Stephen Ekwaro-Osire
Aparecido Carlos Gonçalves
Fisseha M. Alemayehu *Editors*

Probabilistic Prognostics and Health Management of Energy Systems

Springer

Probabilistic Prognostics and Health Management of Energy Systems

Stephen Ekwaro-Osire
Aparecido Carlos Gonçalves
Fisseha M. Alemayehu
Editors

Probabilistic Prognostics and Health Management of Energy Systems

Springer

Editors
Stephen Ekwaro-Osire
Department of Mechanical Engineering
Texas Tech University
Lubbock, TX
USA

Aparecido Carlos Gonçalves
Faculdade de Engenharia de Ilha Solteira
Universidade Estadual Paulista
Centro, Ilha Solteira, São Paulo
Brazil

Fisseha M. Alemayehu
School of Engineering, Computer Science and Mathematics
West Texas A&M University
Canyon, TX
USA

ISBN 978-3-319-55851-6 ISBN 978-3-319-55852-3 (eBook)
DOI 10.1007/978-3-319-55852-3

Library of Congress Control Number: 2017934323

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

All engineering systems—energy systems, transportation systems, micro-mechanical–electrical–systems, and computer chips, all of which are initially designed for longevity and durability—will, in due course, suffer damage initiation at the smallest level, under repeated hygro-thermal–mechanical–electrical–electromagnetic loadings. The materials used in these diverse engineering systems are both mechanical load-bearing as well as multifunctional. These materials are almost always heterogeneous, with very complex microstructures. The science of damage precursors in such materials, and thus in the engineering systems they are made of, is an emerging discipline. The detection of damage precursors, the measurement of the growth of damage (including cracks), and the prediction of the remaining useful life of an engineering system are the emerging science of diagnostics and prognostics. This will also involve sensors and actuators and thus forms the emerging Internet of Things.

The issues of damage precursors, measurement of damage progression by non-intrusive techniques, and the prediction of useful life are all subject to uncertainties, are all stochastic processes, and are data-driven. In the 1960s, in order to cope with data subject to noise, the so-called Kalman filter was discovered. Later, in order to cope with nonlinear physical phenomena, the so-called particle filters were discovered. Damage initiation as well as its development under repeated general loads is all highly nonlinear phenomena. Thus, Bayesian statistical methods are needed to cope with randomness in the measured data as well as in the predictive methodologies. Hence, the science of diagnostics and prognostics is rooted in the disciplines of filtering of noisy data, Bayesian statistics, uncertainties in computation, data-driven modeling, non-intrusive measurement techniques, high-performance near-real-time computations, the Internet of Things, the design of new materials with delayed or nullified damage precursors, and the general discipline of sustainment.

This monograph is a collection of papers dealing with emerging research in all these disciplines. Thus, it belongs to all engineering libraries. All the authors are to be complimented for contributing succinct summaries of the pieces of research which comprise the whole of the subject of this monograph.

Lubbock, TX, USA Satya N. Atluri

Acknowledgements

The edition of this book was supported by the São Paulo Researchers in International Collaboration (SPRINT) under Grant Number 2015/50026-1. The SPRINT grant was governed by the cooperation agreement for research between São Paulo Research Foundation (FAPESP), Brazil, and Texas Tech University, USA.

Contents

Part I
Trends and Applications

Probabilistic Prognostics and Health Management: A Brief Summary

Fisseha M. Alemayehu and Stephen Ekwaro-Osire

Abstract This chapter gives a brief summary of probabilistic prognostics and health management (PPHM) and presents a framework to implement PPHM to predict remaining useful life (RUL) of energy systems efficiently and with minimal uncertainty. The chapter also presents the way forward by indicating that an interdisciplinary research is critical so as consortium of multidiscipline experts will come together and discuss the implementation of the framework for enhanced RUL prediction. The prediction of RUL with minimal uncertainty will significantly lower or avoid the downtime of energy systems and thereby reduce the cost of energy. The reduction in cost will make renewable energy, like wind energy, cheaper.

Keywords Remaining useful life · Uncertainty · Probability · Prognostics and health management

1 Introduction

Energy is crucial to the security and prosperity of nations. Renewable and non-renewable sources of energy are being used to quench the demand of different societies in all over the globe. These sources of energy are mainly thermo-chemo-electro-mechanical systems that are subjected to uncertainty in future loading conditions, material properties, process noise and other design parameters.

The uncertainty in remaining useful life (RUL) prediction of these energy systems, for example the drive train of wind turbines, is high. The determination of RUL in the field of prognostics and health management (PHM) has to consider the uncertainties in the current state of the system, future loading conditions, future

F.M. Alemayehu (✉)
School of Engineering, Computer Science and Mathematics,
West Texas A&M University, Canyon, TX, USA
e-mail: falemayehu@mail.wtamu.edu

S. Ekwaro-Osire
Department of Mechanical Engineering, Texas Tech University, Lubbock, TX, USA

S. Ekwaro-Osire et al. (eds.), *Probabilistic Prognostics and Health Management of Energy Systems*, DOI 10.1007/978-3-319-55852-3_1

system parameters as well as the future process noise [1–3]. The uncertainties in the aforementioned quantities need to be quantified and the RUL estimation should reflect the uncertainty propagation of these quantities.

2 Methodology

A framework that could be implemented in different energy systems for the condition-based PPHM method has been developed and presented in Fig. 1. Efficient uncertainty quantification methods to quantify future loading, system parameters and process noise will be investigated. A physics-based or data-driven system model will be used to predict future states of the system while checking if threshold values that define failure are attained. Improved sensing techniques will be implemented to accurately measure the current condition of the energy system, for example the wind turbine gearboxes (WTG), and efficient filtering methods will enhance the estimated current state in light of the sensed data. Finally, using analytical and/or numerical probabilistic methods, the probability distribution function (pdf) of the RUL will be defined. The RUL results with minimal uncertainty will help the energy operator to decide if preventive/corrective maintenance will be needed to troubleshoot system failure.

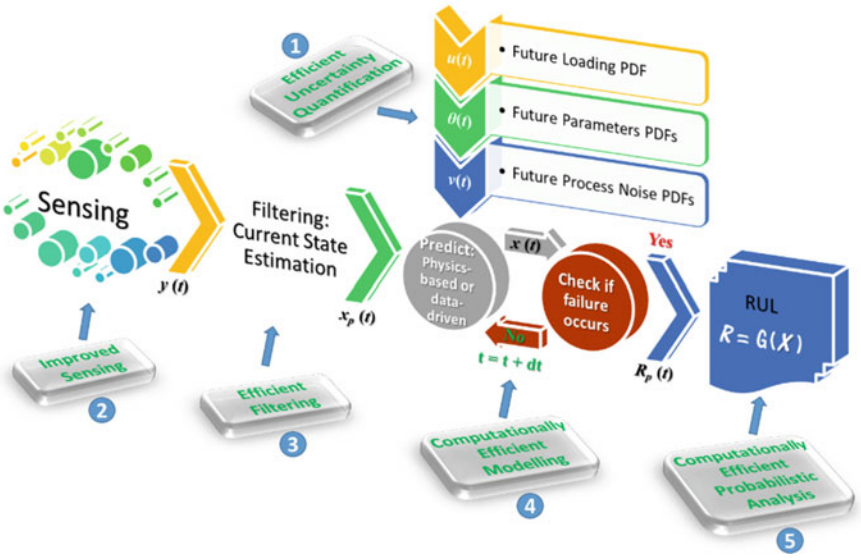


Fig. 1 Condition-based probabilistic prognostics and health management (PPHM)

Each Tablet shown in Fig. 1 is defined as follows:

Tablet 1. Efficient Uncertainty Quantification

The future of the system, i.e. future loading, operating and environmental conditions are not precisely known [1, 2]. PDF of future parameter uncertainties like load variations, degradation or variation in material properties, process noise need to efficiently be quantified. Future uncertainty quantification methodologies such as Maximum Entropy Principles [4, 5], ensemble techniques [6], Maximum Likelihood Evaluation (MLE) [7, 8], and neural networks [9, 10] can be implemented.

Tablet 2. Improved Sensing

As the quality of sensors increase, measurement uncertainty will decrease which in turn will enhance the precision of the current state estimation. In energy system, like in the case of the WTGs, oil cleanliness and lubrication parameters, oil temperature, vibrations, acoustic emissions can be monitored [11] to directly or indirectly [2] be used in the estimation of current state of the system. Measurement noise [12] and resonance problems as well as sensor installation (attachment) and powering are future challenges that the industry faces.

Tablet 3. Efficient Filtering

The present condition or state, i.e. at the time of prediction, has to be precisely estimated so as RUL will be predicted with minimum uncertainty. The state of the system may or may not be measurable using sensors [13]. Hence, sensor measurement data is directly or indirectly, via state quantification process using measurement function, used as an input to filtering approaches so as to improve the estimated current state in light of measured data [1, 12]. Using filtering approaches such as Kalman or Particle filtering techniques, the uncertainty in current state is described with a distribution. Efficient filtering methods and improved sensing can improve the estimate of the current states and thus reduce the uncertainty in RUL prediction [1].

Tablet 4. Computationally Efficient Modelling

Physics-based or data-driven efficient models need to be used to predict the future states. Numerical modeling techniques such as Finite Element Analysis (FEA) [14], Multibody Dynamic (MBD) [8, 15], and Peridynamics [16] or analytical models with exact solutions could be implemented. These models should describe progression of the faults and damages in time. Current state of the system, future loading, system parameters, and process noise are used as model inputs.

Tablet 5. Computationally Efficient Probabilistic Analysis

Efficient probabilistic techniques are critical to estimate uncertainty propagation in RUL estimation. Probabilistic sampling-based methods such as Monte Carlo (MC), Latin Hypercube Sampling (LHS), Adaptive Importance Sampling (AIS), or analytical methods such as First Order Reliability Method (FORM), Second Order Reliability Method (SORM), Advanced Mean Value (AMV) method [2, 12, 17, 18] could be implemented.

3 The Way Forward

Operators of energy systems are extremely interested in knowing the current state of their systems and thereby predict the future in such a way that preventive maintenance will be implemented. Eventually, operators would like to have high degree of certainty in the knowledge of the Remaining Useful Life (RUL) of their system at certain point of time. Such a subject is extremely intertwined that it needs the intervention of multidisciplinary team of experts to untangle the problem and come up with a better solution in determining the RUL with high degree of certainty. Hence, an interdisciplinary research is critical so as consortium of multidiscipline experts will come together and discuss.

The framework presented on Fig. 1 invites the involvement of experts from the fields of engineering, energy, atmospheric science, mathematics, management and others. Currently, the framework is on development stage and results of PPHM of selected energy systems, like the drivetrain of wind turbines, will be published in the future.

References

1. S. Sankararaman, K. Goebel, Why is the remaining useful life prediction uncertain?, in *Annual Conference of the Prognostics and Health Management Society 2013* (2012), pp. 1–13
2. S. Sankararaman, M.J. Daigle, K. Goebel, Uncertainty quantification in remaining useful life prediction using first-order reliability methods. IEEE Trans. Reliab. **63**(2), 603–619 (2014)
3. M.J. Daigle, K. Goebel, A model-based prognostics approach applied to pneumatic valves. Int. J. Progn. Health Manage. **2**, 1–16 (2011)
4. X. Guan, R. Jha, Y. Liu, Probabilistic fatigue damage prognosis using maximum entropy approach. J. Intell. Manuf. **23**(2), 163–171 (2012)
5. J.Y. Jung, C.H. Chin, J. Cardoso, An entropy-based uncertainty measure of process models. Inf. Process. Lett. **111**(3), 135–141 (2011)
6. W.S. Parker, Ensemble modeling, uncertainty and robust predictions. Wiley Interdiscip. Rev.: Clim. Change **4**(3), 213–223 (2013)
7. F.M. Alemayehu, S. Ekwaro-Osire, Loading and design parameter uncertainty in the dynamics and performance of high-speed-parallel-helical stage of a wind turbine gearbox. J. Mech. Design **136**(9), 091002-091002-13 (2014)
8. F.M. Alemayehu, S. Ekwaro-Osire, Probabilistic performance of helical compound planetary system in wind turbine. J. Comput. Nonlinear Dyn. **10**(4), 041003-041003-12 (2015)
9. D. An, N.H. Kim, J.H. Choi, Statistical aspects in neural network for the purpose of prognostics. J. Mech. Sci. Technol. **29**(4), 1369–1375 (2015)
10. J. Liu, A. Saxena, K. Goebel, B. Saha, W. Wang, An adaptive recurrent neural network for remaining useful life prediction of lithium-ion batteries (2010), pp. 0–9
11. S. Sheng, P. Veers, Wind turbine drivetrain condition monitoring—an overview, in *Mechanical Failures Prevention Group: Applied Systems Health Management Conference* (2011)
12. D. An, J.H. Choi, N.H. Kim, Prognostics 101: a tutorial for particle filter-based prognostics algorithm using Matlab. Reliab. Eng. Syst. Saf. **115**, 161–169 (2013)
13. M. Daigle, A. Saxena, K. Goebel, An efficient deterministic approach to model-based prediction uncertainty estimation, in *Annual Conference of the Prognostics* (2012), pp. 1–10

14. F. Zhao, Z. Tian, Y. Zeng, Uncertainty quantification in gear remaining useful life prediction through an integrated prognostics method. IEEE Trans. Reliab. **62**(1), 146–159 (2013)
15. F.M. Alemayehu, S. Ekwaro-Osire, Uncertainty considerations in the dynamic loading and failure of spur gear pairs. J. Mech. Design **135**(8) (2013)
16. S.A. Silling, Reformulation of elasticity theory for discontinuities and long-range forces. J. Mech. Phys. Solids **48**(1), 175–209 (2000)
17. D. Riha, B. Bichon, J. McFarland, S. Fitch, NESSUS Theoretical Manual (2010)
18. B. Thacker, D. Riha, S. Fitch, L. Huyse, J. Pleming, Probabilistic engineering analysis using the NESSUS software. Struct. Saf. **28**(1–2), 83–107 (2006)

Introduction to Data-Driven Methodologies for Prognostics and Health Management

Jay Lee, Chao Jin, Zongchang Liu and Hossein Davari Ardakani

Abstract This book chapter gives an overview of prognostics and health management (PHM) methodologies followed by a case study in the development of PHM solutions for wind turbines. Research topics in PHM are identified and commonly used methods are briefly introduced. The case study in wind turbine prognostics has shown in detail how to develop a PHM system for an industrial asset. With the advancement of sensing technologies and computational capability, more and more industrial applications are emerging. Current gaps and future directions in PHM are discussed at the end.

Keywords Prognostics and health management · Wind energy · Data-driven · Prognostics

1 Overview of Prognostics and Health Management (PHM)

1.1 Definition and the Value of Prognostics and Health Management

Prognostics and health management (PHM) is an engineering discipline that aims at minimizing maintenance cost by the assessment, prognosis, diagnosis, and health management of engineered systems. With an increasing prevalence of smart sensing and with more powerful computing, PHM has been gaining popularity across a growing spectrum of industry such as aerospace, smart manufacturing, transportation, and energy at breakneck speed. Regardless of application, one common expectation of PHM is its capability to translate raw data into actionable information to facilitate maintenance decision making. This practice in industry is often

J. Lee (✉) · C. Jin · Z. Liu · H. Davari Ardakani
NSF I/UCRC for Intelligent Maintenance Systems (IMS), University of Cincinnati, Cincinnati, OH, USA
e-mail: jay.lee@uc.edu

S. Ekwaro-Osire et al. (eds.), *Probabilistic Prognostics and Health Management of Energy Systems*, DOI 10.1007/978-3-319-55852-3_2

referred to as Predictive Maintenance, which, as estimated by Accenture [1], could possibly save up to 12% over scheduled repairs, reduce overall maintenance costs by up to 30% and eliminate asset failures up to 70%. For example, a study performed by National Science Foundation (NSF) indicates that Center for Intelligent Maintenance Systems (IMS), which is a leading research center in the field of PHM, has created more than $855 M of economic impact to the industry with a benefit cost ratio of 238:1 [2] through the development and deployment of PHM technologies to achieve near-zero unplanned downtime and a more optimized maintenance practice.

However, the value of PHM does not stop at maintenance alone. By performing smart analytics to asset usage data, users would be able to gain knowledge about how to achieve optimized performance of the asset. For instance, the State of Health (SoH) and Remaining Useful Life (RUL) of batteries on electric vehicles are highly dependent on the driving behavior. By analyzing the relationship between driving behavior and battery condition, a customized solution could be provided for the improvement of user's driving behavior and thus prolong battery life. Also, by relating process data with product quality metrics, predictive error compensation can be realized for increased product quality assurance. Additionally, asset usage and failure analysis could be fed back to the designers and manufacturers of the asset to nurture customer co-creation for an improved product design.

1.2 Research in Data-Driven Prognostics and Health Management

Besides huge economic potential in various industrial applications, PHM also holds great research value in the fields of signal processing, machine learning and data mining. Research in data-driven PHM requires an interdisciplinary background of computer science, signal processing, statistics, and necessary domain knowledge. In general, there are four major tasks for data analysis and modeling in PHM: pre-processing and feature extraction, health assessment, diagnostics, and prognostics, as indicated in Fig. 1. Prior to doing such tasks, it is critical to perform an overall analysis of the system to find out its critical components and the associated failure modes. Once the critical components of an asset have been determined, a data acquisition system needs to be devised to collect a sufficient set of measurements from the system for further analysis. Below is a description of the four data analysis steps for data-driven modeling of engineering systems:

- The task of pre-processing and feature extraction includes data quality evaluation, data cleaning, regime identification, and segmentation. Even though pre-processing does not directly offer immediate actionable information, it is a critical step and requires both domain knowledge and data processing skills to maintain the valuable parts of the data while removing its unwanted components.

