{M}}^{-1}(t)]_i Y_h \mathbf{Q}^{-1} Y_h^T [\tilde{\mathbf{M}}^{-1}(t)]_i^T \quad (36)$$

then by Schur complement, an equivalent condition for $\Omega(\mathbf{P}) \subset \mathcal{L}(\mathbf{F})$ is given by:

$$\begin{bmatrix} 1 & [\tilde{\mathbf{M}}^{-1}(t)]_i Y_h \\ * & \mathbf{Q} \end{bmatrix} \geq 0, \quad i=1,2,\dots,m \quad (37)$$

The extreme point set of $\tilde{\mathbf{M}}^{-1}(t)$ can be defined as follows:

$$\Sigma \triangleq \left\{ \Psi^j \mid \Psi^j = \text{diag}\{\phi_1, \phi_2, \dots, \phi_m\}, \phi_i = \lambda_i^{-1} \text{ or } \bar{\lambda}_i^{-1}, \right. \\ \left. i = 1, 2, \dots, m; j = 1, 2, \dots, 2^m \right\} \quad (38)$$

then from the convexity of Σ , there always exist $\beta_j \geq 0$,

$$\sum_{j=1}^{2^m} \beta_j = 1 \quad \text{such that}$$

$$\tilde{\mathbf{M}}^{-1}(t) = \sum_{j=1}^{2^m} \beta_j \Psi^j \quad (39)$$

From Eq. (39), it gives

$$\left[\tilde{\mathbf{M}}^{-1}(t) \right]_i = \sum_{j=1}^{2^m} \beta_j (\Psi^j)_i = \sum_{j=1}^{2^m} \beta_j \phi_j \mathbf{k}_i^1 \quad (40)$$

then (37) can be written as:

$$0 \leq \begin{bmatrix} 1 & \sum_{j=1}^{2^m} \beta_j \phi_j \mathbf{k}_i^1 \cdot \mathbf{Y} \\ * & \mathbf{Q} \end{bmatrix} = \sum_{j=1}^{2^m} \beta_j \begin{bmatrix} 1 & \phi_j \mathbf{k}_i^1 \mathbf{Y}_h \\ * & \mathbf{Q} \end{bmatrix} \quad (41) \\ i = 1, 2, \dots, m; j = 1, 2, \dots, 2^m$$

It is sufficient for (18) to guarantee that (41) holds true. This ends the proof. \square

Remark 1: The values of $\alpha_j, j = 1, 2, \dots, 2^m$ are needed online in the adaptive diagnostic algorithm as shown in Eq. (20). One way to obtain them (Wu, Lin & Zheng, 2007) is shown as follows:

$$\alpha_j = \prod_{i=1}^m [z_i (1 - \lambda_i) + (1 - z_i) \lambda_i] \quad (42)$$

where $z_1 2^{m-1} + z_2 2^{m-2} + \dots + z_m = j - 1$, and

$$\lambda_i = \begin{cases} 1, & \mathbf{H}_r = \mathbf{K}_r \\ \text{sat}(\tilde{m}_i^{-1}(t) \mathbf{k}_i^1 \mathbf{K}_r \zeta(t)) - \tilde{m}_i^{-1}(t) \mathbf{k}_i^1 \mathbf{H}_r \zeta(t) \\ \tilde{m}_i^{-1}(t) \mathbf{k}_i^1 \mathbf{K}_r \zeta(t) - \tilde{m}_i^{-1}(t) \mathbf{k}_i^1 \mathbf{H}_r \zeta(t), & \text{else} \end{cases} \quad (43)$$

Since $\Omega(\mathbf{P})$ is an estimation of the domain of attraction, it is desirable to obtain the largest one. This is a volume maximization problem. In general, there are two ways to maximize $\Omega(\mathbf{P})$. Since the volume of $\Omega(\mathbf{P})$ is proportional to $\det(\mathbf{Q})$, one direct way is to construct an determinant maximization problem (Vandenberghe, Boyd & Wu, 1998) as follows:

$$\sup_{\mathbf{Q} > 0, \mathbf{Y}_k, \mathbf{Y}_h, \mu > 0} \log \det \mathbf{Q} \\ \text{s.t. (18) and (19)} \quad (44)$$

The other way is to use a prescribed bounded convex reference set X_R to maximize $\Omega(\mathbf{P})$, which can take its

shape into consideration. Two typical sets of X_R are the ellipsoids and polyhedrons. By taking an ellipsoid $X_0 = \left\{ \zeta(t) \in \mathbb{R}^{n+p} \mid \zeta(t)^T \mathbf{R} \zeta(t) \leq 1, \mathbf{R} > 0 \right\}$ as the reference set, the following optimization problem can be formulated:

$$\sup_{\mathbf{Q} > 0, \mathbf{Y}_k, \mathbf{Y}_h, \mu > 0} \alpha \\ \text{s.t. (a) } \alpha X_0 \subset \Omega(\mathbf{P}) \quad (45) \\ \text{(b) (18) and (19)}$$

Let $\gamma = 1/\alpha^2$, since $\alpha X_0 = \Omega(\gamma \mathbf{R})$, then $\alpha X_0 \subset \Omega(\mathbf{P})$ is equivalent to $\gamma \mathbf{R} \geq \mathbf{P}$. By Schur complement, (45) can be written as:

$$\inf_{\mathbf{Q} > 0, \mathbf{Y}_k, \mathbf{Y}_h, \mu > 0} \gamma \\ \text{s.t. (a) } \begin{bmatrix} \gamma \mathbf{R} & \mathbf{I} \\ \mathbf{I} & \mathbf{Q} \end{bmatrix} \geq 0 \quad (46) \\ \text{(b) (18) and (19)}$$

For a reference set described by a polyhedron $X_0 = \text{conv}\{x_1^0, x_2^0, \dots, x_N^0\}$, $\alpha X_0 \subset \Omega(\mathbf{P})$ is equivalent to $(x_i^0)^T \mathbf{P} x_i^0 \leq \gamma$. Then by Schur complement, the first LMI in Eq. (46) should be replaced by:

$$\begin{bmatrix} \gamma & (x_i^0)^T \\ x_i^0 & \mathbf{Q} \end{bmatrix} \geq 0 \quad (47)$$

In another aspect, the system states cannot be guaranteed to converge to the origin due to the disturbances and actuator faults. Hence, a performance index is needed for the disturbance rejection problem, which can also be described by a prescribed bounded convex reference set. Assumed that this set is denoted by X_∞ , then an optimization problem can be formulated as follows:

$$\inf_{\mathbf{Q} > 0, \mathbf{Y}_k, \mathbf{Y}_h, \mu > 0} \beta \\ \text{s.t. (a) } \Omega(\mathbf{P}) \subset \beta X_\infty \quad (48) \\ \text{(b) (18) and (19)}$$

To address the disturbance rejection and domain of attraction simultaneously, a scaled version of $\Omega(\mathbf{P})$ is defined as follows:

$$\Omega(\mathbf{S}) = \left\{ \zeta(t) \in \mathbb{R}^{n+p} \mid \zeta(t)^T \mathbf{S} \zeta(t) \leq 1, \mathbf{S} = \rho^{-1} \mathbf{P}, \mathbf{S} > 0, 0 < \rho \leq 1 \right\} \quad (49)$$

From the convexity (Hu *et al.*, 2002) of both (18) and (19), it is not difficult to verify that all the trajectories starting from within $\Omega(\mathbf{P})$ will enter $\Omega(\mathbf{S})$ and remain inside it if there exist $\mathbf{Q} > 0, \mathbf{Y}_k, \mathbf{Y}_h, \mu > 0$ satisfying (18), (19) and

$$\rho \bar{A}Q + \rho Q \bar{A}^T + \frac{1}{\mu} \bar{E} \bar{E}^T + \mu \omega_0^2 \rho Q + 2\bar{B}(\Delta_j Y_k + \Delta_j Y_s) < 0, j=1,2,\dots,2^m$$

$$\begin{bmatrix} 1 & \phi_i k_i Y_s \\ * & \rho Q \end{bmatrix} \geq 0, i=1,2,\dots,m$$
(50)

Therefore, to solve the disturbance rejection problem with guaranteed domain of attraction, the following optimization problem can be formulated:

$$\inf_{Q>0, Y_k, Y_s, \mu>0} \beta$$

s.t. (a) $\Omega(S) \subset \beta X_\infty$

(b) $X_0 \subset \Omega(P)$

(c) (18),(19) and (50)

(51)

Remark 2: The controller gain K computed from (51) may be too high to be used in practice. To adjust the controller gain K , since $K = Y_k P$, then the following inequality can be added into the optimization problems:

$$Y_k Y_k^T \leq \sigma I_m, \sigma > 0$$
(52)

By Schur complement, (52) is equivalent to

$$\begin{bmatrix} \sigma I_m & Y_k \\ * & I_n \end{bmatrix} \geq 0$$
(53)

4. OBSERVER-BASED FAULT DETECTION

To activate the adaptive diagnostic algorithm as shown in Eq. (20), the time t_f when a fault occurs is needed to be known. It is the responsibility of fault detection. In this paper, the fault detection is carried out by comparing the output residual with the threshold to be set.

To detect the fault, an observer is defined as follows:

$$\begin{aligned} \dot{\tilde{x}}(t) &= A\tilde{x}(t) + B \text{sat}[u(t)] + L(y(t) - \tilde{y}(t)) \\ \tilde{y}(t) &= C\tilde{x}(t) \end{aligned}$$
(54)

where $\tilde{x}(t)$ and $\tilde{y}(t)$ are estimation of $x(t)$ and $y(t)$ respectively.

Denote $e_x(t) = \tilde{x}(t) - x(t)$, $e_y(t) = \tilde{y}(t) - y(t)$, then an observer error equation can be obtained without incorporating $\omega(t)$

$$\begin{aligned} \dot{e}_x(t) &= (A - LC)e_x(t) + B(I_m - M(t))\text{sat}[u(t)] \\ e_y(t) &= Ce_x(t) \end{aligned}$$
(55)

With (A, C) being detectable, it is not difficult to obtain the observer gain L such that $A - LC$ is stable. Then a fault is detected if $\|e_y(t)\| > \lambda_f$, where λ_f is a pre-specified

threshold. If $\tilde{x}(0) = x(0)$, then $\lambda_f = 0$ is sufficient to detect a fault.

However, when $\omega(t)$ is presented, false alarm may be generated with above detector, even if $\tilde{x}(0) = x(0)$ is satisfied. Increasing λ_f may prevent a false alarm, but it may lead to a detector which is insensitive to a fault of small amplitude. Hence, it is desirable to determine a minimum threshold.

In the presence of the disturbance $\omega(t)$, the observer error equation becomes:

$$\begin{aligned} \dot{e}_x(t) &= (A - LC)e_x(t) + B(I_m - M(t))\text{sat}[u(t)] - E\omega(t) \\ e_y(t) &= Ce_x(t) \end{aligned}$$
(56)

Assumed that $e_x(0) = 0$, then by Laplace transformation, it is obtained that

$$e_y(s) = C(sI_n - A + LC)^{-1} B(I_m - M(t))\text{sat}[u(s)] + G(s)\omega(s)$$
(57)

where $G(s) = -C(sI_n - A + LC)^{-1} E$.

Since no fault occurs when $t < t_f$, that is $M(t < t_f) = I_m$, then Eq. (57) can be written as:

$$e_y(s) = G(s)\omega(s)$$
(58)

Since the disturbance $\omega(t)$ is unknown, then the H_∞ norm of $G(s)$ can be used, which is denoted by:

$$\|G(s)\|_\infty = \sup_{\omega \in \mathbb{R}} \sigma_{\max}[G(j\omega)]$$
(59)

Where \sup denotes the least upper bound, σ_{\max} denotes the maximum singular value of a matrix, and $j = \sqrt{-1}$.

$\|G(s)\|_\infty$ actually gives out the peak gain of $G(s)$ across all frequencies. Hence, a minimum threshold for setting fault alarms can be given by:

$$\min(\lambda_f) = \|G(s)\|_\infty \omega_0$$
(60)

With the minimum threshold, the fault detection can be carried out by

$$\left\{ \begin{aligned} \|e_y(t)\| < \min(\lambda_f): & \text{ No fault occurs} \\ \|e_y(t)\| \geq \min(\lambda_f): & \text{ A fault has occurred} \end{aligned} \right\}$$
(61)

Remark 3: With the threshold given in (60), there still exist a possibility that the fault detector is insensitive to some kinds of fault which may result in small output residuals compared with the threshold. In this case, the

adaptive diagnostic algorithm will not be activated after fault occurring, and the system performance can only be guaranteed by the robustness of the controller designed. Since our emphasis is put on avoiding the false alarm due to disturbance, other fault detection methods which may be more sensitive to the faults will not be discussed in detail in this paper. Actually, as will be shown in next section, the controller is robust enough to guarantee the tracking performance under serious faults while the adaptive diagnostic algorithm is not activated.

5. APPLICATION EXAMPLE

For illustration, the design techniques are applied to the flight control of a Zagi flying wing aircraft (Beard & McLain, 2011). In this example, the control objective is to track the pitch angular and roll angular commands. In the straight and level trim condition with airspeed 10 (m/s) and altitude 50 (m), a linearized model can be obtained as follows:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) \end{aligned}$$

$$A = \begin{bmatrix} 0 & 0 & 0 & 1 & -0.0001 \\ 0 & 0 & 1 & 0 & 0.1665 \\ 0 & 0 & -2.5369 & 0 & 1.3228 \\ 0 & 0 & -0.0000 & -5.6319 & 0 \\ 0 & 0 & 0.1817 & 0 & -3.4009 \end{bmatrix},$$

$$B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 4.8744 & 6.3103 \\ -20.8139 & 0 & 0 \\ 0 & 3.6834 & -1.8480 \end{bmatrix}, C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

where the states $x = [\theta, \phi, p, q, r]^T$ represent the pitch angle (rad), roll angle (rad), and roll rate (rad/s), pitch rate (rad/s), yaw rate (rad/s) in body frame. The controls $u = [\delta_e, \delta_a, \delta_r]^T$ represent the deflection angles (rad) of elevator, aileron and rudder respectively. The control effectiveness matrix B is a normalized control matrix such that the control inputs are constrained by the unitary limits.

To compute the controller gains with the design method in Section 4.1, it is assumed that $E = \mathbf{I}_{5 \times 1}$, $R_0 = 10\mathbf{I}_7$, $R_\infty = \mathbf{I}_7$, $r_0 = 1.5$, $\rho = 1$, $\omega_0 = 1$, $\underline{\lambda}_i = 0.2$, $\bar{\lambda}_i = 1$, $i = 1, 2, \dots, m$, $\sigma = 10^5$. Then by solving the optimization problem (51) with (53), it is obtained that $\mu^* = 0.22$, $\beta^* = 10.1667$, and

$$P = \begin{bmatrix} 0.5505 & -0.3489 & -0.1205 & 0.1068 & -0.0567 & -0.2250 & 0.1294 \\ -0.3489 & 0.4144 & 0.1617 & -0.0543 & 0.0704 & 0.1420 & -0.1615 \\ -0.1205 & 0.1617 & 0.0805 & -0.0169 & 0.0318 & 0.0468 & -0.0628 \\ 0.1068 & -0.0543 & -0.0169 & 0.0330 & -0.0082 & -0.0430 & 0.0186 \\ -0.0567 & 0.0704 & 0.0318 & -0.0082 & 0.0249 & 0.0220 & -0.0265 \\ -0.2250 & 0.1420 & 0.0468 & -0.0430 & 0.0220 & 0.1079 & -0.0549 \\ 0.1294 & -0.1615 & -0.0628 & 0.0186 & -0.0265 & -0.0549 & 0.0764 \end{bmatrix}$$

$$K = \begin{bmatrix} 34.0816 & -17.4027 & -5.4353 & 10.1004 & -2.6318 & -13.7182 & 5.9938 \\ 33.3249 & -43.7326 & -20.7667 & 4.7188 & -10.6075 & -12.9678 & 16.7942 \\ 27.1748 & -36.9970 & -18.2546 & 3.7820 & -6.2969 & -10.5928 & 14.4132 \end{bmatrix}$$

For design of the fault detector, the desired poles for $A - LC$ are assumed to be $\{-1, -2, -3, -4, -5\}$. Then by pole placement, it is obtained that

$$L = \begin{bmatrix} 1.3112 & -0.4096 & -0.0615 & 2.3566 & 0.0939 \\ -0.5300 & 2.1190 & -1.4849 & 0.4902 & 0.3049 \end{bmatrix}^T$$

It follows from (60) that a minimum threshold for setting fault alarms can be given by

$$\min(\lambda_f) = 1.6424$$

To verify the tracking performance of the designed controller under fault situations, the nonlinear model with 6 degree of freedom is used, and it is assumed that the effectiveness factor of the elevator is reduced to be 0.2 at $t_f = 15$, and the effectiveness factors of both aileron and rudder are reduced to be 0.2 at $t_f = 55$. The learning rates for the adaptive diagnostic algorithm are specified by $\gamma_i = 100$, $i = 1, 2, \dots, m$. The reference commands for the pitch angle and the roll angle are both given by the square signals with time period of 20 seconds each, and the amplitudes for both maneuvers are 10 degrees.

Then through simulation with the nonlinear model of the aircraft, the tracking results are given by Fig. 3. For comparison, the tracking results in normal case are also presented in this figure. It is obvious that good performance is achieved for both tracking of the pitch angle and roll angle commands. Though the effectiveness loss of elevator at $t_f = 15$ has impaired the tracking performance, it is recovered quickly. This is actually contributed by excellent function of our integrated fault detection, adaptive diagnosis and reconfiguration algorithm. After malfunction of the elevator, the output residuals exceed the threshold for fault alarm as shown in Fig. 4. Then the adaptive diagnostic algorithm presented in (20) is activated to start the fault estimation process, which is shown in Fig. 5. Due to fast estimation of the effectiveness factor of the elevator, according to the control law in (10), the effectiveness loss is compensated quickly as shown in Fig. 6, which results in good tracking performance under fault condition as shown in Fig. 3.

In addition, from Fig. 4, it can be found out that the residuals in normal case are not equal to zero, which results from the un-modeled dynamics of the aircraft. However, their values are smaller than thresholds. Hence, a false alarm has been avoided by using the fault detection method proposed in Section 4.

For effectiveness loss of both aileron and rudder at $t_f = 55$, the output residuals are smaller than the threshold, and these faults have not been detected. However, good tracking

performance of the roll angle command can still be achieved as shown in Fig. 3 due to strong robustness of the controller designed. Since the effectiveness loss of aileron and rudder is not compensated, the responses of these two actuators are not approaching those in normal case as shown in Fig. 6.

For information, some other state variables from the nonlinear model are also given as in Fig. 7, which indicates that the aircraft has reached new equilibrium points under both the normal case and the fault case. These states are

X_e, Y_e, Z_e for aircraft position in inertial frame, U, V, W for aircraft velocity in body frame, and ψ for yaw angle. From Fig. 7, it can be seen that the main influence of effectiveness loss of elevator is on the pitch rate, while the effectiveness loss of aileron and rudder mainly affect the roll rate, yaw rate, and lateral-directional velocity in body frame. For an intuitive comparison, the 3D trajectories of the aircraft under both normal case and fault case are also presented in Fig. 8.

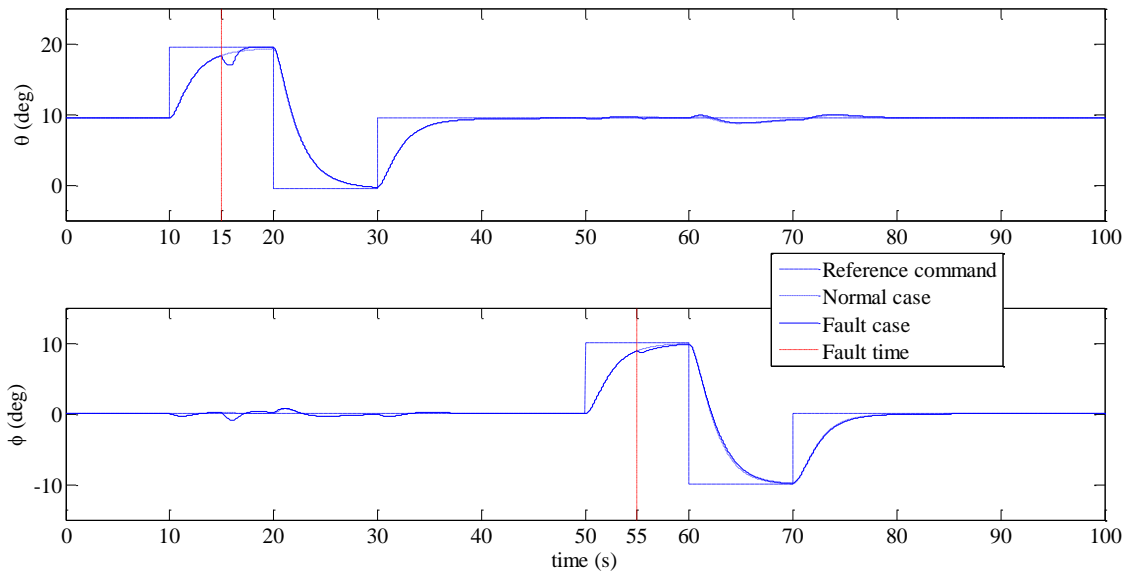


Fig. 3 Tracking of Pitch Angle and Roll Angle Command

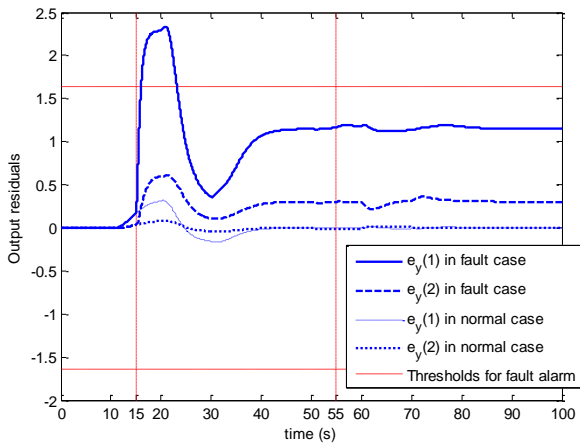


Fig. 4 Output residuals

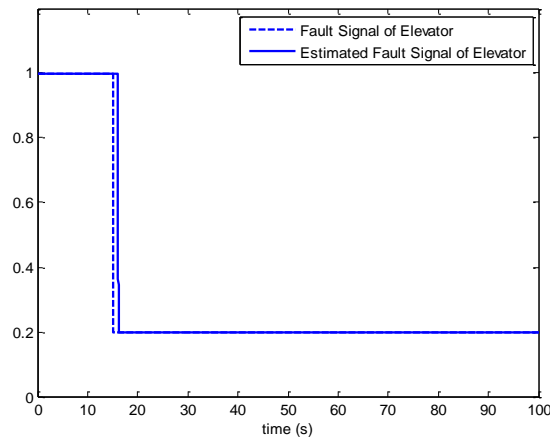


Fig. 5 Effectiveness factor estimation

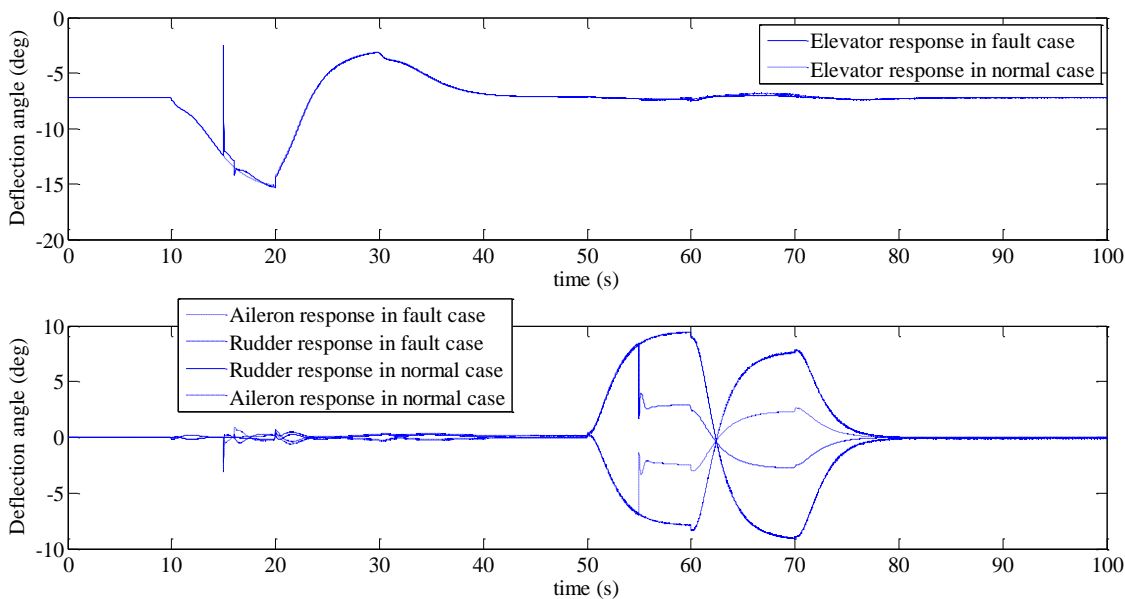


Fig. 6 Actuator outputs

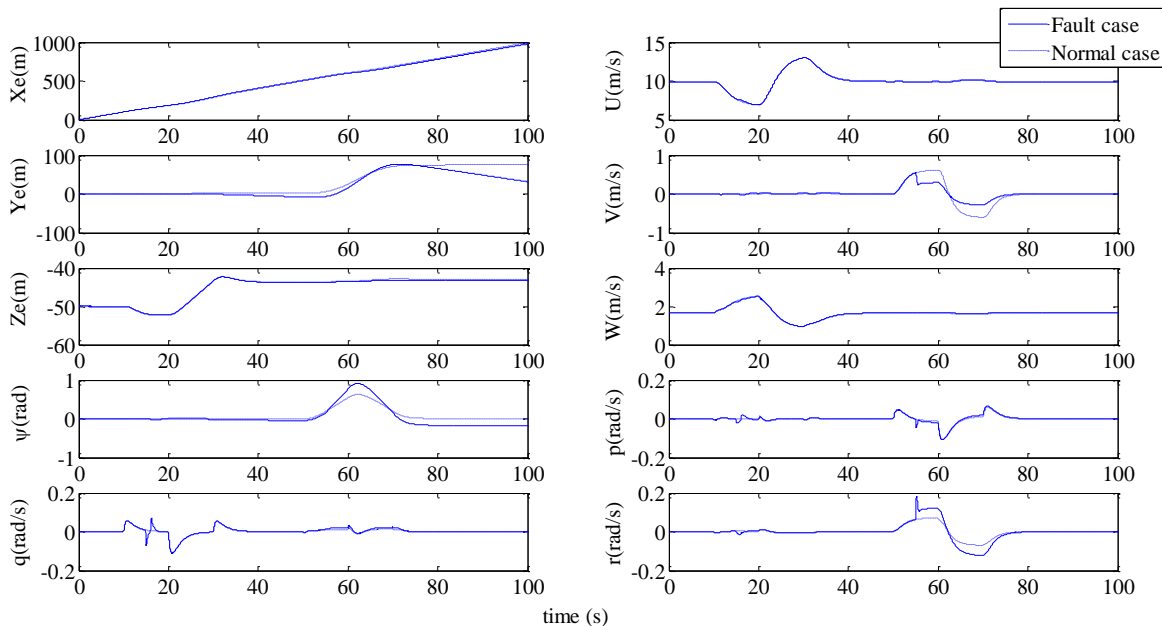


Fig. 7 Other states of aircraft

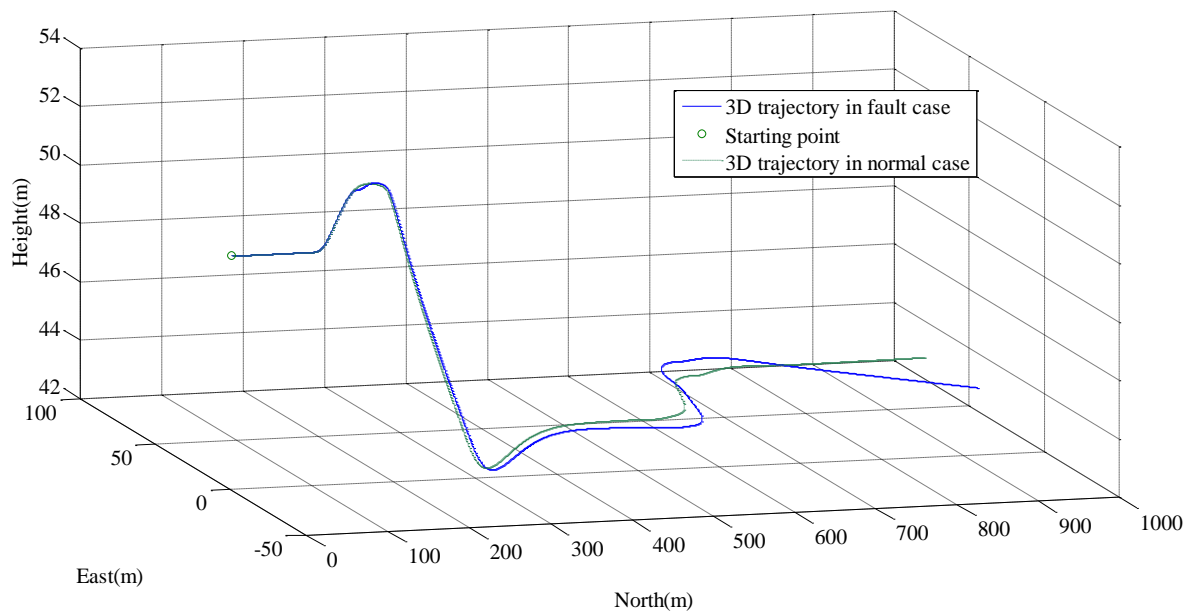


Fig. 8 3D trajectories of the aircraft

6. CONCLUSION

An integrated active fault-tolerant control method against partial loss of actuator effectiveness and saturation is proposed in this paper. LMI conditions are presented to compute the design parameters by integrated design of reconfigurable controller and fault diagnosis module. An observer is designed to detect a fault, and a minimum threshold is set to avoid the false alarm induced by disturbances. The system performance is described by two ellipsoidal sets regarding the domain of attraction and disturbance rejection respectively.

The proposed design techniques are applied to flight control of a flying wing aircraft under actuator faults. The nonlinear model of the aircraft is used for simulation, and satisfactory tracking performance can be obtained. The effectiveness loss of the elevator can be detected and compensated by the proposed integrated design method. However, the fault detector proposed in this paper is not sensitive to the faults of both aileron and rudder, though good tracking performance can still be achieved. This should be improved in our future work.

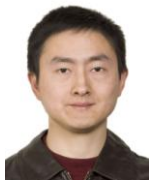
ACKNOWLEDGEMENT

The work reported in this paper is partially supported by the China Scholarship Council (CSC) and the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

- Zhang, Y. M., & Jiang, J. (2008). Bibliographical review on reconfigurable fault-tolerant control systems. *Annual Reviews in Control*, 32(2), pp. 229-252.
- Isermann, R. (2006). *Fault-diagnosis systems: An introduction from fault detection to fault tolerance*. Berlin, Germany: Springer.
- Tarbouriech, S., & Turner, M. (2009). Anti-windup design: An overview of some recent advances and open problems. *IET Control Theory & Applications*, 3(1), pp. 1-19.
- Tarbouriech, S., Pittet, C., & Burgat, C. (2000). Output tracking problem for systems with input saturations via nonlinear integrating actions. *Int. J. of Robust and Nonlinear Control*, 10(6), pp. 489-512.
- Hu, T., Lin, Z., & Chen, B. M. (2002). An analysis and design method for linear systems subject to actuator saturation and disturbance. *Automatica*, 38(2), pp. 351-359.
- Bodson, M., & Pohlchuck, W. A. (1998). Command limiting in reconfigurable flight control. *Journal of Guidance, Control, and Dynamics*, 21(4), pp. 639-646.
- Zhang, Y. M., & Jiang, J. (2003). Fault tolerant control system design with explicit consideration of performance degradation. *IEEE Transactions on Aerospace and Electronic Systems*, 39(3), pp. 838-848.