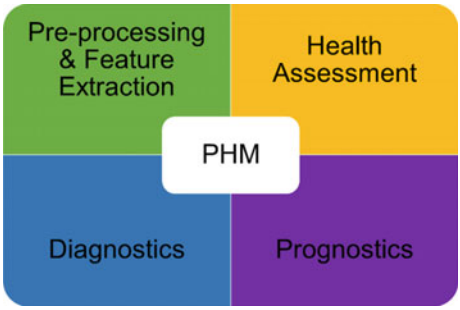


Fig. 1 Research tasks of data-driven PHM analytics

- The task of health assessment consists of estimating and quantifying the health condition of an asset by analyzing the collected data. If there is data of failed condition, then a Confidence Value (CV) could be generated to indicate the probability of asset failure. However, if asset failure data are not available, health assessment could be transformed into either a degradation monitoring problem for gradual faults or a fault detection problem for abrupt faults.
- In PHM, diagnostics refers to the classification of different failure modes by extracting the fault signatures from the data. For rotating machinery, for example, this process consists of enhancing the signal-to-noise ratio of the vibration signals and extracting the cyclo-stationary components which can represent defects in certain components of the machine. A collection of various features can be used along with a clustering or classification algorithm for developing a data-driven model for machine fault diagnosis.
- The task of prognostics refers to the prediction of asset health condition. If a short-term prediction is desired, time-series modeling is often utilized to predict when the machine would go out of threshold. If a long-term prediction is preferred, then the problem becomes a remaining useful life (RUL) prediction with many existing machine learning and statistics tools available. A confidence range will need to be defined for such predictions as the performance of the machine is also highly dependent on the usage pattern and proper maintenance actions that will be taken.

Besides the aforementioned research topics, feature selection and dimension reduction is of vital importance to achieve better PHM results. Health management approaches such as maintenance scheduling and operation management are also within the scope of PHM discipline, but this introduction will only focus on the analytics aspect of PHM.

1.3 Methodology

In this section, a systematic approach for designing and implementing PHM for industrial applications is provided, as described in Fig. 2. The process is separated

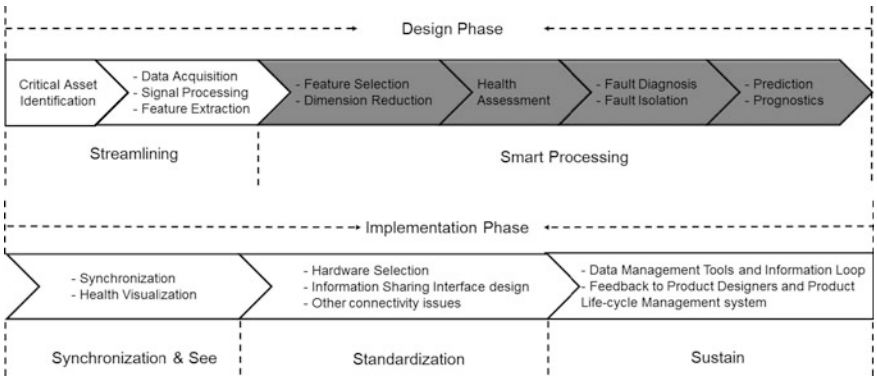


Fig. 2 General procedures for implementing PHM solutions

into five major steps following the "5S" methodology proposed in [3]. Within the 5S methodology, smart processing fulfills the major tasks of a PHM system and comprises the major intelligence of a system.

Smart processing focuses on utilizing data-to-information conversion tools to convert asset raw data into actionable information such as health indicators, maintenance recommendations, and performance predictions. All of this information is crucial for users to fully understand the current situation of the monitored asset and to make optimized decisions. Available data-to-information tools for smart processing includes: physics-based models, statistical models and machine learning/data mining algorithms. Among all three options, machine learning/data mining has its origins in computer practices and holds many advantages in industrial applications [4–6]: (1) Less domain knowledge requirements [7]: without building an exact mathematical model for the physical system, machine learning can also extract useful information by observing the input-output data pairs like the physical models do. This feature makes machine learning useful for complex engineering systems and industrial processes where prior knowledge is inadequate to build satisfactory physical models. (2) Scalable for a variety of applications [8]. (3) Easy implementation: compared with physics-based models, machine learning is more suitable to handle large-scale datasets since it requires less computation resources.

1.3.1 Algorithms

PHM algorithms refer to the data-driven models that serve as the computational core to transforming features into meaningful information. In Fig. 3, an example of this process for health assessment of induction motors is described. The process will be similar for diagnostics and prognostics, but the final visualization tool will be different depending on the purpose of users.

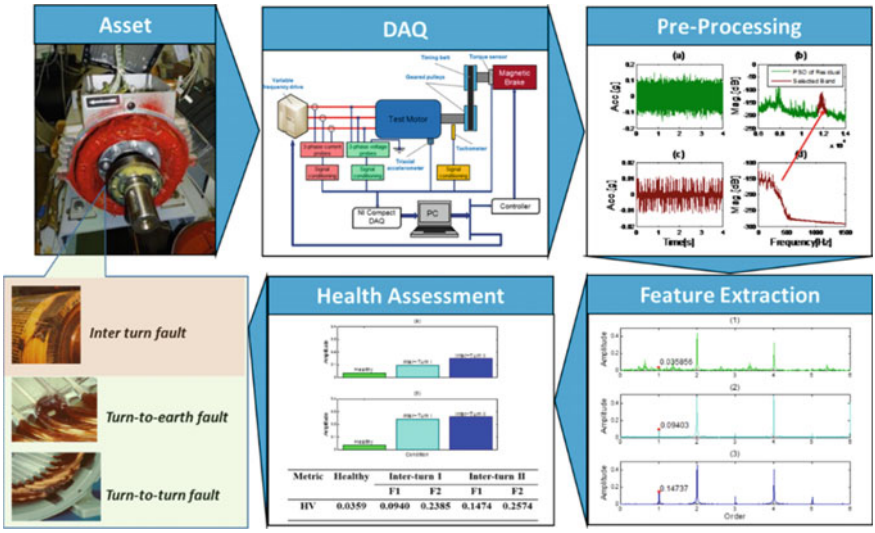


Fig. 3 Example of health assessment process [9]

Table 1 Summary of data pre-processing methods

	Method	Pros	Cons
1	Data quality inspection	Requires prior knowledge of signal type but effective in particular for vibration signal validation	Thresholds are needed for determining whether to include the signal in the analysis
2	Regime identification	Regime identification is important for developing baseline data sets in each operating condition	More sophisticated methods are needed for identifying the operating regime if the system changes operating conditions quickly
	Abnormality removal	Outliers, constant values, and missing values can dramatically increase the false alarm rate	Inclusion of domain knowledge is helpful for outlier detection

1.3.2 Data Pre-processing Algorithms

A summary of data preprocessing tools along with the merits and disadvantages of each method are provided in Table 1. In many applications, one or more of these methods is needed for ensuring that the data is suitable for additional processing and algorithm development. Data quality inspection is particular important to ensure that there are no sensor or data acquisition errors. However, developing an effective algorithm requires some prior knowledge of the signal characteristics or the distribution of the signal.

For more complicated working regimes, it might be necessary to have more advanced techniques for identifying the operating conditions. The regime

information allows one to develop baseline data sets in each operating condition and have a fair comparison and a local health model in each operating condition. Outlier removal from the measured data is also crucial in some of the applications as the presence of outliers can significantly affect the output of the analysis. Various methods exist for removing outlier instances from the signal or extracted features. However, these methods are purely based on the data distribution or characteristics, engineering experience and domain knowledge could be used to further improve the outlier removal algorithm.

1.3.3 Feature Extraction Algorithms

Numerous methods and algorithms are available for extracting characteristics or features from the measured signals, and an overview of some of the available techniques is provided in Table 2. For high frequency type signals such as vibration or current, there are well-established signal processing and feature extraction methods for extracting information from the time and frequency domain representation of the signal [10]. For rolling element bearings, mechanical shafts and gear wheels, there are several specific processing methods for extracting degradation features for these components [11]. Although it is advantageous to use the component specific feature extraction methods for high frequency vibration and current signals, they require a higher sampling rate, more computation, and more costly data acquisition systems.

For applications in which the monitored set of signals consists of trace signals such as temperature, pressure, or other controller signals, a different set of feature extraction algorithms would be recommended. Residual-based processing algorithms such as auto-associative neural networks, or principal component based

Table 2 Summary of feature extraction algorithms

	Method	Pros	Cons
1	Frequency based feature extraction methods	Frequency domain and envelope processing allows for component specific fault features to be extracted	Requires a higher sampling rate and more costly data acquisition
2	Residual based	More suited for low frequency signals and signals with a potential correlation	Residual processing algorithms can involve training a neural network which requires more computation
3	Statistics for each process segment or time slice	Ideal for process signals and provides a simple way of capturing the key aspects of the measured signal	Requires context information for identifying the various time slices of a process signal
4	Time statistics	Requires the least amount of domain knowledge and easiest to implement	Provides less specific information then other methods

methods are example methods that can be used to process trace signals [12]. For these types of algorithms, a baseline is established based on the normal operation of the machine. This baseline is used for comparing the predicted sensor values and processing the residuals as a sign of the drift in machine performance. The extraction of various statistical parameters is a straightforward but effective approach for characterizing the system condition from the available controller signals. In many instances, more insight can be gained by extracting statistics during different time slices, and a time slice could represent a different motion or action that is being performed by the monitored system [13]. If the context information regarding the process signals is not available, then extracting time statistics without any segmentation is a suitable alternative.

1.3.4 Health Assessment and Anomaly Detection Algorithms

A listing of the more commonly used algorithms for assessing machine health is provided in Table 3. The simplest approach for health assessment is to extract a health metric based on the weighted summation of the feature values. This health metric is simple to calculate and statistical thresholds for degradation detection can then be derived based on the distribution of the health value [14].

Various distances from normal health metrics can be used for determining the health condition of the monitored system or component. Mahlanobis distance and principal component analysis based Hotelling's T^2 statistics are distance metrics that incorporate the covariance relationship among the variables; however, Euclidean and other distance metrics are also commonly used [15]. Distance based

Table 3 Summary of health assessment algorithms

	Method	Pros	Cons
1	Weighed combination of features	Simple to implement, easier for setting thresholds based on the health value distribution	Does not account for the correlation relationship in the features
2	Distance from normal	Requires only baseline data sets for training the algorithm, distance methods can also account for the variable covariance relationship	Does not account for whether the features are lower or higher than expected
3	Statistical hypothesis Testing	Simple to implement and can be used to test whether the system is in a normal condition	Data might not fit assumed distribution for the hypothesis testing
4	Regression methods	Provides a mapping between the features and an output defect level or health value.	Requires an output value that is related to the health condition of the system and multiple data sets for training
5	One-class classifiers	Support vector data description algorithms can provide a boundary for detecting anomalies	Requires experience on selecting the appropriate parameters and kernel function

methods do not account for whether the features are higher or lower than normal. This has potential drawbacks in that a system can have a lower vibration then normal and this would still trigger a higher distance based health value and an anomalous condition.

A simple but effective approach for anomaly detection is the use of statistical hypothesis testing. Sequential probability ratio test, rank permutation test, and a T-test, are all examples of some of the more commonly used hypothesis test for anomaly detection [16]. Other anomaly detection based methods include a one-class classifier, such as the support vector data description (SVDD) algorithm [17]. Regarding SVDD, one disadvantage is the lack of guidance in the literature on which kernel functions or settings to use for a given application. Regression or neural network based methods are particular effective if sufficient data is available for developing the regression models. A neural network or regression model can be used to provide a mapping between the feature values and a health value or defect size [18]. The ability for the model to generalize usually requires multiple training data sets.

1.3.5 Health Diagnostic Algorithms

For root-cause analysis and diagnosis, there are many different methods and algorithms to perform this task, and a sample of some of the more commonly used methods are listed in Table 4. Incorporating engineering knowledge and experience into the diagnostic algorithm makes the use of fuzzy membership functions and rules an attractive technique [19]. However, it becomes more challenging to use fuzzy based diagnostic algorithm for new applications in which there is not sufficient experience on the failure modes and their signatures.

The use of a classification algorithm is a popular alternative if there is data from multiple health states including a baseline condition and several of the different failure modes that can occur. The use of neural networks, support vector machines, and Naïve Bayes algorithm are some of the more common classification algorithms used for machine condition monitoring [20]. By learning the relationship between the extracted features and the baseline and failure signatures, the classification

Table 4 Summary of health diagnostics algorithms

	Method	Pros	Cons
1	Fuzzy membership rules	Can include engineering knowledge and experience in the diagnostic algorithm	Requires experience for determining the rules and membership functions
2	Machine learning classifier algorithm	Can learn the relationship between the feature values and the output health label	Requires data sets from each fault class for training the algorithm
3	Bayesian belief network	Models the cause and effect relationship between the feature values and various health states	Determining the BBN structure requires experience or learning the network structure from data

method can accurately diagnose and label the health condition from the monitored system. Another method for diagnostics includes the use of a Bayesian Belief Network (BBN) which provides a network representing the casual relationship between the measured variable and the different failure modes or system conditions that can occur [21].

1.3.6 Prognostics Algorithms

A sample of the more commonly used remaining useful life prediction algorithms is presented in Table 5, along with the advantages and disadvantages of each method. Curve fitting based methods are relatively simple to apply as they do not require a substantial amount of training data or a detailed physical model that describes the fault progression. Neural network or regression based methods can directly relate the feature values with the remaining useful life of the monitored component or system [22]. These methods require substantial training data for learning this relationship and obtaining multiple run-to-failure data sets is not feasible in many applications.

A similarity-based prognostic algorithm is a unique method that matches the previous degradation patterns to the current degradation pattern of the monitored system [23]. The similarity-based prognostic algorithm can be quite accurate; however, it requires several runs to failure data sets in order to obtain a library for performing the degradation trajectory matching. In contrast to the previously described data-driven prognostic algorithms, the incorporation of the physics of failure with a stochastic filtering algorithm is an effective approach. Whether the fault propagation dynamic equations are linear or non-linear, and also whether the measurement noise is Gaussian or not, this group of methods can be used [24]. Applying model-based prediction algorithms using stochastic filtering does have some potential challenges. Only for a subset of applications does one have established models for describing the failure mechanism.

Table 5 Summary of prognostics algorithms

	Method	Pros	Cons
1	Curve fitting methods	Simple to implement, does not require substantial training data sets	Results are dependent on selecting an appropriate curve fitting model form
2	Neural network methods	Provides a mapping between the feature pattern and the remaining useful life	Requires several run-to-failure data sets for learning this relationship
3	Stochastic filtering methods	Incorporates the failure physics and can handle uncertainties in the modeling and sensor data	Requires a physical model to describe the failure mechanism
4	Similarity based prediction method	Accurate and can account for different degradation patterns or initial degradation conditions	Requires several run-to-failure data sets

In the following section, a case study is provided which further elaborates on how a comprehensive PHM system can be applied to a real-world application and how the industry can benefit from such a system.

2 Case Study in Wind Turbine Monitoring System

2.1 Project Background

Wind power industry has been growing exponentially world-wide since 2000. By the year of 2012, there were more than 200,000 wind turbines operating, with a total nameplate capacity of 282,482 MW [25]. The US Department of Energy (DoE) claims that it is technically feasible to meet its goal of 20% of the total energy requirements by 2030, but this will involve extensive research in all aspects such as structural design, manufacturing, operation and maintenance, and construction [26]. In spite of the inspiring facts about wind power industry, there are also some hidden risks and concerns for all the players. In addition to the initial investment for turbine construction, operation and maintenance is estimated to take 20–25% of the total cost for on-shore turbines and 18% for offshore turbines over the lifetime, and can increase to 30–35% share of cost by the end of life [27].

Prognostics and health management (PHM) can play an important rule in promoting the development of wind energy and reducing the operation costs by ensuring wind turbines are more reliable and productive. According to *Wind System Magazine*, 70% of total wind turbine maintenance costs are from unscheduled breakdown. And for a 100 MW scale wind farm, only 1% of availability increase can worth between $300–500 K of revenue per year. This view is also shared by the European Wind Energy Association (EWEA) to suggest that condition monitoring is a critical and integral system to the operation and maintenance [28].

By moving to condition based maintenance for wind turbine applications, there are numerous savings and benefits that can be provided as a service to the customer. The potential benefits include lower maintenance cost, a reduced risk of unplanned downtime, and higher asset utilization and uptime [3]. Different industries, such as semiconductor manufacturing, automotive, and machine tool among others, have benefitted from condition monitoring system integrated with advanced PHM tools. However, a similar level of advancement has not been developed in the wind turbine industry due to the very strict system integration requirements and low public acceptance.