- Zhang, Y. M., Jiang, J., & Theilliol, D. (2008). Incorporating performance degradation in fault tolerant control system design with multiple actuator failures. *International Journal of Control, Automation, and Systems*, 6(3), pp. 327-338.
- Pachter, M., Chandler, P. R., & Mears, M. (1995). Reconfigurable tracking control with saturation. *Journal of Guidance, Control, and Dynamics*, 18(5), pp. 1016-1022.
- Guan, W., & Yang, G.-H. (2009). Adaptive fault-tolerant output-feedback control of LTI systems subject to actuator saturation. *Proceedings of American Control Conference (2569-2574)*, 10-12 June 2009, St. Louis, MO, USA.
- Hu, T., & Lin, Z. (2001). Control systems with actuator saturation: Analysis and design. Boston: Birkhäuser.
- Wu, F., Lin, Z., & Zheng, Q. (2007). Output feedback stabilization of linear systems with actuator saturation. *IEEE Transactions on Automatic Control*, 52(1), pp. 122-128.
- Vandenberghe, L., Boyd, S., & Wu, S.-P. (1998). Determinant maximization with linear matrix inequality constraints. *SIAM Journal on Matrix Analysis and Applications*, 19(2), pp. 499-533.
- Beard, R. W., & McLain, T. W. (2011). *Small unmanned aircraft: Theory and practice*. Princeton, New Jersey, USA: Princeton University Press.



Jinhua Fan received his M.S. degree in 2007 from Dept. of Automatic Control, National University of Defense Technology, Changsha, Hunan, China. Currently, he is currently a visiting researcher in the Dept. of Mechanical and Industrial Engineering at Concordia University, Canada, sponsored by China Scholarship Council (CSC) from Sept. 2010 to Sept. 2012.

His research interests are in fault detection, diagnosis and fault-tolerant control algorithms.



Youmin Zhang received his Ph.D. degree in 1995 from the Department of Automatic Control, Northwestern Polytechnical University, Xian, China. He is currently an Associate Professor with the Department of Mechanical and Industrial Engineering at Concordia University, Montreal, Canada. He held several teaching and research positions in Northwestern Polytechnical University, University of New Orleans, Louisiana State University, State University of New York at Binghamton, The University of Western Ontario, and Aalborg University, respectively. His main research interests and experience are in the areas of condition monitoring, fault diagnosis and fault-tolerant (flight) control systems; cooperative guidance, navigation and control of unmanned aerial/ground vehicles; dynamic systems modeling, estimation, identification and control; and advanced signal processing techniques for diagnosis, prognosis and health management of safety-critical systems and manufacturing processes. He has published 4 books, over 200 journal and conference papers. He is a senior member of AIAA, a senior member of IEEE, a member of IFAC Technical Committee on Fault Detection, Supervision and Safety for Technical Processes (SAFEPROCESS) and AIAA Infotech@Aerospace Program Committee on Unmanned Systems, and a member of Association for Unmanned Vehicle Systems International (AUVSI), Unmanned Systems Canada (USC), Canadian Aeronautics and Space Institute (CASI), Canadian Society for Mechanical Engineering (CSME). He is an editorial board member of several international journals and an IPC member and session chair/co-chair of many international conferences.

Zhiqiang Zheng received his Ph.D. degree in 1994 from University of Liege, Belgium. He is currently a Professor in Dept. of Automatic Control, National University of Defense Technology, Changsha, Hunan, China. His main research interests are in guidance and control for aircrafts, and formation control for mobile robots. He published over 150 journal and conference papers.

Multiple Fault Diagnostic Strategy for Redundant System

Yang Peng, Qiu Jing, Liu Guan-jun , and Lv Ke-hong

Institute of Mechatronics Engineering, College of Mechatronics and Automation

National University of Defense Technology

Changsha City, Hunan Province, 410073, China

Nudtyp7894@163.com, qiuqing@nudt.edu.cn, gjliu342@qq.com

ABSTRACT

It is difficult to diagnose the faults, especially multiple faults, in redundant systems by traditional diagnostic strategies. So the problem of multiple fault diagnostic strategy for redundant system was researched in this paper. Firstly, the typical characters of multiple faults (minimal faults) were analyzed, and the problem was formulated. Secondly, a pair of two-tuples were applied to denote the possible and impossible diagnostic conclusion at different diagnostic stages respectively, and a multiple fault diagnostic inference engine was constructed based on Boolean logic. The inference engine can determine the system diagnostic conclusions after executing each test, and determine whether a repair action was needed, and further determine whether a next test was needed. Thirdly, a method determining the next best test was presented. Based on the proposed inference engine and test determining method, a multiple fault diagnostic strategy was constructed. Finally, a simulation case and a certain flight control system were applied to illustrate the proposed diagnostic strategy. The simulation and practical data computational results show that the presented diagnostic strategy can diagnose multiple faults in redundant systems effectively and it is of certain application value.

1. INTRODUCTION

With the rapid development of aviation projects, the designers have attached more importance to

system reliability and safety. In the aviation domain, redundant techniques are usually adopted to improve system reliability. At the same time, as technology advances, there is a significant increase in the complexity and sophistication of aviation systems, which can easily induce multiple faults in all probability. Hence, studying the problem of multiple fault diagnosis in redundant systems is very important and significant. Unfortunately, there are little literatures referring to the problem at present.

A great variety of aviation systems with redundancy and with little or no opportunity for repair or maintenance during the operation may induce multiple faults, thus, a single failure assumption does not hold for this situation. Furthermore, the combinations of multiple faults may be of great multiplicity, and different fault combinations likely take on the same failure omen due to non-linearity, coupling and time-variance among the system components, and especially due to the redundant design in some systems. Thus, it becomes a difficult problem to diagnose multiple faults in redundant systems.

In the literature in the recent years, many scholars show great interesting on the multiple fault diagnostic problems^[1-6]. Nevertheless, the premise of multiple fault diagnosis is enough sensor data acquired by executing multiple tests simultaneously. In practical application, tests are executed sequentially rather than simultaneously in most cases, so it is imperative important to

study multiple fault sequential diagnostic strategy problem. Shakeri et al [7,8] have studied the problem based on sequential test and presented a multiple fault diagnostic optimization generation method, known as Sure strategies. The paper mainly considers the problem of multi-fault sequential diagnosis in redundant systems.

2. PROBLEM FORMULATION

The diagnostic strategy problem is defined by the five-tuple $(\mathbf{F}, \mathbf{P}, \mathbf{T}, \mathbf{C}, \mathbf{B})$, where $\mathbf{F} = \{f_1, \dots, f_m\}$ is a set of independent failure sources, $\mathbf{P} = [P(f_1), \dots, P(f_m)]$ is the a priori probability vector associated with the set of failure sources \mathbf{F} , $\mathbf{T} = \{t_1, t_2, \dots, t_n\}$ is a finite set of n available binary outcome tests, $\mathbf{C} = \{c_1, c_2, \dots, c_n\}$ is a set of test costs and $\mathbf{B} = [b_{ij}]_{m \times n}$ is fault-test dependency matrix where $b_{ij} = 1$ if test t_j detects f_i , otherwise $b_{ij} = 0$.

The form of multiple faults are of great diversity, moreover, multiple faults refer to complex fault mechanism and relate closely to the practical application environment and the specific objects. In order to simply the problem, the paper mainly considers the multiple faults with additivity. Let's define $\mathbf{FS}(f_i) = \{t_j | b_{ij} = 1, 1 \leq j \leq n\}$ to denote the signature of failure state f_i , it indicates all the tests that monitor failure state f_i , $\mathbf{FS}(f_i, f_j)$ denotes the failure signature of the multiple faults, f_i and f_j . If they both satisfy additivity, then

$$\mathbf{FS}(f_i, f_j) = \mathbf{FS}(f_i) \cup \mathbf{FS}(f_j) \tag{1}$$

Nevertheless, there exist many multiple faults which don't satisfy Eq.(1), especially in systems with redundancy. Consider the digraph model in Figure1. The AND nodes α_1 and α_2 show the system is redundant. If only f_3 or f_4 occurs individually, t_2 will not detect them, yet if they both occur, t_2 can detect them, hence $\mathbf{FS}(f_3) \cup \mathbf{FS}(f_4) \neq \mathbf{FS}(f_3, f_4)$. The fault combination $\{f_3, f_4\}$ related to the AND node α_2 is usually termed minimal fault, which can be considered as a special fault state in multiple fault redundant analysis. The minimal faults for the example are $s_5 = \{f_3, f_4\}$ and $s_6 = \{f_1, f_2, f_3\}$.

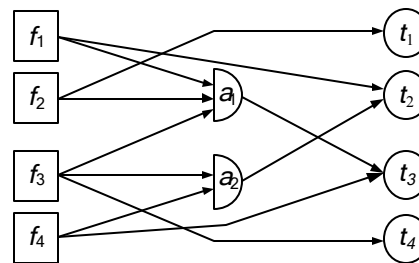


Figure 1. An example system with redundancy

In fault-tolerant systems, the failure sources \mathbf{S} can be derived by adding minimal faults to the single fault set, i.e., $\mathbf{S} = \{s_1, s_2, \dots, s_z\}$, where $s_i = \{f_i\}$ for $1 \leq i \leq m$ corresponds to each single failure source respectively, and $s_{m+1} \sim s_z$ corresponds to minimal fault of the system respectively.

In Figure1, $\mathbf{S} = \{\{f_1\}, \{f_2\}, \{f_3\}, \{f_4\}, \{f_3, f_4\}, \{f_1, f_2, f_3\}\}$. The corresponding fault-test dependency matrix $\mathbf{D} = [w_{ij}]_{n \times z}$ is shown in Table 1.

Table1. The extended dependency matrix of the example

Failure sources	Tests			
	t_1	t_2	t_3	t_4
$s_1 = \{f_1\}$	0	1	0	0
$s_2 = \{f_2\}$	1	0	0	0
$s_3 = \{f_3\}$	0	0	0	1
$s_4 = \{f_4\}$	0	0	1	0
$s_5 = \{f_3, f_4\}$	0	1	1	1
$s_6 = \{f_1, f_2, f_3\}$	1	1	1	1

Through the previous analysis, the multiple fault diagnostic strategy problem in redundant systems can be defined by the five-tuple $(\mathbf{S}, \mathbf{P}^*, \mathbf{T}, \mathbf{C}, \mathbf{D})$, where \mathbf{S} denotes the extended fault states, \mathbf{P}^* denotes the fault probability vector, $\mathbf{P}^*(s_i) = P(f_i)$ for $1 \leq i \leq m$, $\mathbf{P}^*(s_i)$ for $m+1 \leq i \leq z$ equals the product of correlation single fault sets. \mathbf{D} denotes the extended dependency matrix. \mathbf{T} and \mathbf{C} have the same meaning as defined before.

3. BOOLEAN LOGICAL INFERENCE ENGINE

Usually, a test is not enough to unambiguously isolate failure sources. However, according to the outcomes of the test, faults can be divided into possible diagnostic conclusion and impossible diagnostic conclusion. Based on the ideal and using the compact set conception provided by Shakeri, let the two-tuple (\mathbf{X}, \mathbf{G}) describes diagnostic conclusion of the system at different time, where $\mathbf{X} = \{x_k | x_k \subset \mathbf{S}^*\}$ and $\mathbf{G} (\mathbf{G} \subseteq \mathbf{S}^*)$ denote

possible and impossible diagnostic conclusion at present time epoch respectively, and $\mathbf{S}^* = \mathbf{S} \cup \{s_0\}$, s_0 denotes fault-free conclusion. Set \mathbf{X} is complete and is a set cluster consisting of compact sets. Compact set x ($x \in \mathbf{X}$) denotes the possible diagnostic conclusion, which is consistent to the known test outcomes and composed of minimal faults. If there is given (\mathbf{X}, \mathbf{G}) , where $\mathbf{X} = \{x_1, x_2, \dots, x_q\}$, then $x_k \in \mathbf{X}$, $(x_k \cap \mathbf{G}) = \emptyset$. Denote diagnostic conclusion corresponding to PASS and FAIL outcomes of the test t_j by $(\mathbf{X}_{jp}, \mathbf{G}_{jp})$ and $(\mathbf{X}_{jf}, \mathbf{G}_{jf})$ respectively.

The impossible diagnostic conclusion, \mathbf{G}_{jp} and \mathbf{G}_{jf} , can be calculated by:

$$\begin{aligned} \mathbf{G}_{jp} &= \{s_i | s_k \subseteq s_i, \forall s_k \in \mathbf{G} \cup \mathbf{TS}(t_j)\} \\ \mathbf{G}_{jf} &= \mathbf{G} \cup \{s_0\} \end{aligned} \quad (2)$$

Where $\mathbf{TS}(t_j) = \{s_i | w_{ij}=1, \text{for } 1 \leq i \leq z\}$ denotes the signatures of test t_j , indicating all the failure states detectable by test.

The possible diagnostic conclusion, \mathbf{X}_{jp} and \mathbf{X}_{jf} , can be get through the following steps.

First, \mathbf{X}_{jp} and \mathbf{X}_{jf} can be expressed by:

$$\begin{aligned} \mathbf{X}_{jp} &= \sum_{x_k \cap \mathbf{G}_{jp} = \emptyset} x_k \cdot \sum_{s_i \notin \mathbf{G}_{jp}} (1 - w_{ij}) s_i \\ \mathbf{X}_{jf} &= \sum_{x_k \cap \mathbf{G}_{jf} = \emptyset} x_k \cdot \sum_{s_i \notin \mathbf{G}_{jf}} w_{ij} s_i \end{aligned} \quad (3)$$

Then, let (3) expand to and/or expressions, and simply them based on the following logical rules.

$$\begin{aligned} s_k \cdot s_k &= s_k, s_i + s_i = s_i \\ s_0 \cdot s_i &= s_i, s_k + s_k \cdot s_i = s_k \end{aligned} \quad (4)$$

where sign “ \cdot ” denotes logical multiplication operation, $s_1 \cdot s_2$ denotes that the two faults occur simultaneously; sign “ $+$ ” denotes logical add operation, and shows that at least one of the two faults occurs.

The further simplification of (3) can be based on the rule 1.

Rule1: If $s_i \subset s_k$, then $s_i \cdot s_k = s_k$; if $s_i \cup s_j = s_k$, then $s_i \cdot s_j = s_k$. For example, in table 1, $s_3 \cdot s_5 = s_4$, $s_5 = s_5$, and $s_3 \cdot s_4 = s_5$ due to $s_3, s_4 \subset s_5$ and $s_3 \cup s_4 = s_5$

At last, eliminate compact sets which include elements of \mathbf{G} , and get possible diagnostic conclusion, \mathbf{X}_{jp} and \mathbf{X}_{jf} .

Consider the data in Table1. Initially, the diagnostic conclusion (\mathbf{X}, \mathbf{G}) is $(\mathbf{F}^*, \emptyset)$, where $\mathbf{F}^* = \{\{s_0\}, \{s_1\}, \dots, \{s_6\}\}$. Provided four test are executed, and t_1 PASS, t_2, t_3, t_4 FAIL. First, derive the impossible diagnostic conclusion $\mathbf{G} = \{s_0, s_2, s_6\}$ according to (2); then get the possible diagnostic conclusion based on (3) (4), $\mathbf{X} = s_1, s_3, s_4 + s_5$; simply it to $\mathbf{X} = s_5$ according to rule 1. In the form of set, the possible diagnostic conclusion is $\mathbf{X} = \{\{s_5\}\}$.

After getting the diagnostic conclusion (\mathbf{X}, \mathbf{G}) , use the rule 2 to determine whether the repair operations are needed.

Rule2: If $|\mathbf{X}|=1$, all the faults in \mathbf{X} should be repaired; if no test gives any information gain, i.e., $\mathbf{X}_{jp} = \emptyset$ or $\mathbf{X}_{jf} = \emptyset$ for $t_j \in \mathbf{T}$, then all the faults in \mathbf{X} should be repaired too.

Refresh the diagnostic conclusion after repair operations based on rule 3.

Rule3: If fault state s_i has been repaired, then refresh the diagnostic conclusion according to (5).

$$\begin{aligned} \mathbf{G}' &\leftarrow \mathbf{G} \cup \{s_k \in \mathbf{S} | s_k \cap s_i \neq \emptyset\} - \{s_0\} \\ \mathbf{X}' &\leftarrow \mathbf{S}^* - \mathbf{G}' \end{aligned} \quad (5)$$

4. MULTI-FAULT DIAGNOSTIC STRATEGY

Multi-fault diagnostic strategy is constructed as follows: first, judge whether the candidate tests can provide information at the present diagnostic conclusion (\mathbf{X}, \mathbf{G}) . If not, replace all the candidate fault components; otherwise, select the best test according to the heuristic function, then, judge the system states according to outcomes of the test. The heuristic function used to guide test selection is given by:

$$j^* = \arg \max_{t_j \in \mathbf{T}} \left\{ \frac{IG(\mathbf{X}; t_j)}{c_j} \right\} \quad (6)$$

where c_j corresponds cost of t_j , $IG(\mathbf{X}; t_j)$ denotes average mutual information between test t_j and possible diagnostic conclusion \mathbf{X} . The (6) means that the test with maximal diagnostic information per cost should be selected with a priority.

$IG(\mathbf{X};t_j)$ is calculated by:

$$IG(\mathbf{X};t_j) = - \left\{ \frac{P(\mathbf{X}_{j^p})}{P(\mathbf{X})} \ln \frac{P(\mathbf{X}_{j^p})}{P(\mathbf{X})} + \frac{P(\mathbf{X}_{j^f})}{P(\mathbf{X})} \ln \frac{P(\mathbf{X}_{j^f})}{P(\mathbf{X})} \right\} \quad (7)$$

Given $\mathbf{X}=\{x_1, x_2, \dots, x_z\}$ and $x_k=\{s_{kl}, \dots, s_{kq}\}$, so $P(\mathbf{X})$ can be calculated by:

$$P(\mathbf{X}) = 1 - \prod_{k=1}^z \left(1 - \left(\prod_{j=1}^q P(s_{kj}) \right) \right) \quad (8)$$

Especially, if $\mathbf{X}_{j^p}=\emptyset$ or $\mathbf{X}_{j^f}=\emptyset$, then $IG(\mathbf{X};t_j)=0$.

If there exist more than one (≥ 2) compact sets after executing the test t_j , then, do further according to rule 4.

Rule4: Given the dimensions of the possible diagnostic conclusion satisfies $|\mathbf{X}| \geq 2$. If $t_j \in \mathbf{T}$, $IG(\mathbf{X};t_j)=0$, then repair all the faults in \mathbf{X} , refresh diagnostic conclusion after repair according to (5) otherwise, select the next best test according to (6).

Multi-Fault Diagnostic Strategy Generation Algorithm

Step1: Input the basic data ($\mathbf{S}, \mathbf{P}^*, \mathbf{T}, \mathbf{C}, \mathbf{D}$), and create ψ used to store diagnostic nodes. Initially, $\psi = \{(\mathbf{F}^*, \emptyset)\}$, create an empty set D used to store the decision tree.

Step2: Repeat the following steps until $\psi = \emptyset$, output D.

2.1: Select a diagnostic node from ψ , denoting it by (\mathbf{X}, \mathbf{G}) , and put it in D, analyze dimension of \mathbf{X} .

2.2: **IF** $|\mathbf{X}|=1$, **THEN**

IF $x \cup \mathbf{G} = \mathbf{S}^*$, ($x \in \mathbf{X}$), **THEN**

-Action: remove \mathbf{X} from ψ to D.

Return.

IF $x \cup \mathbf{G} \neq \mathbf{S}^*$, ($x \in \mathbf{X}$), **THEN**

-repair all the faults in x,

-generate a new diagnostic node $(\mathbf{X}, \mathbf{G}')$ under the node (\mathbf{X}, \mathbf{G}) and store it in ψ

Return.

ELSE IF $|\mathbf{X}| > 1$, **THEN**, calculate possible conclusion set after each candidate test, e.g., after test t_j , denote the possible conclusion set by \mathbf{X}_{j^p} and \mathbf{X}_{j^f} . Calculate

diagnostic information of each candidate test.

IF no test **give** any information, viz., \mathbf{X}_{j^p} (or $\mathbf{X}_{j^f}) = \emptyset$ for $t_j \in \mathbf{T}$, **THEN**

-repair all the faults in \mathbf{X} ,

-remove \mathbf{X} from ψ .

Return.

IF there exist some tests giving diagnostic information, **THEN**

-select the best test according Eq.(6), denoting it by t_k , store the new diagnostic conclusion $(\mathbf{X}_{k^p}, \mathbf{G}_{k^p})$ and $(\mathbf{X}_{k^f}, \mathbf{G}_{k^f})$ produced by t_k in ψ and the test t_k in D.

Return.

5. APPLICATION STUDY

A simulation example with five failure nodes, five test nodes and an AND node is used to verify the presented algorithm. The multi-signal flow model of the system is shown in Figure2. The minimal fault is $\{f_1, f_3\}$, and the extended dependency matrix with failure state probability is shown in Table2. The minimal fault probability equals to the product of the correlation single faults.

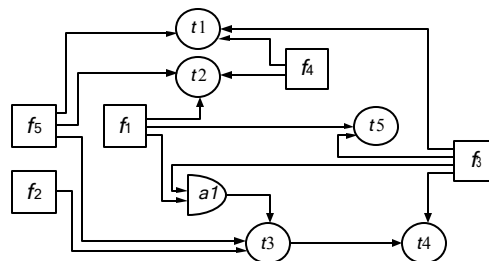


Figure 2. A simulation example with redundancy

Table2. The Extended dependency matrix with fault probability of the simulation example

Fault sources	tests					Fault probability
	t1	t2	t3	t4	t5	
$s_1=\{f_1\}$	0	1	0	0	1	0.014
$s_2=\{f_2\}$	0	0	1	1	0	0.027
$s_3=\{f_3\}$	1	0	0	1	1	0.125
$s_4=\{f_4\}$	1	1	0	0	0	0.068
$s_5=\{f_5\}$	1	1	1	1	0	0.146
$s_6=\{f_1, f_3\}$	1	1	1	1	1	0.002

The corresponding fault diagnostic tree applying the proposed algorithm is shown in Figure3.

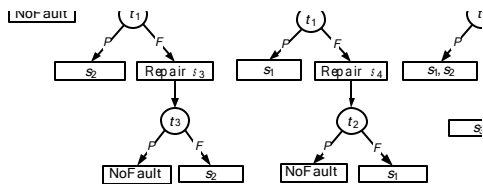


Figure 3. The multi-fault diagnostic tree

The fault tree has 11 leaf nodes, number the nodes sequentially from left to right, and analyze the masking false failures and hidden failures of each leaf node. The results are shown in Table3.

Form the results in Table3, obviously, there are no hidden failures and masking false failures in all diagnostic conclusions. Hence, realize multiple fault diagnosis for systems with redundancy effectively and accurately.

Note that there exist two shaded nodes in Figure3, x_{10} and x_{11} . Let denote the diagnostic conclusions of the two nodes by $(\mathbf{X}_{10}, \mathbf{G}_{10})$ and $(\mathbf{X}_{11}, \mathbf{G}_{11})$ respectively, it can be referred that: $\mathbf{G}_{10} = \{s_0, s_1, s_3, s_6\}$, $\mathbf{G}_{11} = \{s_0\}$, $\mathbf{X}_{10} = \{\{s_5\}, \{s_2, s_4\}\}$ and $\mathbf{X}_{11} = \{\{s_1, s_5\}, \{s_3, s_5\}, \{s_6\}, \{s_1, s_2, s_4\}, \{s_2, s_3, s_4\}\}$ respectively. It is obvious that the possible diagnostic conclusions are not unique, yet, all the tests have been selected, and no test can provide diagnostic information any more, so all the faults in \mathbf{X} should be repaired according rule 4. When the union of possible diagnostic conclusion and impossible diagnostic conclusion equals to \mathbf{S}^* , terminate the diagnostic process.

Table3. The hidden failures and masking false failures for each leafnode

Leafnodes	Passed tests	Repaired faults	Hidden failures	Masking false failures
$x_1 = \{s_0\}$	$D_{P(1)} = \{t_2, t_4\}$	\emptyset	\emptyset	\emptyset
$x_2 = \{s_2\}$	$D_{P(2)} = \{t_1, t_2\}$	\emptyset	\emptyset	\emptyset
$x_3 = \{s_0\}$	$D_{P(3)} = \{t_2, t_3\}$	$\{s_3\}$	\emptyset	\emptyset
$x_4 = \{s_2\}$	$D_{P(4)} = \{t_2\}$	$\{s_3\}$	\emptyset	\emptyset
$x_5 = \{s_1\}$	$D_{P(5)} = \{t_1, t_4\}$	\emptyset	\emptyset	\emptyset
$x_6 = \{s_0\}$	$D_{P(6)} = \{t_2, t_4\}$	$\{s_4\}$	\emptyset	\emptyset
$x_7 = \{s_1\}$	$D_{P(7)} = \{t_4\}$	$\{s_4\}$	\emptyset	\emptyset
$x_8 = \{s_1, s_2\}$	$D_{P(8)} = \{t_1\}$	\emptyset	\emptyset	\emptyset
$x_9 = \{s_3, s_4\}$	$D_{P(9)} = \{t_3\}$	\emptyset	\emptyset	\emptyset
$x_{10} = \{s_2, s_4, s_5\}$	$D_{P(10)} = \{t_5\}$	\emptyset	\emptyset	\emptyset
$x_{11} = \{s_2, s_4, s_5, s_6\}$	$D_{P(11)} = \emptyset$	\emptyset	\emptyset	\emptyset

Consider the digraph model of F18 Flight Control System (FCS) for the left Leading Edge Flap (LEF) in Figure4, which was used as an example in [9]. The minimal faults for the example are $\{FCCA, FCCB\}$, $\{FCCA, CHNL3\}$, $\{FCCB, CHNL2\}$, and $\{CHNL2, CHNL3\}$. The extended dependency matrix is shown in table 4.

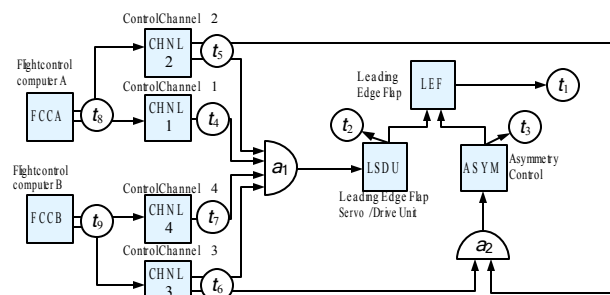


Figure 4. Digraph model of subsystem LEF

Table4. The extended dependency matrix of subsystem LEF

Failure sources	Tests									Probability
	t1	t2	t3	t4	t5	t6	t7	t8	t9	
$s_0 = \{\emptyset\}$	0	0	0	0	0	0	0	0	0	0.9906
$s_1 = \{LEF\}$	1	0	0	0	0	0	0	0	0	0.001
$s_2 = \{LSDU\}$	1	1	0	0	0	0	0	0	0	0.001
$s_3 = \{ASYM\}$	1	0	1	0	0	0	0	0	0	0.001
$s_4 = \{FCCA\}$	0	0	0	1	1	0	0	1	0	0.001
$s_5 = \{FCCB\}$	0	0	0	0	0	1	1	0	1	0.001
$s_6 = \{CHNL1\}$	0	0	0	1	0	0	0	0	0	0.001
$s_7 = \{CHNL2\}$	0	0	0	0	1	0	0	0	0	0.001
$s_8 = \{CHNL3\}$	0	0	0	0	0	1	0	0	0	0.001
$s_9 = \{CHNL4\}$	0	0	0	0	0	0	1	0	0	0.001
$s_{10} = \{FCCA, FCCB\}$	1	1	1	0	0	0	0	0	0	0.0001
$s_{11} = \{FCCA, CHNL3\}$	1	0	1	0	0	0	0	0	0	0.0001
$s_{12} = \{FCCB, CHNL2\}$	1	0	1	0	0	0	0	0	0	0.0001
$s_{13} = \{CHNL2, CHNL3\}$	1	0	1	0	0	0	0	0	0	0.0001

The diagnostic tree of subsystem LEF adopting the proposed reference engine and diagnostic strategy is shown in Figure5.

Obviously, the presented diagnostic strategy can diagnose multiple fault of the LEF correctly. The diagnostic tree is very complex due to many types of multiple faults. The traditional diagnostic strategies based on single fault assumption can't diagnose the multiple faults in redundant systems. For example, in the daily maintenance action of the LEF, if FCCA occurs fault, the single

assumption-based diagnostic strategy can't diagnose it due to the existing AND node α_1 , but if CHNL3 also occurs fault during next flight mission, it will result in severe accident. The proposed diagnostic strategy can efficiently and correctly diagnose these faults in subsystem LEF, so with higher application value in practical engineering.

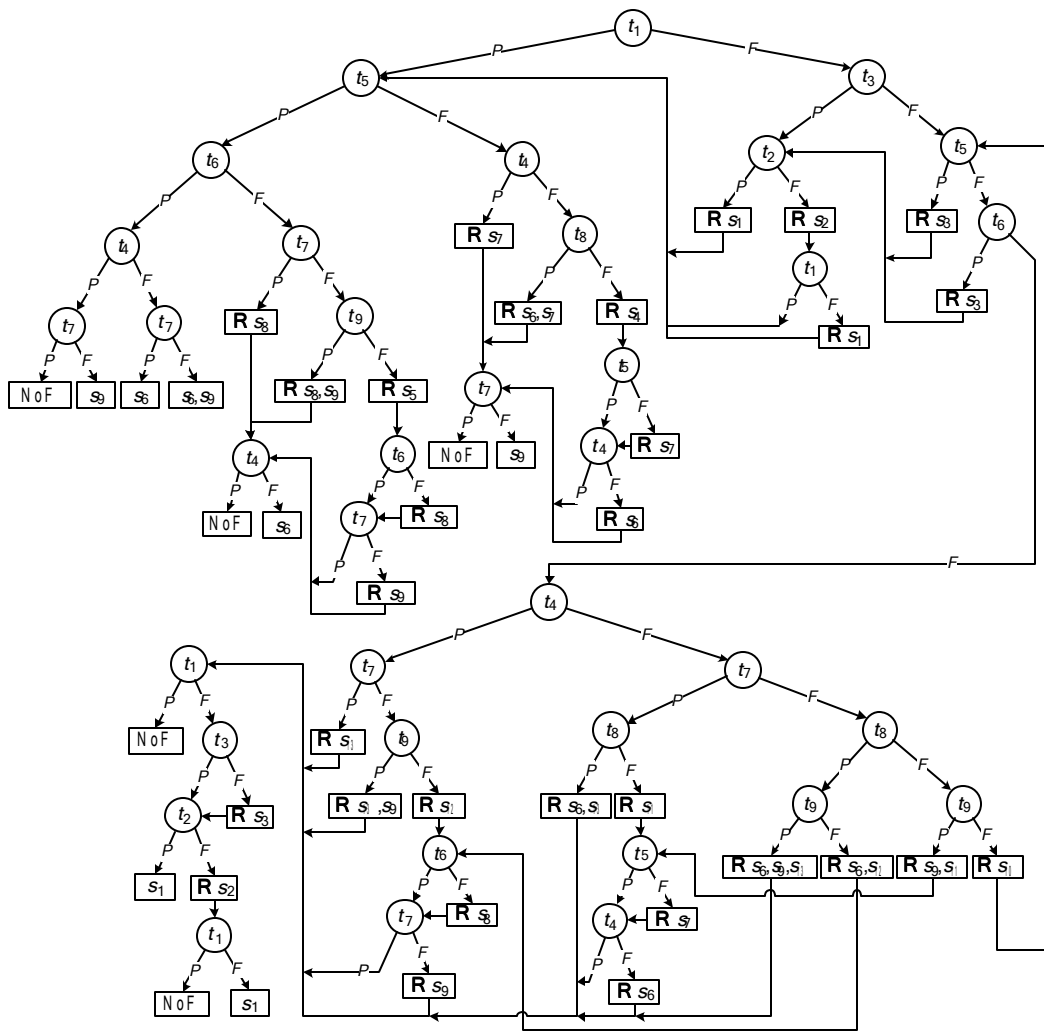


Figure 5. Diagnostic tree of subsystem LEF

6. CONCLUSIONS

The paper mainly considers the multiple fault diagnostic strategy problem arising in systems with redundancy. The paper first formulates the problem, then presents multiple fault inference engine based on Boolean logic and three additional inference rules. The inference engine

can be applied to determine the possible diagnostic conclusion and impossible diagnostic conclusion accurately after executing each test. Based on the knowledge, an efficient multiple fault diagnostic strategy for redundant systems is constructed. An efficient multiple fault diagnostic strategy for the F18 FCS is constructed by the proposed method. The analysis results show that

the strategy can diagnose multiple faults in the FCS, and can avoid missed diagnosis and false diagnosis. Hence, the proposed multiple fault diagnostic strategy can be applied to practical engineering.