The existing monitoring systems for wind turbines mainly fall into two categories: Supervised Control and Data Acquisition (SCADA) system and Condition Monitoring System (CMS). SCADA system has a variety of sensors to collect data from critical components and external environment. The data is used as the inputs for the control systems of pitch angle, yaw, and braking to name a few. CMS mainly have accelerometers and AE sensors mounted on the critical parts of

drivetrain and generator. Vibration level and related features are used to assess the health conditions of the components nearby. However, neither of the two systems are more than data collection system, and very limited and indirect information of operation risks is given to the operator to make optimal maintenance decisions.

2.2 Benefits to Users

The users that can directly benefit from the condition monitoring system are wind turbine OEMs and operators. For OEMs, the benefits include increasing the competitiveness and reliability of their wind turbines, reducing the maintenance and repair costs in warranty period, and with a minimum increase in prices. For operators, they can benefit from increasing availability of the turbines, reducing safety hazards, and reducing operation costs while increasing the revenue.

The benefits to turbine OEM and operators are summarized from the following aspects:

Increase Reliability of wind turbines (OEM, Operator): The condition monitoring system can benefit OEMs to increase the reliability of their wind turbines to increase the competitiveness of their product. This requires reducing the unplanned breakdown by detecting any incipient faults and repairing in the early stages of failure.

Reduce Warranty Costs (OEM): The warranty for wind turbines is usually 5-10 years with coverage of defects or faults in material, wear and tear, gradual deterioration, inherent vice, and latent defects. The price of an extended warranty ranges from $30,000 per turbines per year for a 1.5 MW turbine to $150,000 per turbine per year for a 3 MW turbine (Mike McMullen, *Wind Power Magazine*). However, the cost for warranty on the OEM's side is highly dependent on the failure rate of the components. Repair and replacement of critical components are usually very expensive. Gearbox replacement can cost $250–350 K, generator replacement is $90–120 K, and blade replacement can be $120–200 K [29]. Hence, it is very important to monitor the critical components to avoid severe damage and to reduce warranty costs.

Increase Availability of Wind Turbines (Operators): For operators, it is ideal that the turbine can run 24/365 to maximize the revenue. It can be seen from simple match that for a 100 MW scale wind farm, only 1% of availability increase can be worth between $300–500 K of revenue per year. The goal of increasing availability of wind turbines can be achieved from two aspects: decrease the failure rate of critical component, and decrease the maintenance and repair response time.

Reduce the Redundancy of Maintenance (OEMs and Operators): Whatever information or suggestions are given by the system should be accurate. False alarms and fault misdetection should be controlled to a very low limit. This requires the DAQ system and analytical modules to have a high level of accuracy and optimal settings for threshold.

2.3 Method Development

2.3.1 Identify Critical Subsystems/Components

According to the studies performed by Faulstich and Hahn [29], the component failure frequency and the average downtime per failure are listed in Fig. 4. The components that are suitable for predictive maintenance are those with a low failure rate but with a very high downtime per failure. Hence the components that will be included in the fault localization and diagnosis system are: the drivetrain, the yaw system, the pitch system the rotor blade, and other selected critical electrical components. The total breakdown of those components causes more than 80% of total unplanned downtime of the wind turbine.

2.3.2 Data Acquisition/Signal Selection

Most wind turbines are instrumented with supervised control and data acquisition system (SCADA), while CMS is installed based on customers' (operator) requirement. As shown in Fig. 5, the SCADA system includes variables related to the operating condition of the key components. The sampling frequency is usually very low, and the output data is the statistical values such as mean, peak values and standard deviation. The sampling frequency of the SCADA system varies for different wind turbine settings, but is able to be customized according to data analysis requirement. For data storage, since the data volume is not very large, and SCADA

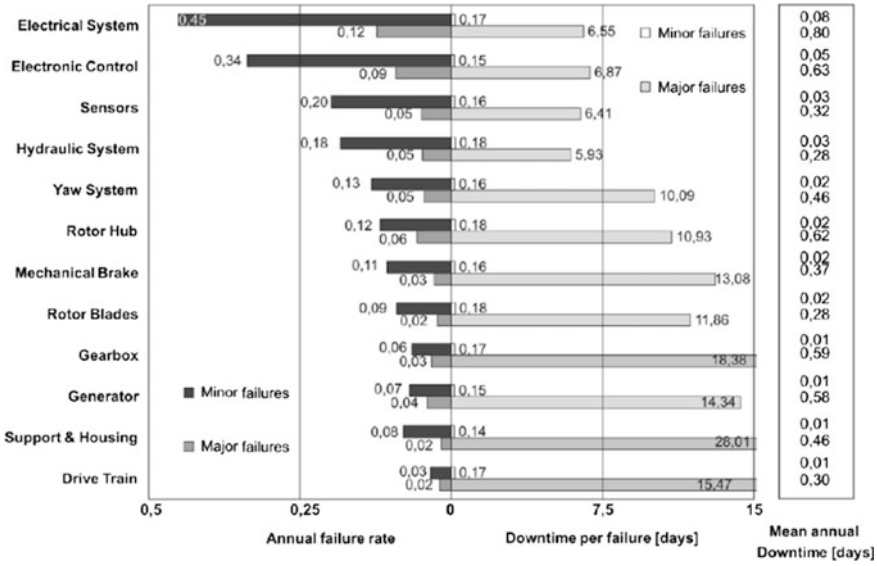
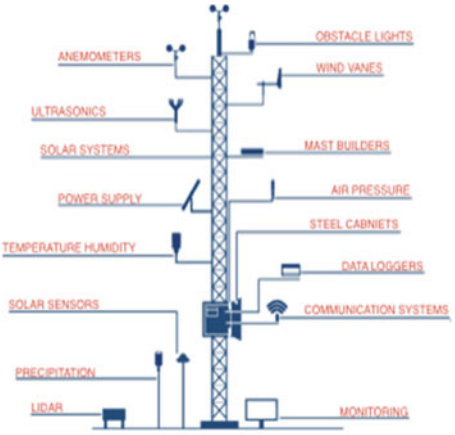


Fig. 4 Wind turbine failure rate and caused downtime [29]



CATEGORY	VARIABLE EXAMPLES
Ambient	Temperature, wind direction, wind speed
Blades	Pitch angle
Controller	Hub temperature, Ground temperature.
Gear	Gear bearing temperature, oil temperature
Generator	Bearing temperature, rotation speed
Grid	Production voltage, current, power factor
Hydraulic	Hydraulic oil temperature
Nacelle	Direction, temperature
Production	Average power, accumulated power
Rotor	Rotation speed
System	Logs of active alarms, turbine state
Hour Counter	Service hours

Fig. 5 List of SCADA variables [27]

Table 6 Design and operations summery of DAQ system

DAQ system	Sampling frequency	Collection interval	Data length	Data storage
SCADA	Range from 1/600–1/30	Continuous collection	1 sample per collection	From beginning of life
CMS	And from 1 K to 10 s KHz	Periodically or event based	From seconds to minutes	From beginning of life

data is important to provide reference information for maintenance actions, the historical data are always stored through lifetime.

The CMS data shall be available for fault localization and diagnosis for critical components. For data format, the sampling frequency determines the range of frequency spectrum while the data length determines the resolution of frequency spectrum in FFT analysis. The sampling frequency should be high enough to capture gear mesh frequencies and bearing characteristic frequencies for high-speed components in drivetrain, and the resolution of the frequency spectrum should be reasonably high. Since CMS collects data in very high frequency, and usually the monitored components are very stable and reliable, there is no need for continuous collection to reduce data storage and processing burden. Hence the data acquisition is usually triggered based on routine or event-based manners. Table 6 shows a summary for the DAQ system of wind turbine.

2.3.3 Multi-regime Modeling for Turbine Global Health Assessment

The global health of wind turbines shall reflect their generation efficiency, namely the efficiency to convert wind energy to electrical energy. Data from SCADA

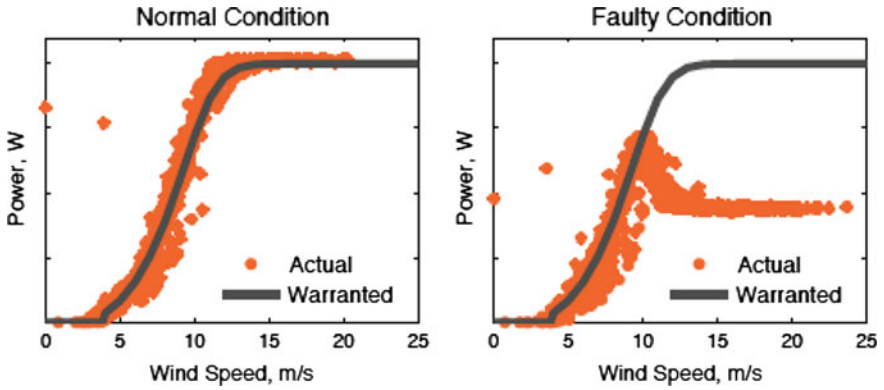


Fig. 6 Wind turbine power curve under normal and faulty conditions [30]

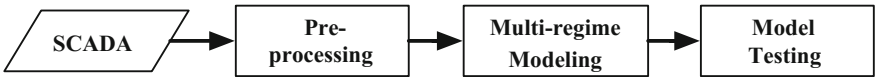


Fig. 7 Flow of wind turbine prognostic modeling

system shall be used to estimate the global health value. Figure 6 shows the power curve of wind turbines under normal and faulty conditions, and it is desired to quantify this kind of change and relate it to the loss of power efficiency.

SCADA parameters, including output power, wind speed, wind direction and pitch angle, is used to model turbine performance. Historical data is fed through a pre-processing module first to remove any outliers and undesired operating regimes for performance analysis. A multi-regime method is used to model the baseline behavior based on data from a selected training duration. Data from subsequent duration is then modeled and tested against the trained model, using distance metrics as a comparison method to quantify the deviation of testing data. Continuous testing can generate frequent evaluation of turbine performance and provide insight of turbine degradation over time with considerable time granularity, which could lead to valuable prediction (Fig. 7).

2.3.4 The Proposed Approach to Assessing Turbine Performance

SCADA variables are first fed into a pre-processing module to go through the following analysis.

1. Outlier filtering based on iterative Grubbs' test.
2. Rule-based non-operational regime filtering, where observations are filtered when wind speed is below cut-in and power output is zero.

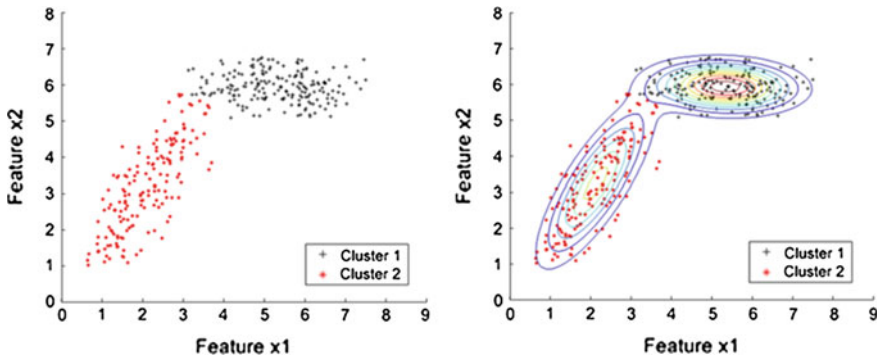


Fig. 8 Gaussian mixture model (GMM) based multi-regime clustering

3. Curtailment event filtering, where observations are filtered when a stalling event due to wind gust is determined to have occurred based on wind speed deviation and pitch angle deviation.
4. Data standardization, which equalizes variable contribution in the multivariate dataset (Fig. 8)

Gaussian Mixture Model (GMM), a probabilistic clustering model, is used to model the training data.

$$H(x) = \sum_{i=1}^{n} p_i h(x; \theta_i). \tag{1}$$

It partitions data into a mixture of Gaussian components with membership assignments where the component parameters and membership weights are estimated using techniques such as Expectation Maximization. The number of cluster, n, is decided based on the "goodness-of-fit" of the model when n is chosen as different numbers. A scoring method such as Bayesian Information Criterion (BIC) is used to evaluate model accuracy.

A testing model can be estimated with same GMM method for new data. An L2 distance between the two mixture models can be calculated, based on computing distance between all possible pairings of Gaussian components of the training and testing model. The normalized L2 distance is considered as a confidence value (CV) of turbine power performance.

$$\|H(x) \cdot G(x)\|_{L2} = \sum_{i=1}^{n} \sum_{j=1}^{m} p_i q_j \left\| h(x; \theta_i) \cdot h(x; \phi_j) \right\|_{L2}, \tag{2}$$

$$CV = \frac{\|H(x) \cdot G(x)\|_{L2}}{\|H(x)\|_{L2} \|G(x)\|_{L2}}. \tag{3}$$

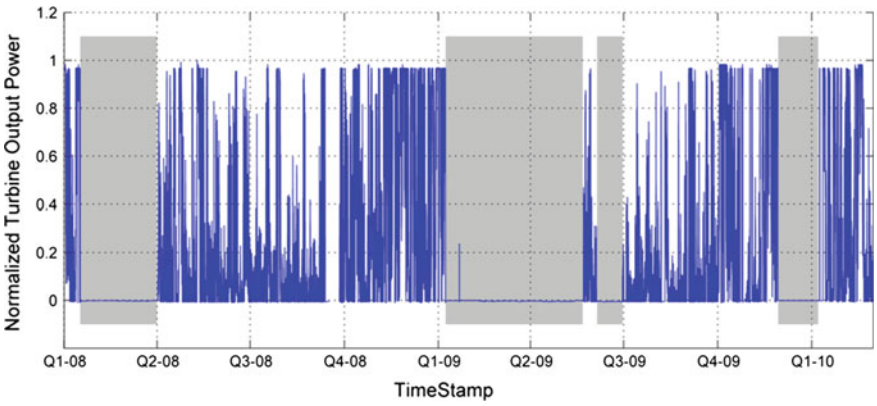


Fig. 9 Active power and failure events

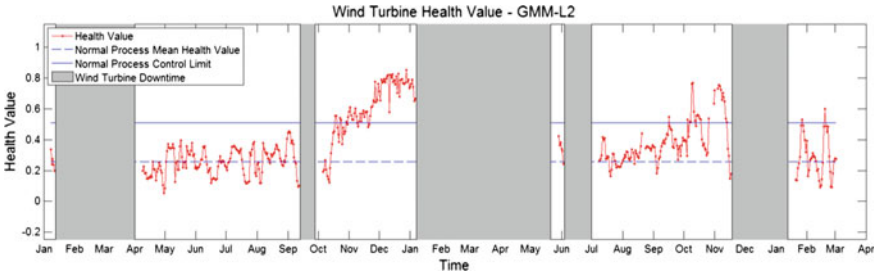


Fig. 10 Health risk increases between downtime

A SCADA dataset acquired from an onshore large-scale turbine is used to validate the proposed methodology for estimating the global health estimator (GHE). The duration of the data is 26 months, during which different parameters are extracted from the SCADA module every 10 min. The actual power output is shown in Fig. 6 where three major downtime events are highlighted in grey shadowed areas: (1) Q1-08 – Q2-08, (2) Q1-09 – Q3-09 and (3) Q4-09 – Q1-10 (Figs. 9 and 10).

2.3.5 Vibration-Based Condition Monitoring for Drivetrain System

Prognostics techniques are mainly developed for key drive train components that cause costly and sometimes catastrophic failures, including rotor, gearbox and generator. In many applications, CMS and SCADA systems are used separately for condition monitoring purposes, mainly due to the issue of data availability shared by different shareholders. A degradation assessment framework is proposed to

integrate both the CMS data and SCADA variables for the evaluation of drivetrain degradation for the scenario when both data resources are available.

The framework of prognostics and diagnosis for drivetrain gearbox based on CMS vibration signals is shown in Fig. 11. In the framework, vibration signals first go through an automatic data quality check process to make sure the DAQ system is problem free. The data quality check process includes a series of check criteria as shown in Table 2. Afterwards, instance selection of vibration signals is performed based on the working regime variables from SCADA data. Vibration signals under certain conditions such as no energy generation will be removed. The selected vibration signals are then applied to the signal processing and feature extraction process to extract gearbox health related features. A collection of features will be extracted from the processed signals from frequency spectrum, TVDFT order tracing, spectral kurtosis filtering, Cepstrum analysis, and envelop analysis. Those features will be used as input for degradation assessment algorithms such as Self-Organizing Map Minimize Quantitation Error (SOM-MQE) and Principle Component Analysis techniques, so that the deviation of current condition from baseline model is quantified by distance metric.

In the training process, the input features are projected to output units by the weight vector in the mapping layer. The output units will compete with each other to determine the cluster of units with the best similarity, and the mapping layer will adjust the weight vector for the wining units. Hence when the training process is

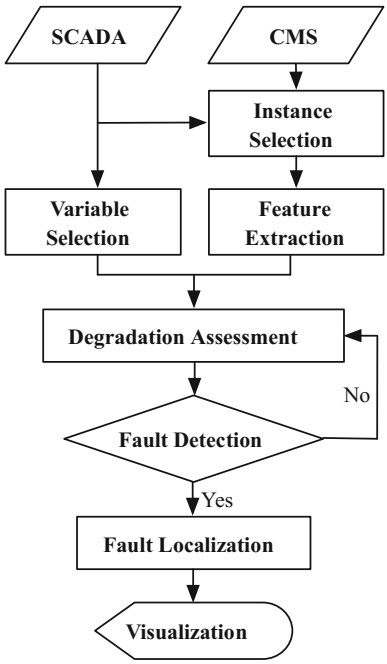


Fig. 11 Framework for drivetrain components health assessment [31]

finished, the units in output layer will gather together to form several clusters, and each cluster corresponds to the same class of input data. If the training data are all under healthy condition, the units will gather to form one (single-regime) or several (multi-regime) clusters that represent the feature characteristics under healthy condition. In the testing process, the input feature vector (x_i) is projected into the output layer with the same weight vector, and its Euclidean distances to the units (w_j) are calculated based on Eq. 4. The unit that has the smallest distance to the projected vector is called the 'Best Matching Unit (BMU)', and the corresponding distance is called 'Minimum Quantitation Error (MQE)'. The MQE value is calculated by (Table 7):

Table 7 Data quality check criteria [32]

Check method	Data processing	Check value	Threshold
Mean check	Mean value of vibration signal	Mean value	Smaller than 1e-5 (should be decently small)
RMS check	RMS value of vibration signal	RMS value	1e-5–0.05 (minimum energy rule and dynamic range rule)
Parseval's theorem-based Energy conservation rule	Time domain RMS and frequency domain RMS level should be close (conservation of energy for FFT)	RMS(x(t)) − RMS (X(f))	Smaller than 0.1%
Statistical distribution rule	Fit normal distribution of vibration signal	Hellinger-like distance and Komogorov distance of empirical and fitted distribution	<0.12 for K-distance <0.1 for H-distance
N-point rule	N neighbor points with the same value	N-point	Depends on sampling frequency. (<1 for our case due to very high sampling frequency)
U-point check	Number of unique points in the vibration signal	Portion of unique points to the length of dataset	>99.99% for our case
Positive and negative point check	Portion of positive and negative points to the length of dataset	Max[P(+), P(−)]	<52% for our case (the value should be close to 50%)
Derivative check	Derivative of vibration signal	RMS value of derivative signal; number of derivative value that exceeds threshold	0.015 for RMS derivative;

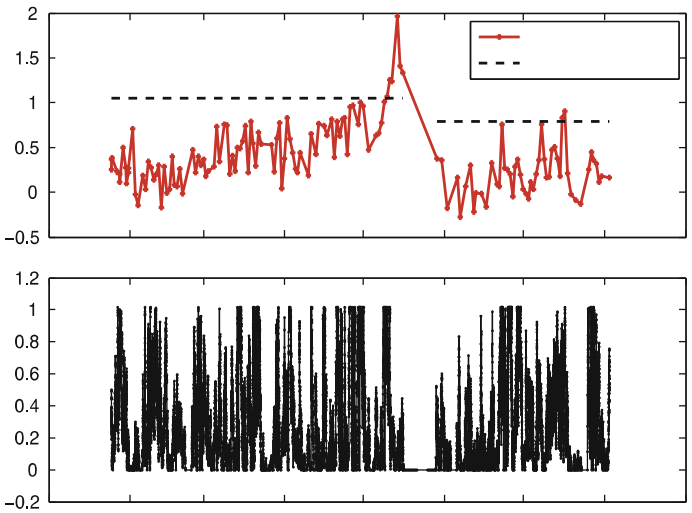
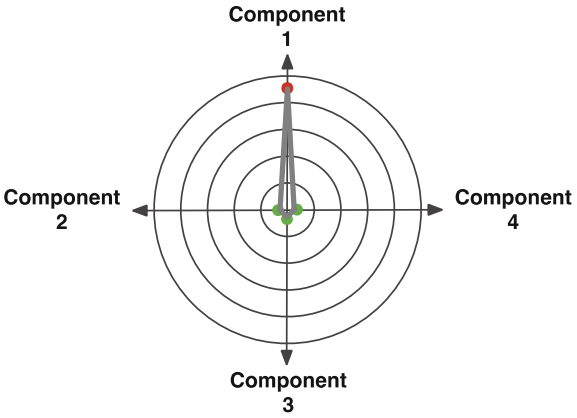


Fig. 12 Fault prognostic with SOM-MQE approach for offshore wind turbine drivetrain (*Note* Legend and axes labels were intentionally removed to keep the confidentiality of the data.)

Fig. 13 Radar chart for fault localization



$$\|x_i - w_c\| = \min_i \{\|x_i - w_j\|\}, \tag{4}$$

$$MQE = \|x - w_{BMU}\|. \tag{5}$$

This method is validated based on the historical data collected from an offshore wind turbine. The data span was 15 months, and had a severe damage at its rotor bearing that has caused downtime of 2 weeks. Figure 12 shows the change of MQE value before and after the breakdown. An incipient fault was detected five days before the severe damage.

As observed from the SOM-MQE result, there is a short duration in the middle of the history when MQE value noticeably exceeded the MQE threshold. The MQE excess occurred about five days before an operation pause due to a failure. The result shows that SOM-MQE is capable of detecting drivetrain anomaly at an early stage. A radar chart is created to view component criticality simultaneously. In this chart, each axis represents the contribution of each component to MQE abnormality. The closer the data point is to the center, the smaller the contribution is (Fig. 13).

3 Industrial Implementation and Gaps

3.1 Available Software and Platforms

As the industry recognizes the potential and business values in different sectors, a lot of companies have developed their PHM solutions/software/platform to achieve predictive modeling for industrial users.

- GE has announced Predix™ as a cloud-based service platform to enable industrial-scale analytics for management of asset performance and optimization of operations [33].
- National Instruments introduced Big Analog Data™ three-tier architecture solution [34], as well as LabVIEW Watchdog Agent™ Toolkit to support smart analytics solutions throughout different big data applications [35, 36].
- Many startup companies emerge recent years providing scalable PHM solutions, such as:
  - Predictronics (http://www.predictronics.com/) provides vertical solutions in various industrial applications from component level to fleet systems.
  - Uptake (http://www.uptake.com/) has been strategically working with Caterpillar and aims at developing a general PHM software.
  - Sparkcognition (http://www.sparkcognition.com/) has products concerning both cyber security and machine prognostics.
  - Trendminer (https://www.trendminer.com/) provides predictive analytics solutions to majorly process industry.

Besides, equipment makers themselves are also developing customized PHM systems for their own machines. For example, Prizm™ by Applied Materials for semiconductor manufacturing equipment [37], or RigWatch® by Canrig for their oil and gas applications [38].

3.2 Gaps and Future Directions

3.2.1 Preprocessing

"Industrial Big Data" is usually more structured, more correlated, more orderly in time and more ready for analytics [6]. This is because "Industrial Big Data" is generated by automated equipment and processes, where the environment and operations are more controlled and human involvement is reduced to minimum. Nevertheless, the values in "Industrial Big Data" will not reveal themselves after connectivity is realized by "Industrial Internet". Even though machines are more connected and networked, "Industrial Big Data" usually possess the characteristics of "3B" [6], namely:

- Below-Surface
  - General "Big Data" analytics often focuses on the mining of relationships and capturing the phenomena. Yet "Industrial Big Data" analytics is more interested in finding the physical root cause behind features extracted from the phenomena. This means effective "Industrial Big Data" analytics will require more domain know-how than general "Big Data" analytics.
- Broken
  - Compared to "Big Data" analytics, "Industrial Big Data" analytics favors the "completeness" of data over the "volume" of the data, which means that in order to construct an accurate data-driven analytical system, it is necessary to prepare data from different working conditions. Due to communication issues and multiple sources, data from the system might be discrete and un-synchronized. That is why pre-processing is an important procedure before actually analyzing the data to make sure that the data are complete, continuous and synchronized.
- Bad-Quality
  - The focus of "Big Data" analytics is mining and discovering, which means that the volume of the data might compensate the low-quality of the data. However, for "Industrial Big Data", since variables usually possess clear physical meanings, data integrity is of vital importance to the development of the analytical system. Low-quality data or incorrect recordings will alter the relationship between different variables and will have a catastrophic impact on the estimation accuracy.

Therefore, preprocessing and how to ensure data quality would be an important issue in PHM. The evaluation of data quality does not have to be limited to the inspection of signal validity, but can also include trend detection to evaluate the predictability, cluster analysis to evaluate potential for fault diagnosis, etc. [39].

3.2.2 Fleet-Based PHM

A fleet refers to a set of assets/machines that share some common characteristics that can be used to group them together according to a specific purpose. e.g. air crafts, vessels, wind turbines, trains, etc. Modern manufacturing enterprise scale is becoming larger and individual asset-based PHM might not be able to sufficiently fit in the changing environment in future.

Fleet-based PHM will be more accurate than conventional individual asset-based PHM:

- Prediction: similarity-based prediction
- Fault detection: peer-to-peer comparison, multiple kernel learning
- Compensation of training data insufficiency
  - Peer comparison without long history of individual baseline data for training.

3.2.3 General PHM Platform

Today's PHM solutions are still very customized and confined to one application. Different applications would have different data acquisition and storage system, different domain knowledge-dependent features and different monitoring purposes. It is very difficult to create a platform that could cover all kinds of applications.

One way of expanding the scope of a PHM solution is to combine several mainstream component/machine level PHM solutions together, and have users choose tools from similar applications. An alternative approach is to build up a standard platform where analytical tools are available but not customized. For such platform, background knowledge about how to use these tools for their application is required.

References

1. A. Alter, P. Banerjee, P. E. Daugherty, W. Negm, *Driving Unconventional Growth through the Industrial Internet of Things*, 2014
2. D.O. Gray, D. Rivers, *Measuring the Economic Impacts of the NSF Industry/University Cooperative Research Centers Program: A Feasibility Study*, 2012
3. J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, D. Siegel, Prognostics and health management design for rotary machinery systems—reviews, methodology and applications. Mech. Syst. Signal Process. **42**(1), 314–334 (2014)
4. M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms* (Wiley, 2011)
5. I.H. Witten, E. Frank, *Data Mining: Practical Machine learning tools and Techniques* (Morgan Kaufmann, 2005)
6. K.P. Murphy, *Machine Learning: a Probabilistic Perspective* (MIT press, 2012)
7. M. Pecht, R. Jaai, A prognostics and health management roadmap for information and electronics-rich systems. Microelectron. Reliab. **50**(3), 317–323 (2010)

8. Z. Ge, Z. Song, *Multivariate Statistical Process Control: Process Monitoring Methods and Applications* (Springer Science & Business Media, 2012)
9. C. Jin, A.P. Ompusunggu, Z. Liu, H.D. Ardakani, F. Petre, J. Lee, Envelope analysis on vibration signals for stator winding fault early detection in 3-phase induction motors. Int. J. Progn. Health Manag. **6**, 12 (2015)
10. A.K.S. Jardine, D. Lin, D. Banjevic, A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mech. Syst. Signal Process. **20**(7), 1483–1510 (2006)
11. R.B. Randall, *Vibration-Based Condition Monitoring: Industrial, Aerospace and Automotive Applications* (John Wiley & Sons, 2011)
12. J.W. Hines, R. Seibert, Technical review of on-line monitoring techniques for performance assessment. State-of-the-Art **1** (2006)
13. G.A. Cherry, Methods for Improving the Reliability of Semiconductor Fault Detection and Diagnosis with Principal Component Analysis, 2006
14. E. Bechhoefer, D. He, P. Dempsey, Gear health threshold setting based on a probability of false alarm, in *Proceedings of Annual Conference of the Prognostics and Health Management Society*, 2011
15. H. Oh, M.H. Azarian, M. Pecht, Estimation of fan bearing degradation using acoustic emission analysis and Mahalanobis distance, in *Proceedings of the Applied Systems Health Management Conference*, pp. 1–12, 2011
16. R. Ganesan, A. N. V. Rao, and T. K. Das, A Multiscale Bayesian SPRT Approach for Online Process Monitoring, in *IEEE Transactions of Semiconductor Manufacturing*, vol. 21.3, pp. 399–412, 2008
17. D. Tax, A. Ypma, R. Duin, Support vector data description applied to machine vibration analysis, in *Proceedings of 5th Annual Conference of the Advanced School for Computing and Imaging (Heijen, NL)*, pp. 398–405, 1999
18. D. He, E. Bechhoefer, Development and validation of bearing diagnostic and prognostic tools using HUMS condition indicators, in *Proceedings of 2008 IEEE Aerospace Conference*, pp. 1–8, 2008
19. D.J. Cleary, P.E. Cuddihy, A novel approach to aircraft engine anomaly detection and diagnostics, in *Proceedings of 2004 IEEE Aerospace Conference*, vol. 5, pp. 3468–3475, (2004)
20. W. Yan, F. Xue, Jet engine gas path fault diagnosis using dynamic fusion of multiple classifiers, in *Proceedings of 2008 IEEE International Joint Conference on Neural Networks*, pp. 1585–1591, 2008
21. L. Yang, J. Lee, Bayesian Belief Network-based approach for diagnostics and prognostics of semiconductor manufacturing systems. Robot. Comput.-Integr. Manuf. **28**(1), 66–74 (2012)
22. N. Gebraeel, M. Lawley, R. Liu, V. Parmeshwaran, Residual life predictions from vibration-based degradation signals: a neural network approach. Ind. Electron. IEEE Trans. **51**(3), 694–700 (2004)
23. T. Wang, J. Yu, D. Siegel, J. Lee, A similarity-based prognostics approach for remaining useful life estimation of engineered systems, in *Proceedings of International Conference on Prognostics and Health Management*, pp. 1–6, 2008
24. M.E. Orchard, A Particle Filtering-Based Framework for On-Line Fault Diagnosis and Failure Prognosis (Georgia Institute of Technology)
25. S. Sawyer, K. Rave, *Global Wind Report–Annual Market Update 2012*, (GWEC, Glob. Wind Energy Council, 2013)
26. U.S. Department of Energy, Wind Power Today 2010, 2010
27. S. Sheng, P.S. Veers, *Wind Turbine Drivetrain Condition Monitoring-An Overview* (National Renewable Energy Laboratory, 2011)
28. P. Gardner, A. Garrad, L.F. Hansen, A. Tindal, J.I. Cruz, L. Arribas, N. Fichaux, *Wind Energy-The Facts Part 1 Technology* (EWEA, Garrad Hassan Partners, UK CIEMAT, Spain, 2009)

29. S. Faulstich, B. Hahn, P.J. Tavner, Wind turbine downtime and its importance for offshore deployment. Wind Energy **14**(3), 327–337 (2011)
30. E.R. Lapira, *Fault Detection in a Network of Similar Machines Using Clustering Approach*, (University of Cincinnati, 2012)
31. D. Siegel, W. Zhao, E. Lapira, M. AbuAli, J. Lee, A comparative study on vibration-based condition monitoring algorithms for wind turbine drive trains. Wind Energy **17**(5), 695–714 (2014)
32. A. Jabłoński, T. Barszcz, M. Bielecka, Automatic validation of vibration signals in wind farm distributed monitoring systems. Measurement **44**(10), 1954–1967 (2011)
33. General Electric, Predix. https://www.ge.com/digital/predix
34. National Instruments, Big Analog Data™ Solutions. http://www.ni.com/white-paper/14667/en/
35. Center for Intelligent Maintenance Systems, Development of Smart Prognostics Agents (WATCHDOG AGENT®). http://www.imscenter.net/front-page/Resources/WD.pdf
36. National Instruments, Watchdog Agent™ Prognostics Toolkit for LabVIEW—IMS Center. http://sine.ni.com/nips/cds/view/p/lang/en/nid/210191
37. Applied Materials, Applied TechEdge™ Prizm™. http://www.appliedmaterials.com/media/documents/techedge-prizm-overview
38. CANRIG, RigWatch® Instrumentation and Equipment Condition Monitoring. http://www.canrigdrillingtechnology.com/rigwatch.php
39. Y. Chen, J. Lee, *Data Quality Assessment Methodology for Improved Prognostics Modeling* (University of CIncinnati, Cincinnati, OH, 2012)

Prognostics and Health Management of Wind Turbines—Current Status and Future Opportunities

Shuangwen Sheng

Abstract The global wind industry has seen tremendous growth during the past two decades. However, the industry is challenged by premature component failures, which lead to increased turbine downtime and subsequently, cost of energy for wind power. To mitigate the impacts from these failures, the wind industry has been exploring various areas for improvements ranging from product design, new materials or lubricants, to operation and maintenance (O&M) practices. Condition-based maintenance or prognostics and health management (PHM) has been explored as one enabling technology for improving O&M practices. This chapter provides a brief overview of wind turbine PHM with a focus on operational data mining and condition monitoring of drivetrains. Some future research and development opportunities in wind turbine PHM are also briefly discussed.

Keywords PHM · Wind turbine · Diagnostics · Prognostics · Operation and maintenance

1 Introduction

Global cumulative wind installation capacity reached 430 gigawatts (GW) by the end of 2015 [1]. However, the industry still experiences premature turbine component failures, led by gearboxes, leading to increased operation and maintenance (O&M) costs and subsequently, the cost of energy for wind power. The cost of failures can become much higher for offshore wind plants. Based on European experiences, on average, the availability of offshore wind plants is about 7% lower than land-based plants, which have an averaged availability of about 98% [2], and the O&M costs for an offshore wind plant is twice the cost of a land-based plant [3]. There is a clear need for the wind industry to improve reliability and reduce O&M costs, especially when turbines are installed offshore.