REFERENCES

- [1] Simpson W R, Sheppard J W. Multiple Failure Diagnosis[C]. Proceedings of the IEEE Autotestcon, 1994:381-389.
- [2] Davis R. Retrospective on diagnostic reasoning based on structure and behavior [J]. Artificial Intelligence, 1993,59: 149-157.
- [3] Tu F, Pattipati K R, Deb S, et al. Computationally efficient algorithms for multiple fault diagnosis in large graph-based systems[J]. IEEE Transactions on Systems, Man and Cybernetics, 2003 33(1):73-85.
- [4] Long Bing, Jiang Xing-wei, Song Zheng-ji. Study on multiple fault diagnostic technique for aerospace craft based on multi-signal model[J]. Journal of astronavigation, 2004, 25(5):591-594.[in chinese]
- [5] Stefano Chessa, Paolo Santi. Operative Diagnosis of Graph-Based Systems with Multiple Faults.[J]. IEEE Transactions on Systems, Man and Cybernetics, 2001, 31(2):112-119.
- [6] Kleer J D. Diagnosing Multiple Faults. Artificial Intelligence. 1987, 32:97-130.
- [7] Shakeri M, Raghavan V, Pattipati K, et al. Sequential testing algorithms for multiple fault isolation[J]. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 2000, 30(1):1-14.
- [8] Shakeri M, Pattipati K R, Raghavan V, et al. Optimal and Near-Optimal Algorithms for Multiple Fault Diagnosis with Unreliable Tests [J]. IEEE trans on SMC,1998:431-440.
- [9] Doyle S A, Dugan J B, Patterson-Hine A. A quantitative analysis of the F18 flight control system[C]. American Institute of Aeronautics and Astronautics Computing in Aerospace 9 Conference proceedings, 1993:668-675.

Province, China, on Sep. 4, 1978. He received the B.S. degree in Communication Engineering from Xi'an Institute of Communication, M.S. degree in Circuit and System and Ph.D. degree in Mechatronic Engineering from the National University of Defense Technology, P. R. China in 2001, 2003 and 2008, respectively. From 2009, he worked as a post doctor in Institute of Mechatronic Engineering, National University of Defense Technology. Current research interests include design for testability, condition monitoring and fault diagnosis, mechanical signal processing, etc.

Qiu Jing was born in 1964. He received B.S. degree from Beijing University of Aeronautics and Astronautics in 1985, M.S. and Ph.D degrees in Mechatronic Engineering from National University of Defense Technology in 1988 and 1998, respectively. His research interests include fault diagnosis, reliability, testability, maintenance and etc.

E-mail: qiuqing@nudt.edu.cn

Liu Guan-jun was born in 1972. He received B.S. and Ph.D degrees in Mechatronic Engineering from National University of Defense Technology in 1994 and 2000, respectively. His research interests include fault diagnosis, testability, maintenance and etc.

E-mail: gjliu342@qq.com



Yang Peng was born in Hubei



Online Abnormality Diagnosis for real-time Implementation on Turbofan Engines and Test Cells

Jérôme Lacaille¹, Valério Gerez²

¹ *Snecma Algorithm Expert Villaroche, France*
jerome.lacaille@snecma.fr

² *Snecma System Engineer, Villaroche, France*
valerio.gerez@snecma.fr

ABSTRACT

A turbofan used in flight or in a bench test cell produces a lot of data. Numeric measurements describe the performance of the engine, the vibration conditions and more generally the behavior of the whole system (engine + bench or aircraft). It seems reasonable to embed an application capable of health diagnosis. This inboard monitoring system should use very light algorithms. The code need to work on old fashion FADEC calculators (Fault Authority Digital Engine Control) built on a technology dating more than 20 years. Snecma, as an engine manufacturer, has a great knowledge of the engine design and its behavior. This knowledge helps to select the best inputs for a good abnormality detection process, hence limiting the need of a too complex solution. In this article, I describe a very simple anomaly detection algorithm designed for embedding on light computers. This algorithm was validated on a bench test cell running a military engine.*

1. THE CONTEXT

During the development process, engine parts or prototypes are tested in a bench cell environment. The engine and the bench itself are connected to many sensors (more than a thousand). Afterward, during standard missions, only a subset of those sensors is kept; but the engine continues to send a lot of messages carrying a potential knowledge about its behavior. Our goal is to fetch a part of this information to be able to detect potential abnormalities. An application to detect anomaly should work during the test process on benches but also during real flights. One of our main

concerns is to develop a code that may be implemented on current computers such as FADEC or eventually ACMS (Aircraft Condition Monitoring System). Today's calculators in use do not have a big computing power, neither much memory. For example, it may be difficult to implement complex signature classification algorithms (Cômes 2010a, 2010b, 2011, Cottrell 2009, Lacaille 2009c, 2011). Most of the work should be done on the fly using mathematic filters with little amount of memory for calibration.

The volume of data available during bench tests is really huge. Analyzing simultaneously too many sensors will damage the quality of a mathematic computation, so we choose to build many small instances of the same algorithm. Each instance deals with a small (but reasonable) amount of measurements; it produces its own diagnostic outputs (detection and anticipation or prognostic). Each separate result is an indication of the behavior of a specific component according to specific faults. All such results are merged together by a higher decision layer. This complex fusion algorithm embeds a selection step which gives a great indication of what detectors (instances) should be implemented for on board work. This article presents the first part of the whole process: the detection layer called CES for *continuous empirical score*.

This work presents a specific code that is both light and efficient. Much solutions proposed to monitor engines and detect abnormalities are built around specific components and the code is generally cut in two parts for embedding acquisition and ground analysis (Lacaille and Nya Djiki 2009, Flandrois 2009, Lacaille 2009a, 2009c, 2011). Other types of monitoring algorithms are fully dedicated to ground analysis; in general they work on a larger time scale using successive flights to detect trends and build prognostics (Cômes 2010a, 2010b, 2011, Lacaille 2011). Those solutions need large databases and even use some

* This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

datamining analysis to preprocess the observations (Seichepine 2011).

The current proposition is a general detector of unusual behavior and it is built in one standalone application dedicated for embedded systems.

2. THE DETECTION ALGORITHM

2.1 Inputs

Looking at a specific component, one finds probable faults, corresponding signature indicators and running constraints. All this data is reported in the FMECA (Failure Mode, Effects, and Criticality Analysis) document which lists all failure modes of all components of each engine system with corresponding occurrence probabilities. Then for one specific component and a small list of potential faults it is possible to isolate a small amount of indicators divided in two subsets.

- The first subset describes the fault signatures; we call it the endogenous subset.
- The second one gives the constraints or the description of the execution context during which such fault may occur. This second set of inputs is called the exogenous subset because it describes the working conditions that should apply for a valid detection.

The two input subsets are not used the same way: the exogenous subset serves the identification of an acceptable context when the endogenous subset is only studied for abnormality detection when the context is accepted.

Each subset is made of indicators that do not necessary correspond to the raw measurements. A small preprocessing stage should be implemented. We select a set of online linear filters (moving averages and autoregressive filters) with the help of company experts. Eventually, mathematic computations, relations between sensor outputs, are also used in place of the initial measurements (Lacaille 2007). On more powerful computers one may try to mathematically reduce dimensionality using PCA (Principal Component Analysis) (Lacaille 2009c) or other advanced algorithms like the LASSO (Lacaille 2011b).

2.2 Outline

This algorithm is based on a very simple assumption: “most of the time the engine is working under normal conditions; then when something unusual happens, it may be easily detected as an outlier”.

How does it work? Look at a new input, the exogenous part of the input describes the context, then look at what happened for the endogenous indicators when this context “nearly” applies. If the endogenous observation resembles the already observed ones, everything is

usual: no abnormality. Otherwise the behavior is unusual: this is an anomaly.

Such algorithm needs some sort of memory to store normal conditions, a distance computation to compute *proximity* of context observations and a *score* which is another distance or likelihood to see if an observation is an outlier.

The computation is controlled by quality estimations:

- To define a context as usual or normal, a minimum amount of observations is needed. If the engine is often in such “running context” the quality of the “normal” flag is high otherwise it is not clear. The *adequacy* measurement computes an indicator of neighborhood. If new measurements are really new, for example “never observed”, the adequacy should be low (distance from current context to other measurements is high); otherwise the adequacy is high when the current and some already observed contexts are similar (distance between current and past context observations is low).
- When the context is clearly identified, the local variance of the endogenous indicators (on similar context/exogenous data) gives a *precision* indication[†]. This precision value indicates the quality of the outlier detection. When the precision is high, the endogenous indicators should be almost constant for a given running context. Hence the detection of an outlier is easy. However, if the precision is low, the variance of the endogenous measurements is high so the detection is a little fuzzier.

Adequacy and precision are used together to build some global quality indicator.

2.3 Definition of the proximity

In the exogenous domain, the main computation is the proximity. This is a distance between a current (exogenous) context and some stored observations.

We will note \mathbf{u} a vector of exogenous measurements. Let \mathbf{u}^* be the current observation and \mathbf{H}_u the set of historic exogenous measurements stored in a little database[‡]. The distance between a current \mathbf{u}^* and any $\mathbf{u} \in \mathbf{H}_u$ is noted $d(\mathbf{u}^*, \mathbf{u}) = \|\mathbf{u}^* - \mathbf{u}\|_u$ where the u -norm corresponds to an Euclidian norm straighten according to the distribution of the exogenous measurements stored in the historic database (Eq. 1).

$$d(\mathbf{u}, \mathbf{u}^*)^2 = (\mathbf{u}^* - \mathbf{u})' \Sigma_u^{-1} (\mathbf{u}^* - \mathbf{u}) \quad (1)$$

[†] High precision corresponds to low local variance.

[‡] This database will be automatically updated; it is not a temporal buffer but a selection of interesting templates.

Where $\Sigma_{\mathbf{u}} = \mathbf{cov}(\mathbf{U})$ is the correlation matrix of the stored exogenous context observations[§]. This distance is coded easily without computation of the correlation matrix and its inverse as shown in the following algorithm:

Let $\mathbf{U} = (\mathbf{u}_1 \dots \mathbf{u}_n)'$ be the matrix of all exogenous measurements stored in the database and $\boldsymbol{\mu}$ be its mean.

Compute $\mathbf{QR} = \mathbf{U} - \boldsymbol{\mu}$ the unary and upper triangular decomposition of the centered observations.

Compute $\mathbf{r} = (\mathbf{u}^* - \mathbf{U})\mathbf{R}^{-1}$ which will be a rectangular matrix with n rows (the number of stored observations) and $m_{\mathbf{u}}$ columns (if $m_{\mathbf{u}}$ is the dimension of the exogenous vector). As \mathbf{R} is a triangular matrix the computation of \mathbf{r} is straightforward.

Finally make \mathbf{d} the vector of n rows by computing for each row the Euclidian norm of the corresponding $m_{\mathbf{u}}$ -dimension vector in \mathbf{r} : $d^2 = \langle \mathbf{r}, \mathbf{r} \rangle$.

$d_i^2(\mathbf{u}^*) = \|\mathbf{u}^* - \mathbf{u}_i\|^2 = \sum_{j=1 \dots m_{\mathbf{u}}} r_{i,j}^2$ for $i=1 \dots n$.

The proximity value is defined as a given quantile of the computed distances $d_i(\mathbf{u}^*)$. Eq. 2 defines $prx(\mathbf{u}^*, \mathbf{H}_{\mathbf{u}})$ as the proximity of \mathbf{u}^* to the history $\mathbf{H}_{\mathbf{u}}$ with percentile parameter $\rho_{\mathbf{u}}$.

$$P(d(\mathbf{u}, \mathbf{u}^*) \leq prx(\mathbf{u}^*)) < \rho_{\mathbf{u}} \quad (2)$$

As the number of observation is finite, this quantile is just an approximation. For example we may select the first value of the “sorted” distances $d'_i < d'_{i+1}$ which realizes the preceding constraint: $i/n \leq \rho_{\mathbf{u}}$ and $(i+1)/n > \rho_{\mathbf{u}}$.

2.4 Definition of the adequacy

The adequacy is an indicator of novelty according to the exogenous observations. It should increase when new observations are already seen, or equivalently if new observations are common to the observations stored in our database.

We keep a buffer of the last observed data $\mathbf{B}_{\mathbf{u}}$. For each observation \mathbf{u}^* in this buffer its proximity to the history $\mathbf{H}_{\mathbf{u}}$ is $prx(\mathbf{u}^*)$. This list of proximities should be compared to distances accepted in the history. We also have an equivalent list of proximities $prx(\mathbf{u})$ for all \mathbf{u} in $\mathbf{H}_{\mathbf{u}}$ but each computed with all observations in $\mathbf{H}_{\mathbf{u}}$ excepted the singleton $\{\mathbf{u}\}$. The proximity value has the dimension of a distance so the sum of all squared

[§] Along this article, a Gaussian approximation is made following an assumption that things may behave “normally” in a local context. However this is just an approximation so computed proportions are not exact observed probabilities, but the dimension of each object is respected.

proximities follows a statistic law equivalent to a χ^2 . The ratio f (Eq. 3) of the two corresponding sums (local distances over normal distances) approximately follows a Fisher law $F(\#\mathbf{B}_{\mathbf{u}}, \#\mathbf{H}_{\mathbf{u}})$ (where $\#\mathbf{B}$ denotes the cardinal of the set \mathbf{B}).

$$f = \frac{\frac{1}{\#\mathbf{B}_{\mathbf{u}}} \sum_{\mathbf{u}^* \in \mathbf{B}_{\mathbf{u}}} prx^2(\mathbf{u}^*)}{\frac{1}{\#\mathbf{H}_{\mathbf{u}}} \sum_{\mathbf{u} \in \mathbf{H}_{\mathbf{u}}} prx^2(\mathbf{u})} \quad (3)$$

The numerator is high if the new observations are far from the stored historic data. The way it is “far”, is normalized by a standard measure of proximity done of normal observation (the denominator).

An adequacy value may be defined as the p-value associated to this statistic test (Eq. 4):

$$adequacy = 1 - P(F < f) \quad (4)$$

2.5 Risk of abnormality

Each time a new observation is available, and if the current adequacy is high, the endogenous measurements should be analyzed. We want to compare this new observation with the ones already observed when the context was similar.

We extract a subset of the stored input database with observations close to the current context. This is inferred from the computation of the proximity values. Our subset is the set of historic observations corresponding to a proximity percentile ρ_x (Eq. 5^{**}). We denote \mathbf{H} the stored set of all historic observations including exogenous and endogenous indicators ($\mathbf{H}_{\mathbf{u}}$ and \mathbf{H}_x are the respective projections of \mathbf{H} on the exogenous and endogenous indicators):

$$\mathbf{H}(\mathbf{u}^*) = \left\{ (\mathbf{u}_i, \mathbf{x}_i) \in \mathbf{H} / \frac{1}{\#\mathbf{H}-1} \sum_{j \neq i} \mathbf{1}_{d_j(\mathbf{u}^*) \leq d_i(\mathbf{u}^*)} \leq \rho_x \right\} \quad (5)$$

This subset of the historic storage contains couples of exogenous and endogenous indicators, but it is defined only from computations on context (exogenous) observations.

The score is computed from the likelihood of the endogenous observations according to a local Gaussian law defined empirically by the selected subset of endogenous historic data. If \mathbf{x}^* is the current endogenous observation, for each $\mathbf{x} \in \mathbf{H}_x(\mathbf{u}^*)$, we compute $d(\mathbf{x}^*, \mathbf{x}) = \|\mathbf{x}^* - \mathbf{x}\|_x$ as we did previously on exogenous observations:

^{**} The bold $\mathbf{1}$ here (Eq. 5, 10 and 12) denotes the indicator function.

$$d(\mathbf{x}^*, \mathbf{x})^2 = (\mathbf{x}^* - \mathbf{x})' \Sigma_x(\mathbf{u}^*)^{-1} (\mathbf{x}^* - \mathbf{x}) \quad (6)$$

This time $\Sigma_x(\mathbf{u}^*)$ refers to the covariance matrix of the selected endogenous observations in $\mathbf{H}_x(\mathbf{u}^*)$. The (\mathbf{u}^*) notation is only there to recall that the subset $\mathbf{H}_x(\mathbf{u}^*)$ contains only endogenous measurement that were chosen with approximately similar context.