S. Sheng (✉)
National Renewable Energy Laboratory, Golden, CO, USA
e-mail: shuangwen.sheng@nrel.gov

S. Ekwaro-Osire et al. (eds.), *Probabilistic Prognostics and Health Management of Energy Systems*, DOI 10.1007/978-3-319-55852-3_3

The wind industry has tried to improve reliability and reduce O&M costs from a wide range of perspectives, such as testing and design [4, 5], tribology and lubricants [6, 7], and O&M [8, 9]. Once turbines are manufactured and installed in a wind plant, the main opportunity for cost reduction lies in improvement of O&M practices. Condition-based maintenance (CBM), or prognostics and health management (PHM), is one enabling technology that the wind industry has investigated for O&M improvement. Condition monitoring is often used interchangeably with CBM or PHM but in this chapter, condition monitoring is treated as a set of various techniques that focus on data sensing, signal processing, fault detection, diagnosis and prognostics; however, CBM or PHM is treated as a framework containing all elements of condition monitoring, as listed earlier, and also adding an O&M decision supporting piece. In the remaining sections of this chapter, PHM, which is typically thought to encompass CBM, will be used for simplicity of discussion.

PHM is defined in [10] as an approach to system life-cycle support that seeks to reduce or eliminate inspections and time-based maintenance through accurate monitoring, incipient fault detection and diagnosis, and prediction of impending faults. A PHM framework or process typically involves activities that are classified in [11] into seven layers: data acquisition, data processing, condition assessment (detection), diagnostics (identification), prognostics, decision support, and human-machine interface. These layers can be grouped into three blocks according to their functions as observation (data acquisition and processing layers), modeling and analysis (condition assessment, diagnostics, and prognostics layers), and decision (decision support and human-machine interface). The benefits of PHM highlighted in [12] include increased productivity, reduced downtime, reduced number and severity of failures (particularly unanticipated failures), optimized operating performance, extended operating periods between maintenance, reduced unnecessary planned maintenance, and reduced life-cycle cost. When root causes for a certain failure mode are identified, improvements in operation and product design can potentially be accomplished. PHM has been successfully applied in fuel cell systems, nuclear power plants, aviation applications, and electronics. The benefits of PHM, as seen in other applications, can help greatly improve O&M practices in the wind industry if it is harnessed to the full potential. The original onset of PHM for wind turbines was the 1980s, when turbines were equipped with supervisory control and data acquisition (SCADA) systems; however, with dedicated add-on instrumentation, PHM emerged about two decades later when such systems became economically beneficial for utility-scale wind turbines. Because turbine SCADA data mining for PHM purposes had only commenced a few years ago and the deployment of drivetrain condition monitoring systems on utility-scale wind turbines is still increasing, it is reasonable to state that the PHM of wind turbines is largely at the nascent stage.

So far, the development of PHM for wind turbines has focused on data acquisition, data processing, condition assessment, and diagnostics. Depending on the specific technologies employed, some are more mature than the others. The O&M decision supporting piece within the PHM framework still appears to be at the research and development (R&D) stage, with the majority of the work on O&M

strategy optimization being done in Europe and focusing on offshore applications [13, 14]. There is a much higher value proposition for PHM when turbines are installed offshore because of the challenges with accessibility and the availability of maintenance vessels. With the evolution of new technologies such as big data, cloud computing, and Internet of Things [15], there are opportunities to implement the entire PHM framework for wind in a more cost-effective manner to help reduce O&M costs—and, subsequently, the cost of energy for wind power without subsidies—to a level competitive with traditional power generations.

In a broad sense, PHM of wind turbines can target almost all its major assemblies, including the rotor, drivetrain, tower, foundation, and even subsea cables. Providing these assemblies are instrumented with appropriate sensors as part of the turbine or substation SCADA systems, potential issues with them are likely to be discovered through SCADA data mining, which is often referred to as performance monitoring. However, this type of analysis may be unable to identify specific issues down to the component level unless the sensor location is unique to the faulted component. This limitation can often be overcome by deploying dedicated condition monitoring systems on turbines. The vast amount of SCADA data generated at a wind plant is the very first resource the wind industry can explore to improve O&M practices. As a result of high downtime and replacement costs, the main focus of PHM in the wind industry has been on drivetrains, not only through SCADA data mining but also dedicated condition monitoring technologies. Given the number of components in a typical wind turbine and the diverse failure modes these components may experience, if economically feasible, it is beneficial to integrate a few technologies for PHM of wind turbines to cover a wider range of failure modes and take advantage of the strengths of each technology.

There is substantial R&D and deployment potential for PHM technologies in the wind industry to help reduce O&M costs and increase the competitiveness of wind power. This chapter provides a brief overview of typical practices of PHM in wind turbines covering both SCADA data mining and dedicated condition monitoring with a focus on drivetrains. The chapter also highlights some future R&D opportunities in PHM of wind turbines.

2 Typical Practices in Utility-Scale Wind Turbines

This section discusses typical PHM practices in commercially operated utility-scale wind turbines. It focuses on drivetrains and includes subsections on SCADA data mining and condition monitoring. Both SCADA data mining and condition monitoring can be integrated in the PHM framework serving the observation and modeling and analysis functions. The drivetrains discussed herein are for geared wind turbines and are considered to include the main shaft bearings, the gearboxes, and the generators.

2.1 SCADA Data Mining

Modern utility-scale wind plants are normally equipped with SCADA systems, which collect various types of data from the turbines and send them to a centralized computer for monitoring and control purposes. The parameters collected by a wind plant SCADA system are typically 10-minute averages and can be classified into [16]:

- Wind parameters (e.g., wind speed, deviation)
- Performance parameters (e.g., power output, rotor speed, and blade pitch angle)
- Vibration parameters (e.g., tower acceleration and drivetrain acceleration)
- Temperature parameters (e.g., bearing temperature).

The SCADA system also provides information on turbine states (e.g. operation, service, and alarm) [17]. For PHM of wind turbines, there are typically two types of analyses based on SCADA data:

- Modeling the correlations among different parameters (e.g. power and wind, for normal operational states) and using these models to identify abnormal turbine conditions
- Conducting statistical analysis of events (e.g., status codes) experienced by turbines [18].

A diagram introduced in [19] is shown in Fig. 1 as a sample schematic for a typical PHM of wind turbine practice based on SCADA data mining. The PHM is meant for real-time monitoring of wind turbines and built on an offline model, which leads to the *Turbine Model* as shown in the figure. The *Turbine Model* is developed to reflect the relationship between model inputs (e.g., wind speed and air density) and output (i.e., turbine active power as used in Fig. 1) under fault-free conditions. The algorithm starts with feeding the most recent time step SCADA data history to the *Data Filtering* block, which selects only those samples meeting a number of status requirements, such as blade pitch limits. The status requirement

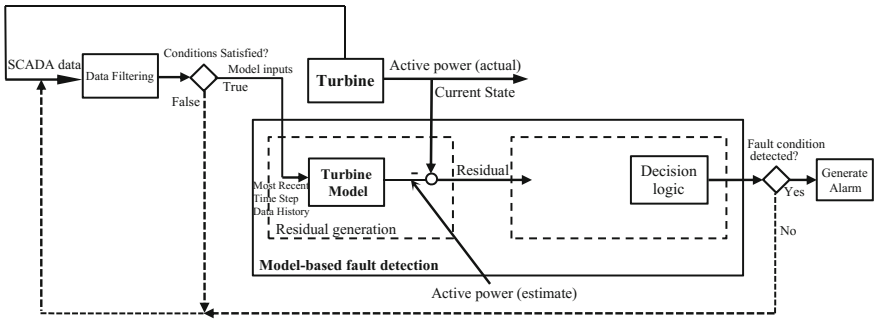
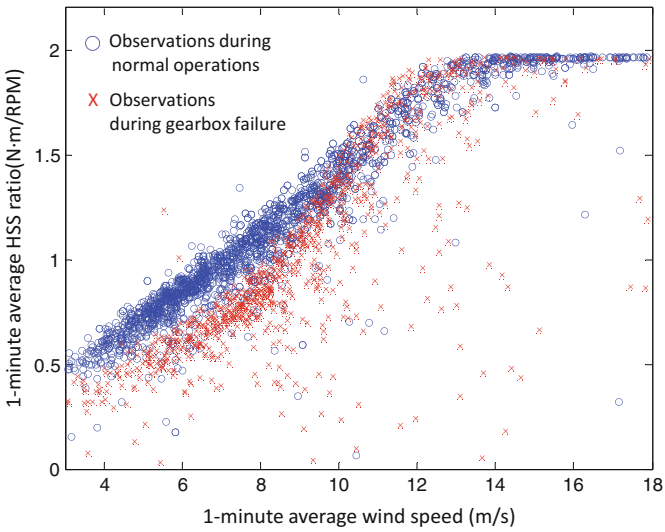


Fig. 1 A sample schematic for PHM of wind turbines based on SCADA data mining (reproduced from [19] with permission)

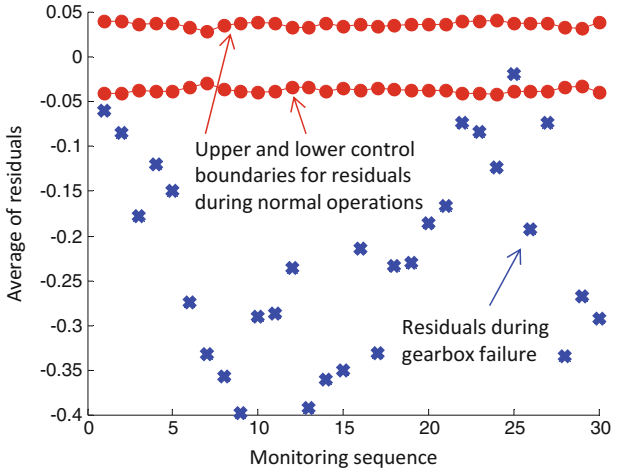
check is implemented by the *Conditions Satisfied* step in the diagram, and the outputs of this step are passed as inputs to the *Turbine Model*. The *Model* generates a predicted mean power, as the *Model* output, for the current state. A residual or difference between the predicted power output from the model and the turbine actual power at the current state is generated; this is termed the *residual generation stage* in Fig. 1. Throughout the period when the turbines are being monitored, the residual values generated at intervals of the SCADA time step, typically l0 min, are combined to form a time series, which is then analyzed at the *residual evaluation stage*, including *residual processing* and *decision logic* (two steps). The *residual processing* step identifies and discards the portion of residuals with high uncertainties. The *decision logic* step determines whether certain characteristics of the processed residuals are met to indicate a fault condition. When any of the fault conditions are met, an alarm is generated. Otherwise, the algorithm continues to iterate.

To illlustrate SCADA data-mining-based PHM for wind turbines, a case study conducted in [20] using data collected from the two-bladed Controls Advanced Research Turbine (CART2) at the U.S. Department of Energy's National Renewable Energy Laboratory (NREL) is presented here. The CART2, which is rated at 600 kilowatts (kW), is used for dedicated wind turbine control research, together with a three-bladed Controls Advanced Research Turbine. The data were collected at 100 hertz (Hz), which is a much higher resolution than a typical commercial SCADA system because of the need for other research activities. The data include measurements taken when a gearbox failure, i.e., gear tooth fracture, occurred. The modeling input was a 1-minute average wind speed in meters/second (m/s), a higher resolution than typical the 10-minute average SCADA data on commercially operated wind turbines, and the output was the high-speed shaft (HSS) ratio, which is defined as the ratio between HSS torque in Newton meters ($N \cdot m$) and the HSS speed in revolutions per mintue (rpm). Figure 2a shows the observations between the chosen inputs and outputs during normal operations and the gearbox failure. It can be seen that after the gearbox damage occurred, the HSS ratios were lower at the low-to-medium wind speed range (i.e., 3–10 m/s) than those during normal operations. A model for fault-free condition was developed using the data collected during normal operations, and it was used to generate a residual control chart in which decision boundaries were defined as three-sigma of the averaged residuals. The HSS ratios calculated based on data collected during gearbox failure are compared with those estimated by the baseline model, and the corresponding residuals are illustrated in Fig. 2b, along with the residual control boundaries. The monitoring sequence in the horizonal axis of Fig. 2b refers to the data pairs formed by a certain input wind speed and its corresponding HSS ratio. It can be seen that most of the residuals during gearbox failure fall outside of the lower control limits, indicating an abnormal turbine condition.

To summarize, the typical benefits of SCADA data mining-based wind turbine PHM include:



(a) Observations from CART2 during normal operations and gearbox failure.



(b) Most residuals during gearbox failure fall outside of control limits established under normal operations.

Fig. 2 A case study illustrating PHM of wind turbines for gearbox failure detection using SCADA data mining (adapted from [20] with permission)

- Readily available data with no additional investments in dedicated condition monitoring instruments
- Easy identification of abnormal turbine conditions by looking at key performance parameters or status codes, triggering further inspections

- Use of temperature as a reliable condition indicator for some turbine components, such as main bearings, generator bearings, or gearbox bearings.

The main limitations of this approach include:

- Difficulties in immediate detection of exact damaged components (e.g., bearings or gears inside gearboxes)
- Possibility of insufficient lead time from temperature-only measurements to save monitored components from irreparable or collateral damage
- Possibility or presence of false alarms caused by varying loads experienced by wind turbines
- Inability to meet full turbine condition monitoring or wind plant PHM needs of accurate fault detection and diagnostics [21].

2.2 Condition Monitoring

Various dedicated condition monitoring technologies can be deployed on wind turbines and these may include [22] vibration analysis, acoustic measurement, oil monitoring, thermography, and visual inspection. These technologies can be classified into two categories: continuous and periodic techniques. PHM of wind turbines based on dedicated continuous condition monitoring techniques is widely recognized by the industry as technically beneficial but economically debatable, especially for aging land-based turbines. For offshore and newly developed wind plants, deploying one continuous condition monitoring technology has almost become a default option. The increasing deployment level of continuous condition monitoring technologies on wind turbines provides the industry a significant opportunity to harness the benefits of PHM to its full potential and reduce O&M costs.

An abstract description of typical continuous condition monitoring systems deployed in wind turbines is illustrated in Fig. 3. These technologies normally target a few mission-critical and cost-prohibitive assemblies of wind turbines, such as gearboxes. The physical measurement varies depending on the specific technology used (e.g., the popular sensors for vibration analysis are accelerometers). Data acquisition is normally implemented by a data acquisition unit (DAU) with microprocessor-based software. The DAU collects signals generated by the physical measurement sensors and converts them into data that can be transmitted to a remote computer for analysis. Some data-buffering storage is typically provided at the DAU. The data collected by the DAU can be transmitted to the remote computer via cabled or wireless connections. The following few steps (i.e., signal processing, fault detection and diagnostics, and prognostics) are normally implemented on the remote computer. The computer is equipped with a dedicated software package that has its own database, a user-friendly interface, and typically a hierarchical architecture representing the monitored wind plants, turbines, assemblies, and

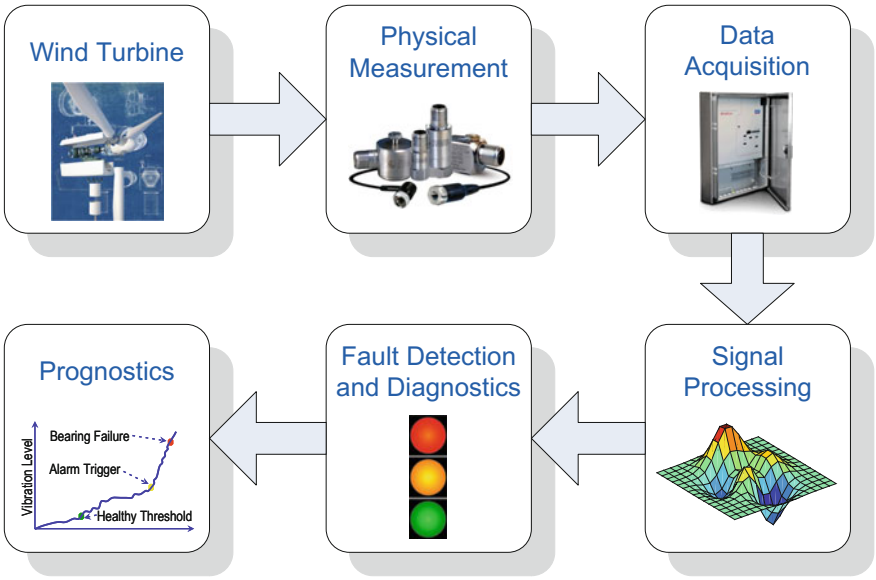


Fig. 3 An abstract description of a typical continuous wind turbine condition monitoring system. (Photos courtesy of: wind turbine [*upper left*], Joshua Bauer, NREL 500057; physical measurement [*upper middle*], IMI Sensors, a division of PCB Piezotronics, Inc.; data acquisition [*upper right*], SKF)

components. The computer also typically provides much more data storage than the DAU does. *Signal processing* may involve data cleaning, filtering, and feature extraction, which drives values from raw data to make it more informative, nonredundant, and normally to a reduced dimension. *Fault detection and diagnostics* can provide information on whether there is a fault, where the fault is, and severe the fault is in the monitored turbine components. *Prognostics* targets fault progression and estimation of remaining useful life [23, 24], which provide critical inputs for O&M optimization. The fault detection, diagnostics, and prognostics tasks normally require human interventions whereas the other tasks can be automatically performed.

The measurements obtained by continuous wind turbine condition monitoring systems may include strains, accelerations, acoustic emissions, oil debris counts, oil condition measurements, electric currents, and voltages. Most of these signals except oil-related measurements can be processed to derive certain features or condition indicators that are useful for subsequent fault diagnostics and prognostics. The processing may include certain preprocessing of data, feature extraction in time-domain (e.g., peak, root mean square values), frequency-domain (e.g., gear-meshing frequencies and sidebands and bearing fault frequencies), and joint time-frequency-domain analyses (via wavelet or short-time Fourier transforms). For oil-related measurements, the sensor outputs normally can be used without the complex processing listed earlier. Fault detection and diagnostics are typically

conducted through various pattern classification, clustering, or regression analysis algorithms. Specifically, it may involve trending the derived features or condition indicators, and the rates of change of these variables. It may also involve identifying the appearance of new frequencies that correspond to certain component faults, the evidence of abnormal modulations of certain operating frequencies, or the violation of thresholds set for certain features or condition indicators. For prognostics, there are two main methods: a data-driven approach, using pattern recognition or machine-learning techniques (e.g., autoregressive models and neural networks [25]), and empirical or physics-based modeling methods, using physical understanding of the governing mechanism of the modeled phenomena, such as crack propagation models developed via fracture mechanics [26]. Among fault detection, diagnostics, and prognostics, detection and diagnostics are normally provided by most commercial condition monitoring systems with prognostics remaining mainly at the R&D or tryout stage.

Among the different continuous monitoring techniques [27], vibration analysis and oil debris monitoring are predominantly used on wind turbines. The gearbox, main bearing, and generator of a wind turbine have been typically monitored due to their cost-prohibitive replacement. It is beneficial to deploy both vibration and oil debris analysis techniques to cover the broad, complex, and diverse failure modes that wind turbine drivetrains experience. In reality, the most commercially operational wind turbines either have one of these two technologies or none. Typical practices with vibration analysis and oil debris monitoring in wind turbines are provided in the next two sections, which focus on commercial solutions for fault detection and diagnostics of drivetrains.

2.2.1 Vibration Analysis

For a wind turbine, vibration analysis is typically used to monitor the drivetrain. It consists of several sensors (typically a few accelerometers and a tachometer), a DAU located in the turbine nacelle, and a data server located at the wind plant or a remote monitoring center. The tachometer can be a dedicated channel for the condition monitoring system or it can be shared with the turbine controller. The communication between the DAU and the data server located at the wind plant can be through Ethernet or fiber optic cables. If no data server is set up at the local wind plant, the DAU normally can be configured to wirelessly transmit the test data to a server located in the remote monitoring center, which can be anywhere around the globe. The data server normally hosts a software package, which is used to review and analyze the collected data, present analysis results, and streamline both raw and processed condition monitoring data into a database. One wind plant, typically consisting of hundreds of turbines, can be monitored by one condition monitoring software package located at the server if there is no problem with communications between the condition monitoring DAUs and the data server.

The main differences among various vibration analysis systems are the number of sensors, measurement locations, and analysis algorithms used, whereas almost all

commercial solutions use accelerometers as their main physical measurement devices. Typically, one to two accelerometers are mounted on the main bearing to measure either radial or axial acceleration. Three to four accelerometers are installed on the gearbox to measure the radial accelerations at different gearbox stages (e.g., planetary and parallel stages). One to two accelerometers are installed on the generator to measure either drive or nondrive end radial acceleration.

For analysis algorithms, various vibration condition monitoring systems may have different approaches covering time and frequency, or joint time-frequency domains [28]. Often, the time domain parameters are used to track the trend of overall vibration level over time at a specific sensor location and to detect faults that have occurred to the monitored component. Some triggering mechanisms can be set up based on the trending of these time domain parameters to enable discrete snapshots of detailed frequency analysis on raw or preprocessed signals. Based on these snapshots, detailed diagnostics of the monitored component can be conducted. Based on data processed by frequency domain analysis algorithms, some statistical parameters can be calculated, such as the amplitude of characteristic frequencies for gears (e.g., meshing frequency) and bearings (e.g., ball-passing frequency), and these parameters can be trended over time in a fashion similar to those calculated based on the raw data for both fault detection and diagnostics purposes. The challenge with traditional frequency domain spectrum analysis is its in effectiveness on nonstationary transient signals, which can be better handled by joint time-frequency domain methods. Modern vibration analysis systems used on wind turbines typically have signal processing methods in all these domains to take advantage of their strengths and improve system performance.

2.2.2 Oil Debris Monitoring

Continuous oil debris monitoring in wind turbines is typically applied to the gearbox, as it is normally the only oil-lubricated assembly in the drivetrain [8]. The main function of this type of sensor is to measure debris shed by gears and bearings and circulated with the lubrication oil. Sensing is typically done using a magnetic field-based principle [29]. When lubrication oil circulates through these sensors, total debris counts, including ferrous and nonferrous types, are recorded. Often, these sensors can estimate the debris sizes and separate them according to different size bins.

One main difference among various oil debris monitoring sensors has to do with the sensor-mounting location, which can be either inline within the main lubrication system or online within a side-stream lubrication system that has a slower flow rate. The sensor-mounting location is determined by the bore size of the sensor, the pressure ranges, and the flow rate that the sensor can handle. Another difference has to do with the minimum detectable debris size. For inline sensors, the size is at the level of hundred microns and a few times bigger than the minimum size detectable by online sensors.

The outputs of the oil debris monitoring sensors can be viewed using dedicated software packages provided by the sensor suppliers, the software platform of a vibration analysis system that can accommodate the outputs from these sensors, or a website, which is typically managed by the sensor suppliers. With some postprocessing algorithms, the debris generation rates [30] can be trended and examined, in addition to the total debris counts, to identify potential gear or bearing faults.

This subsection thus far has focused on continuous condition monitoring technologies. Given that offline oil sample analysis has been a fairly standard practice in the wind industry, it is briefly discussed here. For offline oil sample analysis, an oil sample from a gearbox lubrication system is normally taken at a 6-month interval and sent to a dedicated laboratory for analysis. However, if the continuous oil debris monitoring sensors reveal abnormal conditions, it is better to conduct spot oil sample analyses to help identify failures in progress. The parameters sought in an oil sample analysis typically include particle counts, water content, total acid number, viscosity, and sometimes particle element identification. An analyst at the laboratory reviews the testing results and provides maintenance recommendations to the owner or operator of the test turbine based on limits of metal content set by using historical data collected from similar wind turbine gearboxes. The main benefits of offline oil sample analysis include examining parameters not covered by continuous oil debris monitoring sensors, especially those reflecting oil condition, identifying failure sources through elemental analysis of debris, and enabling root cause analysis for some failures.

2.2.3 Discussions

To summarize, the typical benefits of continuous condition monitoring-based PHM for wind turbines include:

- Detection of turbine high-frequency dynamics that is not achievable with a typical SCADA system via dedicated vibration measurements; this can normally help isolate damaged components
- Unique insights on gearbox oil and component condition gained through oil debris sensors whose results are relatively easy to interpret, or periodic oil sample analysis which can help pinpoint failed gearbox components and assist failure root cause analysis
- Coverage of more failure modes occurring in wind turbines than identified through SCADA data mining.

The limitations of this approach include:

- Additional investment in dedicated instrumentation, monitoring service, or resources for data analysis and interpretation of results
- Challenges with vibration analysis for low-speed-stage turbine components [31]
- Low effectiveness of oil debris monitoring for pinpointing damaged components.

Given the diverse and complex failure modes seen in wind turbine components, an approach that integrates various technologies is recommended, especially one that starts with an initial mining of the SCADA data and then incorporates recommendations given by dedicated condition monitoring technologies, such as vibration analysis or oil debris monitoring.

3 Future R&D Opportunities

Despite the fact that some elements within the framework of PHM of wind turbines (e.g., SCADA data mining and various condition monitoring techniques that are increasingly exploited by the wind industry), plenty of R&D opportunities to realize the full potential of PHM of wind turbines still exist. This section discusses some current R&D activities and future R&D opportunities in PHM of wind turbines according to three areas: data acquisition, signal processing and modeling (i.e., data processing, condition assessment, diagnostics, and prognostics), and O&M (i.e., decision support and human-machine interface).

In terms of data acquisition, some R&D activities are currently carried out by both industrial solution providers and research institutes. A few examples [32] include a shock pulse method for main shaft bearing condition monitoring, gearbox filter element analysis to complement traditional oil sample analysis, and electric signature analysis (currently evaluated only on a test rig or small wind turbines). These approaches could be targeted as potential opportunities for future R&D work. In addition, future R&D efforts could focus on new technologies that are either complementary or superior to the popular sensing solutions for drivetrain condition monitoring or unique and beneficial to other mission-critical and cost-sensitive components in wind turbines. When turbines are installed offshore, special sensing techniques for undersea cables and foundation may be needed.

Along the lines of signal processing and modeling, some current R&D activities [33] include time-frequency analysis based on wavelet transform, Wigner-Ville distribution, or empirical mode decomposition, and data-driven modeling based on neural networks, genetic programming, or regression analysis. Most of these R&D efforts are conducted by research institutes, and they are academically very attractive but often computationally expensive and hard to implement in the field. In terms of modeling effort, validation presents a major challenge with the need for long-term data collection and the lack of publicly available data. Future R&D work may focus on increased use of SCADA data, improved accuracy and certainty of diagnostic decisions including severity-level evaluations, and reliable and accurate prognostics based on performance monitoring, usage monitoring, and load prediction to enable estimation of the remaining useful life of turbine components.

R&D activities on wind plant O&M have focused on offshore wind plants. The reason may be that owners and operators of land-based wind plants are challenged

more by the availability of spare parts and qualified technicians; also, they have not seen many obvious benefits from optimized O&M practices. O&M optimization is more attractive to owners and operators of offshore wind plants than it is to owners and operators of land-based plants because of the high value proposition, which is mainly caused by even lower accessibility and additional logistics and scheduling complexities. A few example R&D activities on offshore wind plants include time-domain Monte Carlo simulation to determine the most cost-effective approach to allocating O&M resources considering environmental conditions, transportation systems, failures, and repairs [34], and Bayesian theory for optimal planning of inspections and maintenance based on a single wind turbine and component considering inspections, repairs, and loss of production [35]. Future R&D efforts in O&M need to focus on a fusion of various data streams or models (e.g., weather forecasting) to optimize O&M practices that can reduce loads and extend the life of turbine components, offer convincing evidence of additional benefits to both land-based and offshore wind plants, and can help develop proper reasoning or expert systems that are able to automate data interpretation and deliver actionable maintenance recommendations.

In addition, a few R&D opportunities span across two or three of the above categories. One area is to conduct R&D work to handle uncertainty within the PHM framework, including uncertainty representation and interpretation, quantification, propagation, and management [36]. For offshore wind plants, the structural load prediction, which is needed for component life estimation, becomes more complex as both wave and wind influences exist. Some novel sensing, sensor integration, or modeling methods may need to be developed. Another area is reliability-centered maintenance, which can target the entire wind plant and recommend the appropriate maintenance for different failures seen in turbine components at the right time. The maintenance of a typical wind plant in the foreseeable future has to combine different strategies, including reactive, preventative, predictive, and proactive measures owing to the number of turbines, the diversity of components on the turbines, and the variation in their failure modes and mission criticality. Yet, another area of interest is root cause analysis. Although very detailed and thorough root cause analysis for all turbine failures may not be economically feasible, it is recommended for frequent failures, as the findings can potentially help improve the turbine operation, control strategy, and even component design.

The PHM of wind turbines can take advantage of emerging technologies (e.g., big data, cloud computing, and the Internet of Things) to become more effective and attractive yet economical. With further R&D in PHM of wind turbines and its relevant areas, and with gradual acceptance by the industry, the technology will help increase the competitiveness of wind power by reducing its O&M costs and subsequently the cost of energy.

Acknowledgements This work was supported by the U.S. Department of Energy under Contract No. DE-AC36-08GO28308 with the National Renewable Energy Laboratory. Funding for the work was provided by the DOE Office of Energy Efficiency and Renewable Energy, Wind and

Water Power Technologies Office. The author would also like to acknowledge the NREL condition monitoring and O&M research partners for their support.

References

1. Global Energy Research Council, *Global Wind Report*, 2015
2. P. Tavner, *Offshore Wind Turbine Reliability* (Supergen Wind Training, Manchester, UK, 2011)
3. International Renewable Energy Agency, *Renewable Energy Technologies: Cost Analysis Series, Volume 1: Power Sector, Wind Power*, May, 2012
4. Y. Guo, J. Keller, W. LaCava, Planetary gear load sharing of wind turbine drivetrains subjected to non-torque loads. Wind Energy **18**(4), 757–768 (2015)
5. Y. Guo, R. Bergua, J. van Dam, J. Jove, J. Campbell, Improved wind turbine drivetrain designs to minimize the impacts of non-torque loads. Wind Energy **18**(12), 2199–2222 (2015)
6. A. Erdemir, A. Greco, J. Keller, S. Sheng, Material wear and fatigue in wind turbine systems. Wear **302**, 1583–1591 (2013)
7. A. Greco, K. Mistry, V. Sista, O. Eryilmaz, A. Erdemir, Friction and wear behavior of boron based surface treatment and nano-particle lubricant additives for wind turbine gearbox applications. Wear **271**, 1754–1760 (2011)
8. S. Sheng, Monitoring of wind turbine gearbox condition through oil and wear debris analysis: A full-scale testing perspective. Tribol. Trans. **59**(1), 149–162 (2016)
9. E. Byon, L. Ntaimo, Y. Ding, Optimal maintenance strategies for wind turbine systems under stochastic weather conditions. IEEE Trans. Reliab. **59**(2), 393–404 (2010)
10. Z.S. Chen, Y.M. Yang, H. Zheng, A technical framework and roadmap of embedded diagnostics and prognostics for complex mechanical systems in prognostics and health management systems. IEEE Trans. Reliab. **61**(2), 314–322 (2012)
11. M. Jouin, R. Gouriveau, D. Hissel, M.C. Péra, N. Zerhouni, Prognostics and health management of PEMFC—State of the art and remaining challenges. Int. J. Hydrog. Energy **38** (35), 15307–15317 (2013)
12. J. Coble, R. Ramuhalli, L. Bond, J.W. Hines, B. Upadhyaya, A review of prognostics and health management applications in nuclear power plants. Int. J. Progn. Heal. Manag. **6** (2015)
13. P. Joschko, A.H. Widok, S. Appel, S. Greiner, H. Albers, B. Page, Modeling and simulation of offshore wind farm O&M processes. Environ. Impact Assess. Rev. **52**, 31–39 (2015)
14. M. Scheu, D. Matha, M. Hofmann, M. Muskulus, Maintenance strategies for large offshore wind farms. Energy Procedia **24**, 281–288 (2012)
15. C. Perera, C.H. Liu, S. Jayawardena, M. Chen, A survey on Internet of Things from industrial market perspective. IEEE Access. **2**, 1660–1679 (2014)
16. K.S. Wang, V.S. Sharma, Z.Y. Zhang, SCADA data based condition monitoring of wind turbines. Adv. Manuf. **2**, 61–69 (2014)
17. C. Kaidis, B. Uzunoglu, F. Amoiralis, Wind turbine reliability estimation for different assemblies and failure severity categories. IET Renew. Power Gener. **9**(8), 892–899 (2015)
18. F. Castellani, A. Garinei, L. Terzi, D. Astolfi, M. Moretti, A. Lombardi, A new data mining approach for power performance verification of an onshore wind farm. Diagnostyka **14**(4), 35–42 (2013)

19. S. Butler, J. Ringwood, F. O'Connor, Exploiting SCADA system data for wind turbine performance monitoring, in *Proceedings of the Conference on Control and Fault-Tolerant Systems*, *Nice, France*, 2013
20. N. Yampikulsakul, E. Byon, S. Huang, S. Sheng, M. Yu, Condition monitoring of wind power system with nonparametric regression analysis. IEEE Trans. Energy Convers. **29**(2), 288–299 (2014)
21. W. Yang, P.J. Tavner, C.J. Crabtree, Y. Feng, Y. Qiu, Wind turbine condition monitoring: technical and commercial challenges. Wind Energy **17**(5), 673–693 (2014)
22. T.W. Verbruggen, Wind *Turbine Operation and Maintenance Based on Condition Monitoring, Energy Research Center of the Netherlands, Petten, The Netherlands*, 2003
23. X.S. Si, W. Wang, C.H. Hu, D.H. Zhou, Remaining useful life estimation—a review on the statistical data driven approaches. Eur. J. Oper. Res. **213**(1), 1–14 (2011)
24. K. Medjaher, D.A. Tobon-Mejia, N. Zerhouni, Remaining useful life estimation of critical components with application to bearings. IEEE Trans. Reliab. **61**(2), 292–302 (2012)
25. P. Baraldi, M. Compare, S. Sauco, E. Zio, Ensemble neural network-based particle filtering for prognostics. Mech. Syst. Signal Process. **41**, 288–300 (2013)
26. F. Zhao, Z. Tian, Y. Zeng, Uncertainty quantification in gear remaining useful life prediction through an integrated prognostics method. IEEE Trans. Reliab. **62**(1), 146–159 (2013)
27. S. Sheng, P. Veers, Wind turbine drivetrain condition monitoring—an overview, in *Proceedings of the Mechanical Failures Prevention Group: Applied Systems Health Management Conference*, *Virginia Beach, Virginia*, 2011
28. Y. Lu, J. Tang, H. Luo, Wind turbine gearbox fault detection using multiple sensors with features level data fusion. J. Eng. Gas Turbines Power **134.4**, 042501-1–8 (2012)
29. R. Dupuis, Application of oil debris monitoring for wind turbine gearbox prognostics and health management, in *Proceedings of the Prognostics and Health Management Society Annual Conference, Portland, Oregon*, 2010
30. S. Sheng, Investigation of various condition monitoring techniques based on a damaged wind turbine gearbox, in *Proceedings of the 8th International Workshop on Structural Health Monitoring, Stanford, California,* 2011
31. D. Coronado, K. Fischer, Condition monitoring of wind turbines: state of the art, user experience and recommendations, in *Fraunhofer Institute for Wind Energy and Energy System Technology IWES Northwest, Bremerhaven, Germany*, 2015
32. S. Sheng, *Prognostics and health management of wind turbines: Current status and future opportunities, presented at the Probabilistic Prognostics and Health Management of Energy Systems Workshop, Ilha Solteira, Brazil*, 14–15 Dec 2015
33. S. Sheng, *Improving component reliability through performance and condition monitoring data analysis, presented at Wind Farm Data Management & Analysis North America, Houston, Texas*, 25–26 Mar 2015
34. Y. Dalgic, I. Lazakis, I. Dinwoodie, D. McMillan, M. Revie, Advanced logistics planning for offshore wind farm operation and maintenance activities. Ocean Eng. **101**, 211–226 (2015)
35. J.J. Nielsen, J.D. Sørensen, On risk-based operation and maintenance of offshore wind turbine components. Reliab. Eng. Syst. Saf. **96**(1), 218–229 (2011)
36. S. Sankararaman, K. Goebel, Uncertainty in prognostics and systems health management. Int. J. Progn. Heal. Manag. **6** (2015)

Overview on Gear Health Prognostics

Fuqiong Zhao, Zhigang Tian and Yong Zeng

Abstract This chapter is dedicated to an overview of prognostics methods for gear health management. By noticing that most prognostic methods are application dependent and new methods keep emerging, this study is necessary for providing the latest status of prognostics capability specific to gears. The reviewed frameworks and/or methods are grouped into data-driven, physics-based and integrated ones. Their respective merits and drawbacks are outlined. The opportunities and challenges are also discussed for future research.