Finally, for each couple $(\mathbf{u}_i, \mathbf{x}_i)$ in $\mathbf{H}(\mathbf{u}^*)$ we have a distance measure $d_i(\mathbf{x}^*) = d(\mathbf{x}^*, \mathbf{x}_i)$ on exogenous data (Eq. 6) and an equivalent proximity distance $d_i(\mathbf{u}^*)$ on endogenous data (Eq. 1). To take into account the proximity in the score computation we weight the endogenous distances by the context proximity^{††}:

$$score(\mathbf{u}^*, \mathbf{x}^*) = \frac{\sum_{(\mathbf{u}_i, \mathbf{x}_i) \in \mathbf{H}(\mathbf{u}^*)} \frac{d_i^2(\mathbf{x}^*)}{1 + d_i^2(\mathbf{u}^*)}}{\sum_{(\mathbf{u}_i, \mathbf{x}_i) \in \mathbf{H}(\mathbf{u}^*)} \frac{1}{1 + d_i^2(\mathbf{u}^*)}} \quad (7)$$

As this measure has the dimension of a χ^2 with m_x degrees of freedom (m_x is the dimension of the endogenous vector) the anomaly risk indicator is defined by

$$risk = P(\chi_{m_x}^2 \leq score) \quad (8)$$

2.6 Risk precision

The preceding computation gives a risk of abnormality, but this essentially depends on the observations already observed. Hence it is necessary to follow another quality indicator that gives a precision for this result. Our choice is to use an indicator based on the local variance of the endogenous data according to the current context.

Let σ_j be the variance of one of the endogenous observations x_j on \mathbf{H} and the equivalent $\sigma_j(\mathbf{u}^*)$ on $\mathbf{H}(\mathbf{u}^*)$. The ratio of those two variances is a Fisher, so we take the corresponding p-value and the mean on all components j . This is given by (Eq. 9) where generic variable F is a Fisher stochastic variable with the adequate number of freedom degrees $\#\mathbf{H}(\mathbf{u}^*)$ and $\#\mathbf{H}$:

$$p_j(\mathbf{u}^*) = P\left(F < \frac{\sigma_j^2(\mathbf{u}^*)}{\sigma_j^2}\right) \text{ for all endogenous variable } x_j \quad (9)$$

$$\text{and } precision = 1 - \frac{1}{m_x} \sum_{j=1}^{m_x} p_j(\mathbf{u}^*)$$

^{††} This is one proposition, other weighted computation are possible for Eq. 7.

This precision value is a number between 0 and 1 that increases when the variance on the local context decreases^{††}.

2.7 Update of the database

Each time a new observation $(\mathbf{u}^*, \mathbf{x}^*)$ is acquired one first computes the adequacy, the risk and its precision; but the database should also be updated. The observations stored in this database represent the normal behavior of the monitored system. As this set should maintain a small size, one must focus only on the “best” observations and store them as templates.

Hence a test is made to check if the new observation is more relevant than the worse one already stored. If this is the case the new observation replaces the other.

A “worse” observation is selected as the one that is the least useful for our purpose. That’s an observation which may be suppressed from the history without “much” loss in the evaluation of the proximities and risks^{§§}. At first, this observation is selected among the ones (set \mathbf{H}^-) with the lower values of proximity. A new percentile ρ^- is defined for this purpose (Eq. 10):

$$\mathbf{H}^- = \left\{ (\mathbf{u}_i, \mathbf{x}_i) \in \mathbf{H} / \frac{1}{(\#\mathbf{H}-1)} \sum_{j \neq i} \mathbf{1}_{prx(\mathbf{u}_j) \leq prx(\mathbf{u}_i)} \leq \rho^- \right\} \quad (10)$$

Then the observation with the lowest risk is selected.

$$(\mathbf{u}^-, \mathbf{x}^-) = \arg \min_{(\mathbf{u}, \mathbf{x}) \in \mathbf{H}^-} [risk(\mathbf{u}, \mathbf{x})] \quad (11)$$

This observation is replaced by the current one $(\mathbf{u}^*, \mathbf{x}^*)$ if the current context is still unknown (belong to the set \mathbf{H}^+ defined by Eq. 12) and if the current score is greater than the “worse” one. The first constraint limits the number of templates belonging to the same context. The second condition ensures that the new added observation corresponds to something really different.

A fourth percentile threshold ρ^+ is used for the context constraint:

$$\mathbf{H}^+ = \left\{ (\mathbf{u}_i, \mathbf{x}_i) \in \mathbf{H} / \frac{1}{(\#\mathbf{H}-1)} \sum_{j \neq i} \mathbf{1}_{prx(\mathbf{u}_j) \leq prx(\mathbf{u}_i)} > 1 - \rho^+ \right\} \quad (12)$$

^{††} We certainly may find a better multivariate solution here. It should take endogenous correlation into account.

^{§§} As a first implementation of this algorithm, a stochastic metropolis algorithm was programmed to update the database. It adds more freedom and converges to better solutions but the random process makes it difficult to validate on a real time implementation. It was then decided to temporarily replace the stochastic method by a least precise but more easily controllable deterministic rule.

Then the new observation (\mathbf{u}^* , \mathbf{x}^*) replaces the “worse” one if it belong to \mathbf{H}^+ and if $score(\mathbf{u}^*, \mathbf{x}^*) > score(\mathbf{u}^-, \mathbf{x}^-)$.

At the beginning, the database is initialized with the first observations. Intermediates states of the history \mathbf{H} may be stored to help the engineers understand the behavior of the detector and optimize the configuration thresholds.

3. FIELDDED IMPLEMENTATION

There are two reasons for the deployment of HM algorithms in the test benches:

To monitor the installation and the tested machine: in spite of automated and human monitoring of safety parameters, a slow degradation of a body of the machine or the bench cell (engine, gearbox, torque transmissions ...) may lead to a sudden and unexpected failure.

The economic impact may therefore prove prohibitive regarding the program developments underway. Indeed, apart from the exorbitant repair cost, time penalizes programs. It is therefore essential to anticipate such events by deploying a system that allows, not to replace what already exists in terms of real-time monitoring, but defensively to detect any abnormality, known or not, which may lead to a destructive event.

Maturation of algorithms: by definition, the HM algorithms must evolve continuously. Indeed, even if their developments for embedded applications require a high TRL, the unexpected, related to new applications or exceptional operating conditions ever encountered, requires them to evolve in light of this new experience. Similarly, these same algorithms deployed on ground applications need to be matured. This maturation in the test cells is beneficial both for the monitoring of these facilities themselves and the embedded systems.

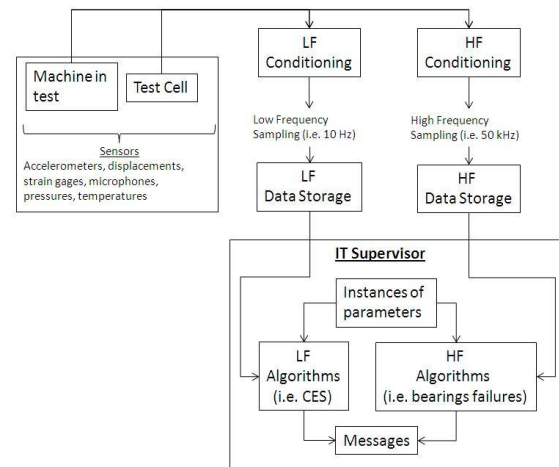


Figure 1: Example of a Health Monitoring system deployed in a test bench.

We distinguish the machine from the facility bench because it may be subject to special supervision regardless of the tested machine. The sensors are installed on the machine and on the equipments of the bench. Depending on their type, signals they deliver are digitalized and stored in databases at:

- High frequency for dynamic measurements issued by the accelerometers, displacement sensors, strain gauges, microphones and unsteady pressure sensors.
- Low frequency for temperature, static pressures, rpm measurements and dynamic signals processed. In the example using the CES algorithm, these types of parameters are exploited.

In the following sample application we focus on the health of a shaft for transmitting torque from low frequency settings. (CES algorithm, left branch of Figure 1). Other implementations of the LF algorithm are also under investigation (Lacaille 2010b). For HF implementations see (Klein 2009) and (Hazan 2010a and 2010b).

3.1 Current benchmark

One accelerometer and one thermocouple are mounted on the ball bearing which bears the shaft to be monitored. One room microphone is located in the test cell near the bench equipment. Sensors measure the radial displacement of the shaft relative to a fixed structure.

The context parameters which influence the above parameters are:

- The shaft speed, calculated from the signal of a phonic wheel linked to the rotating shaft.

- The position of an air intake valve which directly affects the torque felt by the shaft.
- The torque measurement itself.
- The pressure of a piston chamber which loads axially the ball bearing of the shaft and hence affects its dynamic response.

Conditioning systems of sensors and storage bases can be either specific or common to several benches.

A supervisor computer, in turn:

- Hosts and executes the algorithms from a dedicated development and maturation platform (Lacaille 2009b, 2010b).
- Receives and manages the high and low frequency data.
- Generates alert messages.

3.2 Implementation

The algorithm consists of standard modules, all customizable and available from a tools library of the development and maturation studio (see implementation graph on Figure 2).

- The “Read File” module reads the low frequency data files as they arrive in a file directory managed by the supervisor.
- In this example, the “Average” module splits each original signal into indicators using moving averages. This has the effect of smoothing the original signal.
- The “Instances” blocks contain the heart of the CES algorithm and are run in parallel.
- The “Demultiplexing” module separates endogenous and exogenous parameters.
- The CES module delivers the abnormality *risk* and a quality indicator computed as the geometric mean of the previously computed *adequacy* and *precision*.
- The “Message” module issues anomalies whenever the adequacy is above a given threshold (confidence regarding the current situation) and when the risk exceeds another threshold. (A third threshold limits the number of detections waiting the risk to down-cross its value before launching a new alert.)

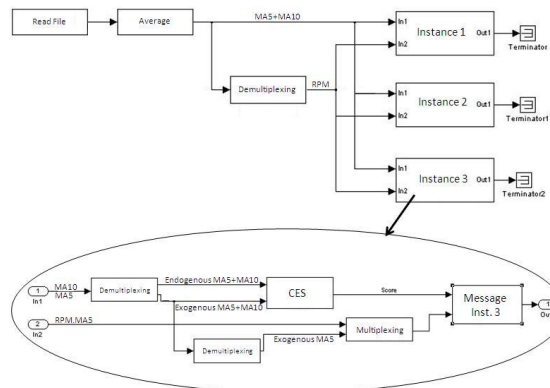


Figure 2: Implementation of the CES algorithm.

3.3 Algorithm’s instances

In the implementation, each instance of parameters is dedicated to a special supervision such as:

Change of the dynamic behavior of the transmission shaft: in this case (Figure 3), the endogenous parameters are by-revolution tracked signals vibrations levels in low frequency computed from the bearing’s accelerometer and sensors that directly measure the radial displacement of the shaft. These levels are representative of the dynamic response of the shaft according to its imbalance, which can itself be a consequence of geometric imperfections or thermal expansion.

Exogenous Parameters from LF storage system		
RPM	Shaft speed	rpm
VPOS	Valve position	degrees
TORQUE	Mechanical torque in the shaft	N.m
PISTPRES	Pressure in a piston chamber	Pa
Endogenous Parameters from LF storage system		
BRGACATRK	accelerometer A, 1/Rev tracked level	ips pk
BRGCBTRK	accelerometer B, 1/Rev tracked level	ips pk
SHTDPATRK	shaft displacement A, 1/Rev tracked level	mils da
SHTDPBTRK	shaft displacement B, 1/Rev tracked level	mils da

Figure 3: Instance configuration for change detection in the dynamic behavior of the transmission shaft.

Change of the mechanical behavior of the ball bearing: (Figure 4) the endogenous parameters are low frequency signals coming from the bearing temperature and energy levels by frequency bands. The energy levels are computed from the wideband signals of the accelerometer and the room microphones. For this instance setting, assumption is made that whether a gradual bearing spalling occurs, levels of endogenous parameters will increase (energy levels in specific frequency bands).

Exogenous Parameters from LF storage system		
RPM	Shaft speed	rpm
VPOS	Valve position	degrees
TORQUE	Mechanical torque in the shaft	N.m
Endogenous Parameters from LF storage system		
BRGTMP	Bearing temperature	° C
BRGACARMSFB1	accelerometer A, RMS level in frequency band 1	g rms
BRGACARMSFB2	accelerometer A, RMS level in frequency band 2	g rms
BRGACARMSFB3	accelerometer A, RMS level in frequency band 3	g rms
BRGMCRMSFB1	microphone RMS level in frequency band 1	dB
BRGACRMSFB2	microphone RMS level in frequency band 2	dB
BRGACRMSFB3	microphone RMS level in frequency band 3	dB

Figure 4: Instance configuration for change detection in the mechanical behavior of the ball bearing.

For these two instances, the context is given by the shaft speed, the position of an air intake valve and the torque transiting through a transmission shaft between the output of a multiplier box and the machine under test.

It is worth noting an additional context parameter for the instance in Figure 3 compared to Figure 4: the piston pressure of the axial loading for the ball bearing which has a direct influence on the tracked levels coming from the accelerometer and the displacement sensors.

The amount of endogenous parameters is not exhaustive and is linked to the amount of sensors that can be much greater than what is presented for those two instances.

(A third instance shown on Figure 2 is not developed here.)

3.4 Experimental observations

Early in the test campaign, when the algorithm starts from zero, the adequacy was never up and fluctuates as the context data have never been encountered. However, when the adequacy exceeds a preset threshold, it stays high. At this point when the risk reached the detection threshold, a message is generated (the star on Figure 5).

The algorithm finished its calibration when no new templates are added to the database. The adequacy keeps the maximum value and almost all observed detections correspond to acknowledged anomalies.

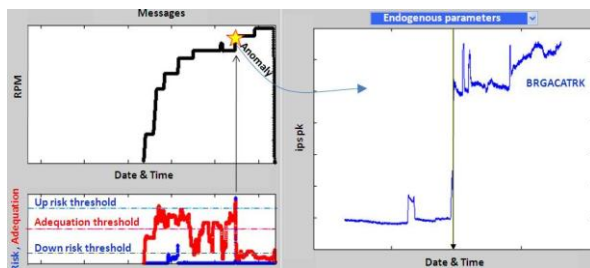


Figure 5: Results are displayed on screen and may be interpreted.

4. CONCLUSION

In this article we described a light algorithm able to detect unusual behavior of a system made from an engine and/or a bench cell. The original point in this code is the management of the input data as a couple of sensors subsets dedicated to the context identification and the monitoring itself. We may also quote the way the detection is controlled by different quality indicators taking into account both the context identification and the precision of the estimation. This algorithm was first installed on a bench for maturation of the code but also to monitor the bench.

An offline test was build for statistic computation of the main performances indicators for such detection algorithm. This test was build from data recorded during a machine test bench of 15 days. We registered all real problems observed during the test (such as stall) and we add synthetic degradations based on expert knowledge. The process was repeated 28 times with random positioning of the simulated degradation. A cross validation scheme was applied: it gives a false alarm rate of less that 1% (with a precision of ± 4%) and a detection probability of more than 55% (± 20%).

The definitions of the KPI are given below:

$$\begin{aligned}
 PFA &= \frac{P(Healthy/Detected)}{P(Detected)} \\
 &= \frac{P(Detected/Healthy)P(Healthy)}{P(Detected)} \quad (13)
 \end{aligned}$$

which gives

$$PFA = \frac{\alpha(1-P(Faulty))}{\alpha(1-P(Faulty)) + (1-\beta)P(Faulty)} \quad (14)$$

and

$$POD = P(Detected/Faulty) = 1 - \beta \quad (15)$$

where α is the type I error and β the detection test power (Lacaille 2010a).

The PFA result corresponds to the requirements for such algorithm on bench test application, however the POD is a little low but is greatly improved by the fusion layer and an optimization of the threshold parameter is in progress.