Keywords Gears · Prognostics · Condition monitoring · Failure mode · Remaining useful life prediction · Vibration analysis

1 Introduction

Gears are commonly seen components and are widely used in machinery and equipment for heavy-duty tasks. When transmitting power, gears typically need to carry a heavy load on their teeth, which makes gears prone to fatigue fracture especially at locations where the load is applied and where there exists large bending stress. The working environment can also accelerate the health deterioration of gears. For example, sliding wear appears and grows due to inadequate lubrication on mating faces, and as a result, material is removed from the gear faces in form of metal particles. These particles are immersed in the oil accelerating the wear process, and tooth geometry changes due to material loss causing overwhelming vibration which could immediately call a halt on the service life of the

F. Zhao (✉) · Z. Tian
Department of Mechanical Engineering, University of Alberta, Edmonton, Canada
e-mail: fuqiongzhao@gmail.com

Y. Zeng
Concordia Institute for Information Systems Engineering,
Concordia University, Montreal, Canada

S. Ekwaro-Osire et al. (eds.), *Probabilistic Prognostics and Health Management of Energy Systems*, DOI 10.1007/978-3-319-55852-3_4

gears. Failures of gears, if unexpected, could cause malfunction of the power transmission subsystem and even serious damage to the whole system.

Because of the important role of gears for ensuring the whole system health and safety, methods to accurately and timely detect gear faults have been a focus that has prevailed in the literature for decades. Except for the sudden failures, there is considerable time for the fault to develop before actual failure occurs, and by noticing this fact, gear remaining useful life (RUL) prediction becomes more attractive to provide extra economic benefits by exploiting gear service life or to better improve safety by adjusting operating conditions. The two aforementioned aspects, fault detection and RUL prediction, are the main tasks in the modern context of equipment prognostics and health management (PHM) [1]. Gear fault detection has received massive research attention and many well-established methods exist [2–5]. In contrast, RUL prediction in PHM is still in its infancy and needs further development for practical applications. Some researchers have done excellent reviews on general prognostics methods [6–13]. However, most of the prognostic methods are application dependent and tailoring a general method into a particular application still needs extra efforts, if possible. In this book chapter, we review prognostics methods that are devised specifically for gears. By giving an overview of the existing frameworks and methods for gear failure prediction, we suggest future directions of gear prognostics methods for better equipment health management.

First, a short introduction to PHM, an advanced framework for modern system health management, is presented. Traditional maintenance strategy after the system is deployed to field is reactive in nature, which simply responds to the fault occurrence. The corrective maintenance actually implicitly allows for outage, and thus is unavoidably accompanied with the high risk of downtime as well as high operational cost. A more intelligent alternative is time-based maintenance, where actions are taken upon failure, while otherwise maintenance is scheduled every optimal time interval [14]. The time-based maintenance can prevent some unexpected failures, but it is still resource-wasteful due to the unexploited service life of some units. It is worth mentioning that failure time distribution of unit population is usually needed to determine the optimal time interval in time-based maintenance. Therefore, it is population oriented which is useful at the stage of product design and setting warranty policy. However, testing and assessment based on population only provide an average performance of interest, and can hardly be used to characterize the performance of an individual. In contrast, at the stage of product deployment, the real concerns for maintenance and operation is indeed the performance of individuals. While addressing such concerns, PHM attracts research interest and efforts from both academia and industry.

The basic idea of PHM is to detect/diagnose faults (diagnostics), use the diagnostic information afterwards to predict failure progression (prognostics) and then to plan maintenance/operation/logistics actions beforehand (maintenance optimization) based on the predicted failure time. In this way, PHM aims to achieve zero unexpected failures, full usage of service life and minimum maintenance cost. One distinct feature of PHM (compared to traditional statistics based life models) is

to use sensors acquiring signals which could be indicative to the system degradation. The framework of PHM opens opportunities for accurate gear life prediction. For example, sensor data (vibration, operational data, etc.) may be utilized to adjust predictive models or to estimate the current state of damage, which could then lead to a better prediction as a result of taking account of the latest health condition. From another perspective, PHM also offers an amenable way for uncertainty quantification when a Bayesian filtering is used for data assimilation. In this chapter, a focus will be given to the cutting-edge prognostics methods for gears within the PHM framework with the traditional ones also covered to better illustrate the comparison and evolution of algorithms.

The remainder of this chapter is organized as follows. In Sect. 2, various gear failure prediction models are reviewed with a focus to demonstrate the state-of-the-art approaches in PHM for gear health prognostics. These approaches are deterministic or stochastic, physics-based or data-driven, and population-oriented or individual oriented. Various methods show their respective strengths and drawbacks. Section 3 will discuss the opportunities and challenges encountered in the course of developing prognostic algorithms, and based on which we suggest areas for future research to improve their efficiency, accuracy and robustness. Conclusions are given in Sect. 4.

2 Gear Health Prognostics Methods

The objective of prognostics is to predict RUL of gears before they fail to meet operation requirements. Owing to a large number of uncertainty sources accompanied by prognostics, it is acknowledged that prognostics should be conducted in a stochastic way. More specifically, one should be able to tell the confidence in the predicted RUL. Apart from that, a good prognostic algorithm is also expected to have a mechanism of uncertainty reduction to increase the confidence for decision-making in maintenance or mission planning [1].

In this section, we will review four categories of prognostic methods for gears. Section 2.1 will review the first one, which is based on Weibull analysis of material rupture before the concept of PHM was even proposed and dates back to 1940s. This method (along with other well-known $s-N$ curve, $\varepsilon-N$ curves and their modified variants) copes with micro-cracks or micro-pits in the damage initiation stage. At the same time, the other three categories are spotted and recognized within the PHM context: the physics-based, the data-driven and the integrated methods. They are invented for damage propagation stage during which the damage grows from a scale that is sufficient to be detected until failure. In Sect. 2.2, we review several physics-based models of damage propagation pertinent to different failure modes of gears. Section 2.3 is dedicated to the counterpart data-driven prognostic methods for gears. Features extraction for prognostic purpose from gearbox condition monitoring data are also reviewed because it is an essential step in developing data-driven methods. Section 2.4 covers the prognostics methods that

combine the physics-of-failure damage propagation model and the condition monitoring data to achieve an adaptive and robust prediction.

2.1 Gear Fatigue Life Statistical Models

Back in 1939, the theoretical work by Weibull [15] developed a fundamental formula to estimate the probability of material rupture at any given stress over a volume. This paper noticed the variation of ultimate material strength and chose to use a distribution function rather than a constant to represent it. Based on this result, a general fatigue life model was proposed by Lundberg and Palmgren [16] in 1947 for rolling-element bearings.

Coy et al. [17] later applied this Lundberg-Palmgren model to surface fatigue life estimation for spur and helical gears as the gear life is reached when pitting appears due to fatigue contact. The output of this model is the gear life at a given survival probability (i.e., reliability) under a given transmitted load, and by feeding the gear mesh contact stress, shear stress and stressed volume. For example, the 90% probability of survival can be written in Eq. (1), where τ_0 is the critical stress, z_0 is the depth of the critical stress, V is the stressed volume, h, c, K_1 are material dependent exponents and e is Weibull slope.

$$L_1 = \left(\frac{K_1 z_0^h}{\tau_0^c V}\right)^{1/e}. \tag{1}$$

To use this model, failure experiments are needed to determine the parameters (exponents and material constant) in the life model; gear mesh contact analysis under certain lubrication condition is also required to obtain the stress undertaken by the material. Thus, this model is built on theories of both material and statistics, so it has the merit of physics-based methods along with the capability to account for the overall uncertainty in population. This model is instrumental as a pioneer for adopting a probabilistic way for gear fatigue life prediction. However, as suggested by the authors, more tests are needed to obtain statistical significance of the experimentally determined parameters. As mentioned before, this model is population-oriented, and it makes no differentiation among individual gears. A large amount of variance in the predicted lifetime is expected with this model.

2.2 Physics-Based Gear Prognostics

Fault initiation and propagation are physical processes that take place in the structure, joint or component made of various types of materials. Researchers naturally intend to understand these processes from physical perspectives, such as

understanding the material property under repetitive thermal and dynamic loading. Unlike the fatigue life model in Sect. 2.1, prognostics in PHM try to understand how the fault propagates or grows with time after an initial fault is detected. The physics-based prognostics method resorts to physical laws that govern the fault propagation process, and if the physics behind fault propagation process is well understood, physics-based methods will give the predictions with the highest accuracy among all the categories. To apply such methods, the first step is to identify the failure mode of interest so that the proper fault progression model is selected.

Gears have several main failure modes: tooth fracture due to crack growth, surface fatigue (pitting/spall) due to rolling contact, and surface wear due to sliding contact. By noticing the scarce publications on pitting progression model for gears, this section will only review the models for the other two failure modes, leaving surface fatigue for a later section when we discuss future opportunities.

2.2.1 Tooth Fracture

As cyclic loading continues during gear mesh, cracks will initiate at tooth fillets subject to maximum bending stress. The propagation of cracks will eventually cause tooth breakage and result in gear failure. Paris' law shown in Eq. (2) is commonly used to describe crack growth with time [18]. It predicts the crack size (a) increment per loading cycle (N).

$$\frac{da}{dN} = C(\Delta K)^m. \tag{2}$$

The important quantity to calculate in Paris' law is the stress intensity factor (SIF), ΔK, which determines the stress distribution near the crack tip in linear elastic fracture mechanics. Many publications are devoted to calculating SIF using different numerical techniques, and interested readers can refer to [19–23] for these techniques in computational fracture mechanics. Many papers applied Paris' law for gear life prediction [24–27]. As a physical law, Paris' equation has many variants to incorporate additional factors that affect crack growth in a gear, such as load ratio, toughness [28], hardness [29], closure retard [30] and random loading [31].

The failure criteria of gears bearing a crack is usually defined by the critical value of crack size or SIF. Hence, SIF calculation is a key to obtain accurate crack propagation prediction. The residual stress due to tooth case hardening was considered in the finite element (FE) models to compute SIF in [32]. Authors in [33, 34] investigated several factors that may influence the crack growth trajectory in the gear tooth, including backup ratio, initial crack location, fillet geometry, rim/web compliance, gear size and pressure angle. Most publications have assumed a constant load applied at a fixed position when calculating SIF. However, in actual gear meshing, the load moves along the tooth, changing in both amplitude and position, so in order to account for moving load during tooth mesh, authors in [35] developed

a quasi-static numerical method to calculate cycles for cracks to propagate to a critical value by breaking the tooth engagement into multiple steps. Even though many factors may affect crack propagation, Paris' law was applied in the absence of any latest information of the crack state in the above-mentioned research neglecting the characteristics of a specific gear in a specific operating condition. In [36] and [37], authors developed the gear RUL prediction system which combined the gear dynamic model, the fracture FE model and the crack estimation algorithm together to achieve the improved prognostics accuracy for a specific gear. The novelty is the incorporation of a module to estimate the current crack size using the measured transmission error [38] or vibration index [39]. As a result, the current health condition is updated before applying Paris' law so that the life prediction will get accurate. However, Paris' law is still applied in a deterministic way and the predicted RUL is a single value with no confidence evaluation.

2.2.2 Sliding Wear

When gear teeth mesh with each other, the tangential velocity is different from two mating teeth surfaces. The relative movement will cause sliding between the two surfaces and as a consequence of the direct asperity contacts, the metal material will be removed from the surface which then defines the sliding wear. Sliding wear will gradually change the tooth geometry (e.g., the tooth thickness becomes thinner), increasing vibrational level of the gearbox and accelerating the formation and growths of other faults. A widely accepted wear model is the Archard's model [40] shown in Eq. (3)

$$\frac{dh}{ds} = kp \tag{3}$$

which describes the wear depth (h) increase at a point on the tooth surface during one unit sliding distance. In Eq. (3), k is the wear coefficient and p is the contact pressure on the mating surface. From this model, it is obvious to conclude that there is no sliding wear at the pitch point of a spur gear because no sliding motion occurs there. It is also worth mentioning that the wear process is highly influenced by the lubrication condition. Most gears run in a partial elastohydrodynamic lubrication (EHL) regime, where the gear teeth moving speed is fast enough to create a film but the film is inadequate to prevent direct asperity contact between two surfaces. Similar to the Paris' law, Archard's model also has its generalizations to account for more factors [41, 42].

Wu and Cheng proposed a sliding wear model for partial-EHL contacts in [43], a model that accounted for many factors including contact pressure, sliding velocity, contact area, thermal desorption and oxidation. Afterwards, this model was later applied to a spur gear wear process where the dynamic loading was considered and the sliding wear volume was calculated in one mesh period. Later, Flodin and Andersson used Archard's model to predict sliding wear on spur and helical gears

[44, 45]. Hertz theory and Winkler's mattress model were applied to calculate the contact pressure on meshing teeth. However, no experiment results were present in their work. Following this work, authors in [46] developed methods that were able to take manufacturing/assembly imperfection and intentional surface modification of gears into wear prediction. Gear wear experiments were also conducted to validate the model. In these methods, the wear coefficient k was treated as a constant obtained from the experiment on some training units. When using this value to predict wear process of other units, there must be some errors because of inevitable uncertainties in material, lubrication and loading condition. To mitigate this drawback, Zhao et al. [47] proposed integrated method for gear wear prediction which will be discussed in later section.

Physics-based methods are generally accurate if the required information is available, but they typically require intensive efforts to build the fault progression model and then to determine the parameters in it. Physical models are usually not available for complex systems or for certain failure mechanisms that is not well understood. Even if the model is available, sometimes the computational resources are too demanding to afford in practice. In addition, physics-based methods are blind to the current health status of the gears, and so any model error can be amplified to an unacceptable level as time proceeds. Furthermore, the deterministic way to treat physical parameters also renders the risk of physics-based methods in predicting fault progression of a specific unit. Last but not least, physics-based methods lack a measure of confidence in the predicted results.

2.3 Data-Driven Gear Prognostics

The fast development in sensor technology provides a large amount of condition monitoring data from which the gear health status can be evaluated and tracked. Data-driven methods are able to extract useful intelligence from huge datasets consisting of sensor signals and/or operational data to achieve desired PHM purposes. The distinct feature of data-driven methods is that the prognostic models are obtained by training on and only on the data, with no input from physics of failure nor assumption of mathematical model for degradation process. The rationale behind the data-driven methods is that, as system performance degrades with usage, its health status can be manifested by or be hidden in the condition monitoring data.

In data-driven methods, before the predictive models are trained, there is a critical step called feature extraction where the features for training are obtained by signal processing and learning techniques. A qualified trending feature should be sensitive to degradation over time (e.g., monotonically increases with damage increases), immune to noise, and robust to changes in environmental/operating conditions. As is well known, vibrational analysis of gears plays an essential role in gear fault diagnostics. There has been a large volume of literature on gear fault detection methods, but traditional features that are used to detect gear faults are not always effective to serve as prognostic indicators because the amplitude of fault

detection features may not be as sensitive to the extent of the fault. The situation is worsened when the gears are subject to non-stationary operating conditions (e.g., time-varying speed and load). Therefore, sophisticated signal processing and machine learning techniques are usually needed for developing good indicators to predict the damage growth trend.