NOMENCLATURE

ACMS	Aircraft Condition Monitoring System
AQV	Adequacy Quality Value
CES	Continuous Empirical Score
DB	Database
FADEC	Fault Authority Digital Engine Control
FMECA	Failure Mode, Effects, and Criticality Analysis
HM	Health Monitoring
KPI	Key Performance Indicator
LASSO	Least Absolute Shrinkage and Selection Operator
PFA	Probability of False Alarm
POD	Probability of Detection
TRL	Technical Readiness Level

NOTATIONS

\mathbf{u}	Vector of context (exogenous) indicators
\mathbf{x}	Vector of endogenous indicators
m_u, m_x	Dimensions of exogenous and endogenous vectors
\mathbf{H}	History storage database
$(\mathbf{H}_u, \mathbf{H}_x)$	DB projection on respectively exogenous and endogenous indicators
$(\mathbf{u}^*, \mathbf{x}^*)$	Current observation
$(\mathbf{u}^-, \mathbf{x}^-)$	“Worse” observation in the DB (the least useful observation)
$\mathbf{H}(\mathbf{u}^*)$	Observations in the neighbor of the current context
prx	A quantile distance to the current history DB
<i>adequacy</i>	Confidence to be already observed
<i>risk</i>	Probability of abnormality
<i>precision</i>	Reliability of the risk value
$d_t(\mathbf{u}), d_t(\mathbf{x})$	Component of the proximity (distance to one observation of the DB) in exogenous or endogenous projection
ρ_u	Percentile threshold for the definition of the proximity value
ρ_x	Percentile threshold for the definition of the context neighborhood
ρ^-	Percentile threshold for the selection of the “worse” stored observation in update process of the DB
ρ^+	Percentile threshold for context replacement constraint in update process

REFERENCES

- (Blanchard et al., 2009) S. Blanchard, J. Lacaille, and M. Cottrell. *Health monitoring des moteurs d'avions*. In Les entretiens de Toulouse, France, 2009.
- (Cômes and all, 2010a) E. Cômes, M. Cottrell, M. Verleysen, and J. Lacaille. *Aircraft engine health monitoring using self-organizing maps*. In ICDM, Berlin, Germany, 2010.
- (Cômes and all, 2010b) E. Cômes, M. Cottrell, M. Verleysen, and J. Lacaille. *Self organizing star (sos) for health monitoring*. In ESANN, Bruges, 2010.
- (Cômes and all, 2011) E. Cômes, M. Cottrell, M. Verleysen, and J. Lacaille. *Aircraft engine fleet monitoring using Self-Organizing Maps and Edit Distance*. In WSOM 2011, Espoo, Finland.
- (Cottrell and all, 2009) M. Cottrell and all. *Fault prediction in aircraft engines using self-organizing maps*. In WSOM, Miami, FL, 2009.
- (Flandrois and Lacaille, 2009) X. Flandrois and J. Lacaille. *Expertise transfer and automatic failure classification for the engine start capability system*. In AIAA Infotech, Seattle, WA, 2009.
- (Hazan, 2010a) A. Hazan, M. Verleysen, M. Cottrell, and J. Lacaille. *Trajectory clustering for vibration detection in aircraft engines*. In ICDM, Berlin, Germany, 2010.
- (Hazan, 2010b) A. Hazan, M. Verleysen, M. Cottrell, J. Lacaille, *Linear smoothing of FRF for aircraft engine vibration monitoring*, ISMA 2010, Louvain.
- (Klein, 2009) R. Klein. *Model based approach for identification of gears and bearings failure modes*. In PHM Conference, San Diego, CA, 2009.
- (Lacaille, 2007) J. Lacaille. *How to automatically build meaningful indicators from raw data*. In AEC/APC, Palm Spring, CA, 2007.
- (Lacaille and Nya Djiki, 2009) J. Lacaille and R. Nya Djiki. *Model based actuator control loop fault detection*. In Euroturbo Conference, Graz, Austria, 2009.
- (Lacaille, 2009a) J. Lacaille. *An automatic sensor fault detection and correction algorithm*. In AIAA ATIO, Hilton Beach, SC, 2009.
- (Lacaille, 2009b) J. Lacaille. *A maturation environment to develop and manage health monitoring algorithms*. In PHM, San Diego, CA, 2009.
- (Lacaille, 2009c) J. Lacaille. *Standardized failure signature for a turbofan engine*. In IEEE Aerospace Conference, Big Sky, MT, 2009.
- (Lacaille, 2010a) J. Lacaille. *Validation of health-monitoring algorithms for civil aircraft engines*. In IEEE Aerospace Conference, Big Sky, MT, 2010.
- (Lacaille 2010b) J. Lacaille, V. Gerez, R. Zouari, *An*

Adaptive Anomaly Detector used in Turbofan Test Cells, PHM 2010, Portland, OR.

(Lacaille, 2011a) J. Lacaille, E. Côme, *Visual Mining and Statistics for a Turbofan Engine Fleet*, in Proceedings of IEEE Aerospace Conference 2011, Big Sky, MT.

(Lacaille 2011b) J. Lacaille, E. Côme, *Sudden change detection in turbofan engine behavior*. In proceedings of the 8th International Conference on Condition Monitoring and Machinery Failure Prevention Technologies, Cardiff, UK.

(Seichepine, 2011) N. Seichepine, J. Ricordeau, J. Lacaille, *Datamining of flight measurements*. In proceedings of AIAA@Infotech 2011, Saint Louis, MO.



Jérôme Lacaille is senior expert in algorithms for Snecma. He joined the company in 2007 with responsibility for developing a health monitoring solution for jet engines. Jérôme has a PhD from the Ecole Normale Supérieure, France in Mathematics. Jérôme has held several positions including scientific consultant and professor. He has also co-founded the Miriad Technologies Company, entered the semiconductor business taking in charge the direction of the Innovation Department for Si Automation (Montpellier - France) and PDF Solutions (San Jose - CA). He developed specific mathematic algorithms that were integrated in industrial process. Over the course of his work, Jérôme has published several papers on integrating data analysis into industry infrastructure, including neural methodologies and stochastic modeling.



Valerio Gerez is a mechanical engineer who works for Snecma since 1982. He has almost 30 years of experience in aircraft engines in the areas of quality and especially in Engine dynamics, both in test cells and in Aircrafts. In 2006, he joined the Diagnostic and Prognostic Department and now manages R&D HM projects for Snecma future applications and the deployment of algorithms in test cells.

Proficy Advanced Analytics: a Case Study for Real World PHM Application in Energy

Subrat Nanda¹, Xiaohui Hu²

¹GE Global Research, 1, Research Circle, K15A61, Niskayuna NY, 12309.

subrat.nanda@ge.com

²GE Intelligent Platforms, 325 Foxboro Blvd, Foxboro, MA, 02035

xiaohui.hu@ge.com

ABSTRACT

GE monitors a large number of heavy duty equipment for energy generation, locomotives and aviation. These monitoring and diagnostic centers located world-wide sense, derive, transmit, analyze and view terabytes of sensory and calculated data each year. This is used to arrive at critical decisions pertaining to equipment life management - like useful life estimation, inventory planning and finally assuring a minimum level of performance to GE customers. Although a large number of analytical tools exist in today's market, however there is a need to have a tool at disposal which can aid not just in the analytical algorithms and data processing but also a platform for fleet wide deployment, monitoring and online processing of equipment. We describe a Prognostics & Health Management (PHM) application for GE Energy which was implemented using GE Intelligent Platform products, and explore some capabilities of both the application and the analytics tool.

1. INTRODUCTION

GE monitors a very large number of heavy duty equipment for energy generation, locomotives and aviation. The main purpose of this monitoring analysis is to analyze the usage and condition of these equipment components, and to assist the users at Monitoring & Diagnostics (M&D) center to perform proactive maintenance activities, root causing existing problems and assist in planning for future downtime periods. This information can be used to help them to plan parts inventory and logistics, thus ensuring a higher reliability and availability for GE customers.

Various systems and sub-systems require the use of advanced data driven techniques to integrate the large amount of field data captured with the existing empirical and physics based models Jammu, Vinay(2010), et al.,

Nanda, Hu This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

These data driven methods are also used to assist the engineers to unearth hidden relations and patterns in key parameters of interest Vachtsevanos (2006), et al. We present in this paper a platform offering from GE Intelligent Platforms called Proficy Cause+ and Troubleshooter™ (GE Intelligent Platforms), (CSense Systems (Pty) Ltd) and discuss some of the key lessons learnt by applying it to monitor and develop prognostics & health management tools for GE equipment.

This paper is organized as follows: we provide a brief outline of the PHM schema developed and flow of information in Section 2. We then discuss the different sources of captured or derived information and patterns searched and the data pre-processing philosophy that are used by GE M&D in order to develop a robust and reliable PHM system. We also discuss in brief an anomaly detection algorithm to highlight one of the many anomaly detection methods used to monitor critical equipment and related alarm generation. We next introduce GE Intelligent Platforms Proficy platform in Section 3 and describe how it enables one to perform offline analysis, integrate with existing PHM platforms and perform field implementation. Section 4 contains some preliminary methods for anomaly detection that were used for a PHM implementation case study. Finally Section 5 has some pointers to what we are planning to do in near future and conclusions based on this study. The information used in this case study consists of data from GE Energy equipment that was anonymized and scaled to avoid disclosure of proprietary information. This does not in any way affect the validity of the methodology or implementation as described in the paper.

2. PROBLEM DEFINITION

In order to enable GE's Monitoring and Diagnostic Centre to detect insipient anomalies and failures in energy generation turbines and subsequently take corrective action, automated monitoring of the parameters of interest such as the following: operating temperatures, pressure ratios,

power produced, measured variables compared to their set point values, performance levels relative to empirically derived operating profiles, ambient conditions, etc. Hence it is required to automate the process of detecting anomalies in performance levels and pick up sudden shifts, and automatically provide possible root causes for observed anomalies. Further, if future performance levels can be predicted accurately, it would enable field engineers to suggest suitable control settings to power plant personal.

The PHM system developed has the following schema and flow of information (Figure 1)

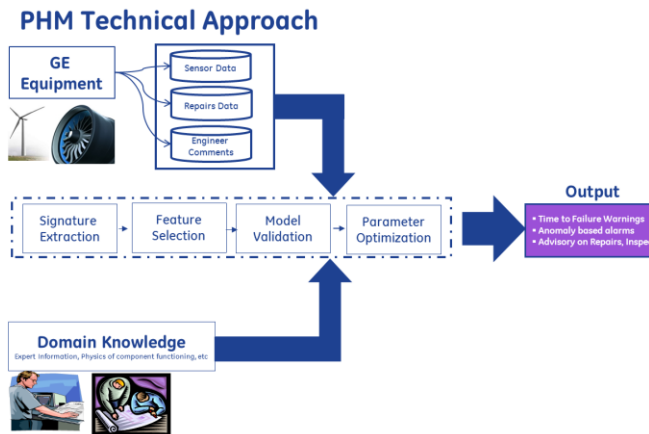


Figure 1: Schema of PHM system developed

In the following subsections, we outline some of the required components for a practical PHM system deployment. This includes capturing multiple data sources, pre-processing and anomaly detection methods. We describe some of these details in the following sections.

2.1 Major Sources of Information captured for PHM

The data captured for remote monitoring & control is of varied types and can be broadly classified into following types:

- Historical operational data:** various sensor signals are stored in both central and distributed data bases. These contain equipment performance and component wise data at various time intervals. Different PHM applications may require customized data sampling rates depending upon the specific failure modes in the components for which PHM applications are developed.
- Controller Data:** Onboard and centrally located controllers use and generate multiple logical values for accurate controlling process. These calculations are stored and are used for developing predictive or diagnostic methods.
- Engineer observations:** Various free-form and structured textual information are recorded by GE field

engineers and M&D personal when responding to customer calls.

- Repair & Maintenance Data:** Detailed descriptions relating to previous repairs and maintenance procedures and inspections performed on equipment are stored in various data repositories.

2.2 Types of PHM Signals & Patterns Monitored

A typical PHM system can be used to perform real time or offline processing to provide accurate equipment remaining life estimate thus enabling subsequent decisions to be taken. This requires multiple types of derived values, some of which are mentioned below and are monitored on a continual basis by the onsite controller or central analysis modules:

- Statistical Quantities:** Various statistical measures like higher order moments of key parameters, moving statistical calculations, etc. See (Casella, G & Berger R. L, 1990) for more details.
- Evolving physical quantities:** In addition to static measures or feature calculations, time evolving nature of the major parameters are critical to detect failures.
- Deviation from expected values:** Most engineering parameters have pre-defined set values and are tracked for deviations from their set point values. A significant deviation and the direction of deviation is an indicator for certain failure modes and insipient failures in critical components.
- Model residuals:** Increasing residuals between empirically derived models and observed values can give insights into impending failures and isolation using appropriate classification models.

2.3 Data Preprocessing for PHM Modeling

Prior to any subsequent PHM models being developed, raw data captured typically has to be processed to ensure that proper variations are captured, no biases are introduced and the underlying distributions generating real life data are modeled correctly. The following are some of the processes that one might consider during a PHM application:

- Filtering:** raw sensor data is filtered on multiple dimensions to include relevant time periods, operating modes of equipment, ambient conditions and failure modes.
- Frame Specific Segmentation:** sensor and controller data are sampled based on specific frame types to average out against biases arising from multiple designs and operating ranges of key parameters.

- c. Smoothing: protecting against biases in model development and parameter learning. Also, to prevent faulty rules due to outliers arising out of data quality issues, certain statistical or model based smoothening modules have been implemented.
- d. Data Quality: In addition to the above mentioned data preprocessing, it is also required to resolve data quality issues such as missing data, faulty sensor readings and out of range values.

Due to the nature of multiple key health indicators that are monitored, there was a need felt to implement different anomaly detection algorithms that can capture different failure modes. Different types of data driven methodologies have been developed in order to develop a robust PHM system – including soft computing (Bonissone, P., & Kai Goebel), reliability and remaining useful life estimation (see Ebeling (2005), machine learning methods and the fusion of these methods with physics concepts. In the following section we outline one such anomaly detection method that was implemented for our PHM case study. These anomaly detection methods would isolate abrupt changes in operation patterns, which are then used for subsequent analysis and decision making process.

2.4 Anomaly Detection Algorithm

Both online and batch mode implementations have been tested for anomaly detection, and an example algorithm is described below. Interested user can see following references for advanced analytics methods used: Kumar, Vipin et al(2005), (Russel S & Norvig P, 2002); (Duda R.O, Hart P.E & Stork, D.G 2001)

Multi-variate Hypothesis testing method Hotelling's T-Square was proposed by Harold Hotelling (Hu, Xiao, Qui, Hai and Iyer, Naresh, 2007), (M. Markou & S. Singh, 2007). It is a multivariate technique that captures the changes in data from multiple parameters by using their covariance information. This is a generalization of the Student's t statistic used in multiple hypothesis testing.

Consider a time series as:

$$X(t) = (X_1(t), X_1(t), \dots, X_m(t))^T$$

Where m is the number of variables sampled at each time t , $X_i(t)$ are the parameter sampled at regular intervals. Assuming $X(t)$ is a multi-variate normal $\sim N(\mu, \Sigma)$, where μ and Σ are the multivariate mean and co-variance matrix respectively.

The mean and variance μ and Σ are estimated as follows:

$$\bar{X} = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_m)^T \quad (1)$$

$$W = \frac{1}{n-1} \sum_{i=1}^n (X(t) - \bar{X})(X(t) - \bar{X})^T \quad (2)$$

Hotelling T2 Statistic for $X(t)$:

$$T^2 = (X(t) - \bar{X})W^{-1}(X(t) - \bar{X}) \quad (3)$$

The Null Hypothesis states that $X(t)$ is NOT different from previous n samples or that no change occurred. Large values of T-Square statistic imply that the null hypothesis was not true. This was implemented using thresholds in Matlab™ (Mathworks). The test comprised of testing the Hotelling-T Square Statistic if it exceeded a present threshold, thereby confirming a change/anomaly at given time instant.

This statistic uses the statistical distance and incorporates multivariate variance-covariance matrix, to detect significant shifts and linear relationships.

We optimized the various parameter settings and threshold values for the anomaly detection module by analyzing different turbine frame types. This was then run for each of the monitored equipment health indicator and each time it was run, we fused the outputs from multivariate Hotelling-T square with other anomaly detection algorithms. This was done in order to capture both local anomalies (uni-variate sense) and system level anomalies (multivariate sense) across more than a single monitored parameter at same time instant.

2.5 Alarm Generation Process

Due to a large number of equipment that are being monitored and their key parameters, there is a definite need to keep the fleet wide alarm rates to have a high probability of detection with a very low false alarm rate. For this purpose, we implemented a time based alarming process and optimized it with respect to the field data, observed failure rates and deployed the algorithms with an ability to change the alarming settings based on fleet requirements.

3. PROFICY ADVANCED ANALYTICS TOOLSET

Proficy® is a suite of commercial-off-the-shelf software solutions that are designed to help solve the operations challenges in infrastructure and/or manufacturing industries. Proficy software suite offers the depth and breadth of capabilities from control and optimization.

Proficy software Suite provides a whole solution from data collection, data management, data visualization to data analysis for remote monitoring and diagnosis.

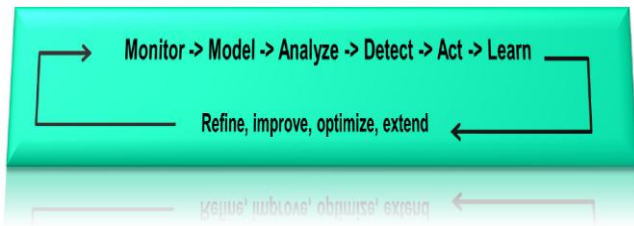


Figure 2. Remote Monitoring & Diagnosis Process

Proficy troubleshooter/cause+ (Figure 3) suite is the key software to perform advanced data analysis and knowledge discovery. It provides a platform to facilitate all steps of advanced analytics

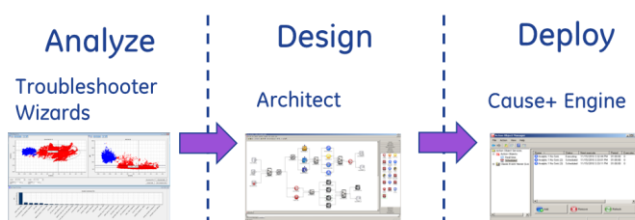


Figure 3. Troubleshooter/Cause+ Functionalities

Analyze: troubleshooter wizards provide rich analytics models and powerful visualization tools for subject matter experts to speed up the process of data exploration and knowledge discovery.

Design: Troubleshooter architect provides integrated development and simulation environment for application engineers to design and debug analytic solutions with pre-built database interface, various data preprocessing/post-processing techniques, and internal or external algorithms/analytics models

Deploy: Cause+ provides a deployment environment for operation managers to run the analytics solutions in real-time, event-triggered, or scheduled manners.

3.1 Offline Analysis

The Troubleshooter Wizards (continuous or discrete) guide users through troubleshooting processes in the developer environment. Using the Wizards, various tools are available to identify the causes of process deviation using historical data. Preparing data is made quick and easy using graph and trend views, and modeling the industrial process is intuitive. From there, knowledge about the process can be gleaned effectively and combined with the knowledge of expert personnel to develop an integrated solution to process problems.

This solution can then be further customized and tested within Architect, and deployed in real-time in the cause+ engine by Action Object Manager.

3.2 Solution Design

The Proficy Architect environment (Figure 4) enables the development of solution blueprints, used to visualize the process in the simulated mode. It contains user-friendly libraries and simple-to-configure blocks with which solution are developed. Various features in the menu and the explorer-type view are available for easy navigation in large solutions. A powerful troubleshooting compiler produces simulation and debugging on the execution ability of a blueprint after development.

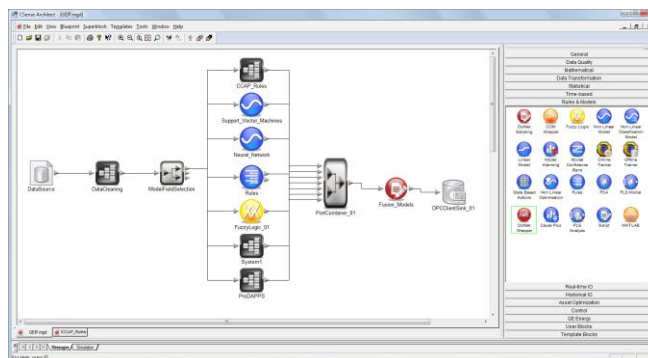


Figure 4: Proficy Architect IDE

This environment enables great flexibility in developing specific condition monitoring and decision-support solutions. The user can choose from high-level rapid prototyping tools to lower-level programming functions in scripting or in any of a number of programming languages that supports Component Object Model (COM, such as Visual C#/C++ and Visual Basic).

3.3 Online Deployment

The Action Object Manager provides a simple and easy way to deploy and monitor Action Objects. It allows users to easily maintain all your Action Objects from one central point of access. An Action Object (AO) is the name given to an executing blueprint. This blueprint could have been created and deployed from a number of services, including Architect, Troubleshooter Wizards, or other Proficy tools.

Proficy Advanced Analytics also provides a toolkit for fleet asset monitoring. The toolkit can apply the same action object to a large number of assets programmatically, which make fleet monitoring easier and more reliable.

3.4 Integration of Proficy with Existing Platforms

The Proficy Advanced Analytics software provides a series of interface tools (Figure 5) to integrate existing algorithms into the system (CSense Systems (Pty) Ltd). Some of the tools are General script, .NET script, COM Wrapper, .NET Wrapper, Matlab script. So users can plug their existing algorithms/ modules directly into the Proficy Advanced Analytics environment and seamlessly integrate with other part of the system. For example, Matlab code can be

plugged into the system directly (Matlab license is needed to run the code). Compiled Matlab objects can also be integrated. The COM Wrapper block allows the user to integrate external COM components within the Proficy Advanced Analytics environment. In this way the user can add custom functionality, own code/libraries, or third party libraries/components to the solution.

Those tools provide great flexibility to end users, who can easily re-use their existing modules develop and deploy PHM platforms. In summary, Proficy advanced analytics provides a series of tools from offline-analysis to online deployment so PHM users can focus on the most creative but challenging part of the job.



Figure 5: Tools in Proficy Advanced Analytics

4. IMPLEMENTATION

The PHM case study described for GE Energy was implemented on Proficy® platform. We tested the various PHM models developed on 200 units operational data for a six month duration at multiple time resolutions.

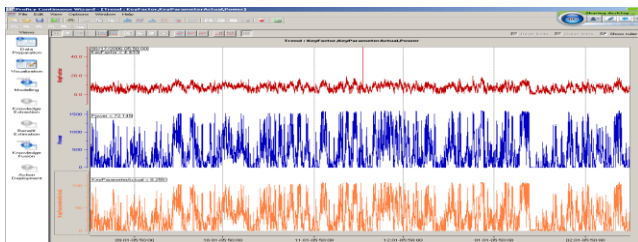


Figure 6: Multiple trending in Proficy®

Some of the data exploratory and model analysis done as explained in Section 2 earlier and was performed to understand the distributions of key parameters of interest, detecting outliers and evolution of key metrics over time. Figure 6 above depicts some of the basic plots explored in Proficy® toolset. As shown, multiple variables can be examined at instantaneous periods of time, understanding

basic distributions and plots such as scatter plots, line plots, overlay plots and histograms.

Deviations of various key parameters from their calculated values and set points can be crucial in a PHM advisory system. An example of such deviations is shown in Figure 7a.

Also, as depicted in Figure7b, some of these critical differentials are tracked over a period of time and early warnings can be picked up to perform enhanced monitoring of high risk units. Some of these can also be used to monitor certain frequently occurring failures and plan for inventory and spare parts.

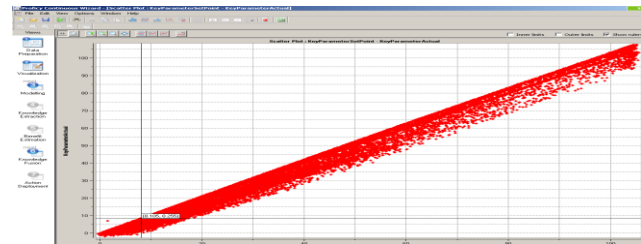


Figure 7a: Predicted Vs Actual Values

A total of over 1 million rows of operational data were analyzed in the above mentioned PHM case study. Initial results indicated a probability of detection over 80%. This is significant as there was little monitoring capability available for some of the turbine generation capability earlier and given that this is work in progress, we are hopeful to increase the probability of detection rates to very high values, keeping the false alarm rates under control.

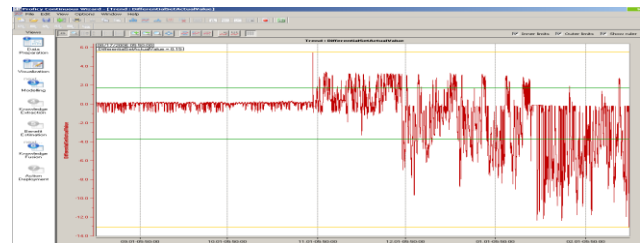


Figure 7b: Tracking evolution of deviations

The main objective is to understand and capture the patterns of increasing deviations and raise appropriate alarms for user to be able to perform exception based monitoring. This would ensure a high productivity and increased reliability.

5. CONCLUSION AND FUTURE DIRECTION

As depicted in this case study for GE Energy, a PHM system using real failures on key equipment was implemented using some of the existing platforms and using GE Intelligent Platform.

On the PHM algorithms side one of the key lessons learnt was to develop PHM algorithms with a high degree of explain-ability to end user: this ensures easy acceptance by field personal and relation of physics of failure with

advanced PHM methods seamless. As this is still work in progress, we plan to improve the quality of results by fusing multiple methods developed individually and using all sources of information available to monitor the equipment.

On the Proficy side, we plan on linking the current GEIP algorithmic capability with GE SmartSignals® algorithms and linking some of the existing legacy algorithms with the Proficy toolset. Also in pipeline is to enhance the native prognostic methods capability within Proficy, thus increasing its analytical power to include advanced methods.

The authors would like to acknowledge that Matlab is a trademark of Mathworks (<http://www.mathworks.com>) and .NET is a trademark of Microsoft (<http://www.microsoft.com>).

REFERENCES

- GE Intelligent Platforms. <http://www.ge-ip.com>
Mathworks: <http://www.mathworks.com>
CSense Systems (Pty) Ltd. User's manual for Proficy Advanced Analytics 5.0 (2011)
Casella, G (1990); Berger R. L Statistical Inference.
Duda R.O (2001); Hart P.E; Stork, D. G: Pattern Classification. John Wiley & Sons.
Hu, Xiao(2007), Qui, Hai and Iyer, Naresh: Multivariate Change Detection for Time Series Data in Aircraft Engine Fault Diagnostics, IEEE.
M. Markou (2003), S. Singh, Novelty Detection : A Review Part 1: Statistical Approaches, Signal Processing, Vol. 83(12), pp2481-2497.
Kumar, Vipin et al(2005): An Introduction to Data Mining.
Russel S(2002), Norvig P; Artificial Intelligence, A Modern Approach, Prentice Hall of India, 2nd edition.
- Jammu, Vinay(2010), et al; Review of Prognostics and Health Management Technologies, GE Global Research
Bonissone, P., & Kai Goebel, Soft Computing Techniques for Diagnostics and Prognostics
Ebeling (2005), C. E., An Introduction to Reliability and Maintainability Engineering, Tata McGraw-Hill Publishing Company, New Delhi.
Vachtsevanos(2006), G., Frank Lewis, Michael Roemer, Andrew Hess and Biqing Wu, Intelligent Fault Diagnosis and Prognosis for Engineering Systems, John Wiley & Sons, Hoboken, New Jersey, 2006

Subrat Nanda is a Lead Engineer at GE Global Research Center. He earned his B.S degree in Production Engineering from Nagpur University in India in 2001 and M.S degree in Autonomous Systems from University of Exeter, England UK in 2003. His research interests are mainly in applying machine learning and pattern recognition methods to industrial problems such as remote monitoring and PHM, Bayesian reasoning, bio-inspired optimization and fusion of data mining methods with physics.

Xiaohui Hu, Ph.D. Dr. Hu is Principal Engineer at GE Intelligent Platforms. He earned his B.S. degree in electrical engineering from Tsinghua University, Beijing, China in 1995 and Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, US in 2004. His research interests are remote monitoring and diagnosis, computational Intelligence, swarm intelligence, and data mining and knowledge discovery.

Author Index

- A**
- Addali, A. 168
Allen, David L. 190
An, Dawn 48, 300
- B**
- Balaban, Edward 15
Banerjee, Avisekh 419
Banks, C. 375
Banks, Jeff C. 490
Beaujean, Pierre-Philippe J. 123
Bechhoefer, Eric 275
Biswas, Gautam 31, 175
Boškoski, Pavle 368, 427
Bole, Brian 67
Bower, Gregory 469
Bregon, Anibal 198
Brown, Douglas W. 516
- C**
- Celaya, José R. 15, 31, 443
Chatterjee, Kaushik 1
Chebel-Morello, B. 257
Chen, Z. S. 510
Cheng, Jialin 143
Choi, Joo-Ho 48, 300
Chung, Jaesik 385
Cui, Jiuzheng 463
- D**
- Daigle, Matthew 15, 198, 323
DeCastro, Jonathan 56
Dempsey, Paula 275
Denaï, M. 257
Duc, Le Minh 149
Dunn, Christopher T. 40
- E**
- Eftekharnjad, B 168
- F**
- Fan, Jinhua 561
Feng, Qiang 463
Flynn, Eric B. 40
Foslien, Wendy 435
- G**
- Gašperin, Matej 368
Galvão, Roberto K. H. 540
Galvao, Roberto K. H. 293
Ge, Z. X. 510
- Gerez, Valerio 579
Ghavami, Peter K. 222
Goebel, Kai 15, 31, 67, 231, 323, 443
Goel, Alok 409
Goel, Nishith 419
Gola, Giulio 267
Gomes, João P. P. 540
Gomes, Joao P. P. 293
Gouriveau, Rafael 555
Grosvenor, Roger I. 502
Guan, Xuefei 94
Guanjun, Liu 572
- H**
- Hadden, George D. 175
Hafiychuk, V. 375
Hajrya, Rafik 532
Harrison, Craig 310
Hashemi, Ali 239
He, David 275
He, Jingjing 94
Hettler, Eric 56
Hoffman, Paul 453
Hu, Chao 385
Hu, Xiaohui 588
Hu, Z. 510
- I**
- Isom, Joshua 282
Lung, Benoit 547
- J**
- Javed, Kamran 555
Jha, Ratneshwar 94
Jing, Qiu 572
Juričić, Dani 368, 427
- K**
- Kapur, Kailash 222
Kehong, Lv 572
Kessler, Seth S. 40
Kim, Kyusung 435
Kim, Nam H. 300
Kim, Taejin 385
Klein, Renata 159
Ko, Sangho 48
Koopmans, Michael T. 209
Koul, Ashok K. 419
Koutsoukos, Xenofon D. 175
Kulkarni, Chetan 31
Kumar, Amar 409, 419

L	
Léger, Jean-Baptiste	547
Létourneau, Sylvain	479
Lacaille, Jérôme	579
Le, Daniel	361
Leão, Bruno P.	540
Leem, Sang Hyuck	48
Liu, Yongming	94, 398
Lu, Tsai-Ching	190
Luchinsky, D. G.	375

M	
Mack, Daniel L.C.	175
Masad, Eyal	159
Matsuura, Jackson P.	293
Mayer, Jeffrey	469
Mba, D.	168
Mbaya, Timmy	104
Mechbal, Nazih	532
Mehta, Ankit	247
Meicke, Stephen	209
Mengshoel, Ole	104, 310
Miller, J.	375
Ming, Tan Cher	149
Mjit, Mustapha	123
Modarres, Mohammad	1, 453
Monnin, Maxime	547
Mylaraswamy, Dinkar	175

N	
Nanda, Subrat	588
Narasimhan, Sriram	15, 323
Nystad, Bent H.	267

O	
Orchard, Marcos E.	8

P	
Paasch, Robert	209
Parthasarathy, Girija	435
Peng, Yang	572
Pflumm, J. Scott	490
Pisu, Pierluigi	239
Preston Johnson	114
Prickett, Paul W.	502

Q	
Quach, Patrick	231

R	
Rabbath, Camille-Alain	247
Rabiei, Masoud	453
Ramasso, Emmanuel	85
Raptis, Ioannis A.	77
Reichard, Karl	469

Ricks, Brian	310
Rooteh, Ensieh Sadat Hosseini	134
Roychoudhury, Indranil	15, 198, 323
Rudyk, Eduard	159

S	
Sadeghzadeh, Iman	247
Saffari, Mahdi	527
Saha, Bhaskar	15, 231, 323
Saha, Sankalita	15, 31, 323, 443
Salman, Mutasim A.	190
Saxena, Abhinav	443
Saxena, Bhavaya	409
Schumann, Johann	104
Sedaghati, Ramin	527
Senoussi, H.	257
Serir, Lisa	85
Smelyanskiy, V. N.	375
Srivastava, Alka	409
Stiharu, Ion	527
Sun, Bo	463

T	
Tang, Liang	8, 56, 67
Tang, Xidong	361
Tian, Zhigang	134, 143
Todd, Michael D.	40
Tumer, Irem Y.	209
Tyson, R.	375

U	
Uluyol, Onder	435

V	
Vachtsevanos, George	67
Vachtsevanos, George J.	8, 77, 516
Vendittis, David J.	123
Vergé, Michel	532
Vianna, Wlamir O. L.	293
Vismara, Mattia	334
Voisin, Alexandre	547

X	
Xiang, Yibing	398

Y	
Yang, Chunsheng	479
Yang, Y. M.	510
Yoneyama, Takashi	293, 540
Youn, Byeng D.	385

Z	
Zaluski, Marvin	479
Zemouri, Ryad	555
Zeng, ShengKui	463

Zerhouni, N. 257
Zerhouni, Noureddine 85, 555
Zhang, Bin 56
Zhang, Guicai 282

Zhang, Yilu 190
Zhang, Youmin 134, 143, 247, 561
Zheng, Zhiqiang 561