In the remainder of the section, we will review popular data-driven methods for gear health prognostics, including statistical machining learning methods and dynamic systems.

2.3.1 Data-Driven Methods: Statistical Matching Learning Methods

Once a qualified indicator is selected, the damage propagation process can be represented as a time series which is an important tool in failure prediction. The neuro-fuzzy (NF) approach and recurrent neural networks (RNN) are two commonly used techniques for time series prediction. NF combines neural networks (NN) and fuzzy logic to circumvent the drawbacks of NN (i.e. a lack of transparency and a slow training rate). It was also found that in [48] that RNN was a better predictor than feedforward neuro-networks (FNN). Therefore, Wang et al. [49] investigated NF system and RNN for gear prognosis. In the proposed NF systems, the time step span is constant (i.e. a number of previous indicator values are fed into NF systems to predict the value at the next step). The method was evaluated by datasets of various gear failure modes: worn gear, chipped gear, cracked gear and pitting gear. A feature of wavelet amplitude pattern based on the overall residual signal was developed as the prognostic indicator for the first three faults, and a normalized kurtosis based on the overall residual signal was used to track the pitting. The signal processing method to extract such features can be found in [50]. The results have shown that the NF predictor accurately caught the feature trend and tracked it very well whereas RNN failed to adapt itself to the new system dynamics after the fault was initiated. Therefore, the study concluded that a properly trained NF system performs better than RNN in both forecasting accuracy and training efficiency. The author later developed an extended neuro-fuzzy (ENF) network which could achieve more accurate prediction in [51].

Following the same route, Samanta and Nataraj investigated the other two time-series prediction techniques for gear prognostics: adaptive neuro-fuzzy inference system (ANFIS) and support vector regression (SVR) in [52]. Both techniques were designed to achieve one-step-ahead prediction in the examples. The same datasets as in [49, 51] were used to compare the performance, and it concluded that SVR performed better than ANFIS at the price of higher training time.

Tian and Zuo [53] proposed an extended recurrent neural network (ERNN) to predict the health condition of gears. The incorporation of the Elman context layer in the proposed networks was to enhance its ability to model nonlinear time series. The authors also added self-feedbacks to the Jordan context neurons to improve the dynamic property of the predictors. In addition, output error was taken into account by the way of feeding it back to the hidden layer using the Jordan context neuron.

All the innovative treatments will make the proposed ERNN more stable and accurate. The feature used for trending is root mean square (RMS) of a vibration signal collected from an accelerated run-to-failure test of a gearbox. Before being fed as inputs, a Weibull curve fitting was conducted on these discrete RMS values so that the inputs are relatively smooth out. The trained ERNN is able to predict a time series of RMS in the future. In the proposed ERNN, there are two neurons in the input layer which implies that the data point in a time series only strongly depends on the two preceding values; with one output and the prediction error fed back into a Jordan network, this approach achieves one-step ahead prediction.

The authors in [54] developed a neural network (NN) approach and with a dynamic window selecting the number of training data as time proceeds. This approach could achieve time span adjustment and multi-step ahead prediction. The feature used for tracking gear pitting progression is the sideband index extracted from the signal after narrowband interference cancellation [55]. This approach attempts to mitigate the heavy reliance on the existence of failure histories; however, the statistical significance of the predicted results need further investigation because it affects the reliability of the algorithm.

Hussain and Gabbar [56] proposed a novel feature extraction technique based on the psychoacoustics phenomenon followed by a wavelet smoothing. The predictors of ANFIS and nonlinear autoregressive model with exogenous inputs (NARX) were tested on the vibration signals obtained from a planetary gearbox inside a wind turbine. It can be seen that the feature gradually increases with time and the authors attributed the reason of vibration increase to the oil loss. It is worth mentioning that the vibration data used in [56] called the National Renewable Energy Laboratory (NREL) through a consortium named Gearbox Reliability Collaborative (GRC) [57]. This consortium is dedicated to the improvement of the reliability of the gearbox in the wind turbines. It is acknowledged that the issue of reliability and maintenance of the gearbox in the wind turbine is of critical concern to the owner and operators in wind industry due to its harsh working environment and high inaccessibility. The complexity of the planetary structure of the gearbox in the wind turbine and the future uncertainty in operating conditions further increase the difficulties in developing effective diagnostics and prognostics methods for it.

In [58], the author used a Gaussian mixture to simulate the vibration signal of a gearbox. By noticing the ineffectiveness of using kurtosis of the residual signal to trend the crack growth, it suggested using a physical index instead. However, the interim model for simulation has no physical meaning. Tian et al. [59] investigated health indicator extraction using a one-stage gearbox dynamic model. It identified RMS based on the residual signal segments as a sensitive indicator for early crack detection as well as for subsequent crack growth trends. In addition, it was found that discrete wavelet transform (DWT) techniques can increase the sensitivity of the indicator. Apart from vibration signals, oil debris monitoring and acoustic emissions have also been used to detect and trend the gear fault propagation [60, 61]. Because the particles emerged in the lubrication oil can lead to excessive wear of gears, it is also an important perspective to directly investigate the degradation and RUL of contaminated lubrication oil used in the gearbox, as done by authors of [62].

2.3.2 Data-Driven Methods: Dynamic System

There is an increasing to model the damage propagation and observation processes as a dynamic system. Recently, there is an increasing volume of literature reported in PHM area that applies the Kalman filter [63] and the particle filter [64] to obtain posterior states and parameters of interest. Dynamic systems have natural interface with data and can achieve real-time implementation as new data arrive. The state transition and observation equations may be obtained using a data-driven process.

The experiments conducted on the spiral bevel gear test facility at NASA provided a gear condition monitoring dataset, which is often used by researchers to validate their gear prognostic models. The purposes of the experiments were to investigate the performance of gear material, tooth design and the effect of lubrication additives on gear fatigue strength. Interested readers can refer to [60] for a detailed description of the test rig and test procedure.

For each failure mode of gears, there exists the associated condition index (CI) to detect the incipient fault. However, there is no universal CI that is effective for all the failure modes. For example, residual RMS works well for tooth pitting but is not effective for gear eccentricity. By noticing this fact, Bechhoefer and He [65] developed a single health index (HI) by fusing multiple correlated condition index (CI) based on gearbox vibration data. The selected six CIs include residual RMS, energy operator RMS, FM0, narrowband kurtosis, amplitude modulation kurtosis and frequency modulation RMS, all of which have good sensitivity to the fault. Based on the six CIs, a single HI can be constructed in different ways: order statistic of CIs, a summation of CIs or the total energy of CIs.

In [66], authors used particle filters to track the pitting growth in gears and used the aforementioned HI [58] and oil debris mass (ODM) in [60] for validation. In the framework of the proposed method, the autoregressive integrated moving average (ARIMA) method was applied to define gear degradation state transition equation using ODM data while the observation equation was obtained by a double exponential smoothing model fitting a single vibrational HI and ODM.

Authors in [67] treated gear feature evolution process as a linear dynamic model. The feature was the component from a Hilbert transform of the vibration data. Designed for on-line applications, the Expectation-Maximization algorithm was first applied to estimate the model parameters using the data in the past time window, and then the linear state-space model was employed to predict future data points in the time series.

Wang et al. [68] utilized two Hidden Markov Models (HMMs) to design two health indicators for gear early fault detection and degradation trending, respectively. Particle filters were also used to track the health indicator evolution which followed an exponential decay. The health indicator in the state-space model was selected as a probability rather than any form of condition monitoring data, and which was proved to have less sensitivity to varied work load compared to RMS of the residual error signals.

The predictive models in data-driven methods are purely dependent on data originating from various sources that are available to us. Therefore, the data

availability is a critical prerequisite. Furthermore, the data quality is also demanding, which requires a qualified trending feature to be extracted from huge condition monitoring data as the first step. In addition, the performance of data-driven method is highly impacted by the noise in the data (e.g., large variance, outliers). In particular, when a gearbox is operating under time-varying conditions (e.g. varying loading and speed), the training stage in the machine learning techniques is difficult to implement because the varying operating conditions cannot be exhausted. In summary, with qualified data, the data-driven approach can be easily applied to complex systems. It is also worth mentioning that in practical applications, the failure threshold setting requires extra efforts because the feature has no direct physical meaning.

2.4 Integrated Gear Prognostics

Data-driven and physics-based methods are two main directions for failure prognostics method development. Their respective merits and drawbacks motivate the integrated prognostics methods which combine data and physics of failure to benefit from both. In the integrated prognostics methods, the fault progression model can be updated with the current fault state estimated from the condition monitoring data. The state-of-the-art integrated methods usually have update processes with data assimilation. The model update process could achieve uncertainty reduction and/or better robustness.

Kacprzynksi et al. [69] developed an integrated prognostics tool which predicts the gear life through fusion of physics-of-failure models and diagnostics information to achieve an improved prediction accuracy. The total probability of failure at a given time was defined as the product of two independent events: crack initiation and crack propagation to failure. The diagnostics information was fused though a mapping between the vibrational features (residual Kurtosis and residual peak to peak) and crack size. Hence, the current crack size estimation is able to shrink the uncertainty in failure time prediction. Uncertainty from multiple sources were also considered when applying Paris' law including uncertainty in loading, material properties, modeling uncertainty and the crack estimation uncertainty. The results showed a variance reduction in failure probability when diagnostics information was present.

Zhao et al. [70] proposed an integrated prognostics framework for gears with a crack at tooth root that combines physical models and condition monitoring data. Physical models include Paris' law, the fracture mechanics model and the one-stage gearbox dynamics model. To account for the uncertainty in crack propagation, material parameters in Paris' law are treated as random variables. Bayesian inference was applied to update the distribution of material parameters in Paris' law each time a crack size was observed. With more observations becoming available, RUL

predicted by the updated Paris' law became more accurate and precise (reduced uncertainty). In this way, not only is Paris' law applied in a stochastic way but tighter confidence bounds are also obtained in the predicted results. This framework also accounted for the effect of dynamic load on crack propagation through the one-stage gearbox dynamics model. Later, the authors investigated uncertainty quantification problem in the proposed integrated prognostics framework [71]. Stochastic collocation methods based on polynomial chaos expansion (PCE) was applied to improve computational efficiency in both likelihood calculation and RUL prediction. Compared to traditional Monte Carlo and analytical methods for uncertainty propagation, PCE exhibits the desired properties considering the optimal balance between computational accuracy and efficiency. Afterwards, this integrated prognostics framework and PCE based uncertainty quantification method were extended to dealing with time-varying operating conditions in [72].

As stated in Sect. 2.2, physics-based methods use deterministic parameters in their predictive model, which could cause errors for a specific unit. As far as the failure mode of sliding wear is concerned, most of the existing approaches applied Archard's model, seen in Eq. (3), as the wear depth propagation model. The wear coefficient k has been treated as a constant value obtained from experiment. Being aware of the variability in the wear coefficient, Zhao et al. [47] developed an integrated method for wear prediction by treating this coefficient as a random variable with the mass loss of the gear weight considered as the condition monitoring data. The wear coefficient can be updated in a Bayesian framework whenever the weight of gear is measured. Validation was conducted using a run-to-failure test on a planetary gearbox, and the results show that the integrated method can effectively capture the characteristics of the wear coefficient for a specific gear and lead to an accurate prediction on gear mass loss due to sliding wear.

The idea of constructing a single HI for gear prognostics was adopted in [65]. Instead of using a data-driven process to obtain the state transition equation, a state-space model was built based on Paris' law, where the extended Kalman filter was applied to obtain predictions on the HI, and the authors applied the method proposed in [73] to set the failure threshold of HI. The results showed that the bounds on the predicted HI became narrower as time proceeds which demonstrated the uncertainty reduction in the presence of new data.

Integration of condition monitoring data and physical models can enhance the predictive capability to a large extent, and with the presence of unavoidable errors in both models and data, their integration may achieve the optimal performance. If applied properly, sufficient data sources from condition monitoring and operational records can further increase the flexibility of integrated methods by devising innovative ways of integration. Additionally, the integrated prognostics method is individual oriented because condition monitoring data are specific to each individual unit. The associated uncertainty reduction mechanism will make the prediction more specific to this individual unit with more confidence.

3 Opportunities and Challenges in Gear Prognostics

Although prognostics attract increasing attention and many efforts have been taken to develop effective methods for RUL prediction, prognostics capabilities are far from perfect in meeting the requirements in real-world applications. The following challenges need to be addressed to improve the performance of PHM for gears.

- Increase the fidelity of physical models. To save the efforts needed to build high-fidelity physical models, existing methods usually simplify the actual working condition that the gears undergo, causing the discrepancy between predicted results and actual ones. For example, effects of dynamics, lubrication and load variation should be considered in crack or wear propagation modeling.
- Develop damage propagation model for more failure modes of gears. Beside the two failure modes reviewed in this Chapter, gears also have other failure modes, most of which rarely have well-established physical laws to describe their evolution. The theory of damage mechanics has already been used to study the pitting evolution in bearings [74], and it would be beneficial to investigate whether it can be applied to gears.
- Develop prognostics for multiple concurrent failure modes that may be dependent and interactive with each other. There is very little research conducted on this topic, and we foresee the difficulty of applying physics-based methods to deal with this problem. However, data-driven methods may as well face the challenges because an overall health indicator is virtually impossible when the vibration signature of simultaneous faults is unknown.
- Increase reliability and robustness of prognostic algorithms. It still needs further development of qualified prognostic indicators that are closely related to fault growth with small noise and insensitive to operating conditions.
- Improve computational efficiency of modeling and uncertainty quantification to facilitate real-time application of PHM system.
- Develop an intelligent PHM system that has synthesized functionalities of fault detection, fault classification, fault assessment, fault tracking and decision making. Integrate PHM outputs into the control module of the monitored system to achieve minimum specialist involvement.

4 Conclusions

We have reviewed prognostics methods for gear health management in this chapter. Because these methods are specific to gears, there must be further analysis involved for gear health monitoring or physical models for gear kinematics and mesh contact. As such, along with physical models describing damage propagation in gears, we also review the extracted features based on gear condition monitoring for

prognostic purpose. The merits and drawbacks of four categories of prognostics methods are discussed. Opportunities and challenges in the future research are also suggested.

References

1. G. Vachtsevanos, F.L. Lewis, M. Roemer, A. Hess, B. Wu, *Intelligent Fault Diagnosis and Prognosis for Engineering Systems* (Wiley, 2006)
2. P.D. McFadden, Detecting fatigue cracks in gears by amplitude and phase demodulation of the meshing vibration. J. Vib. Acoust. Stress Reliab. Des. **108**, 165–170 (1986)
3. W.J. Wang, P.D. McFadden, Decomposition of gear motion signals and its application to gearbox diagnostics. J. Vib. Acoust. **117**, 363–369 (1995)
4. D. Brie, M. Tomczak, H. Oehlmann, A. Richard, Gear crack detection by adaptive amplitude and phase demodulation. Mech. Syst. Signal Process. **11**, 149–167 (1997)
5. P. Vecer, M. Kreidl, R. Smid, Condition indicators for gearbox condition monitoring systems. Acta Polytechnica **45**, 35–43 (2005)
6. A.K. Jardine, D. Lin, D. Banjevic, A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mech. Syst. Signal Process. **20**, 1483–1510 (2006)
7. A. Heng, S. Zhang, A. Tan, J. Mathew, Review—rotating machinery prognostics: state of the art, challenges and opportunities. Mech. Syst. Signal Process. **23**, 724–739 (2009)
8. L. Liao, F. Kottig, Review of hybrid prognostics approaches for remaining useful life prediction of engineered systems, and an application to battery life prediction. IEEE Trans. Reliab. **63**, 191–207 (2014)
9. J.W. Hines, A. Usynin, Current computational trends in equipment prognostics. Int. J. Comput. Intell. Syst. **1**, 94–102 (2008)
10. K.L. Tsui, N. Chen, Q. Zhou, Y. Hai, W. Wang, Prognostics and health management: a review on data driven approaches. Math. Probl. Eng. **2015**, Article ID 793161, 17 pp. (2015)
11. E. Zio, Prognostics and health management of industrial equipment, in *Diagnostics and Prognostics of Engineering Systems: Methods and Techniques*, ed. by S. Kadry (IGI Global, 2012), pp. 333–356
12. X.S. Si, W. Wang, C.H. Hu, D.H. Zhou, Remaining useful life estimation—a review on the statistical data driven approaches. Eur. J. Oper. Res. **213**, 1–14 (2011)
13. J.Z. Sikorska, M. Hodkiewicz, L. Ma, Prognostic modeling options for remaining useful life estimation by industry. Mech. Syst. Signal Process. **25**, 1803–1836 (2011)
14. A.K.S. Jardine, A.H.C. Tsang, *Maintenance, Replacement, and Reliability: Theory and Applications* (CRC Press, 2005)
15. W. Weibull, A statistical theory of the strength of materials. Ingeniors Vetenskaps Akademiens, Handlingar **151** (1939)
16. G. Lundberg, A. Palmgren, Dynamic capacity of rolling bearing. Ingeniors Vetenskaps Akademiens, Handlingar **196** (1947)
17. J.J. Coy, D.P. Townsend, E.V. Zaretsky, Dynamic capacity and surface fatigue life for spur and helical gears. J. Lubr. Technol. 267–274 (1976)
18. P.C. Paris, F. Erdogen, A critical analysis of crack propagation laws. J. Basic Eng. **85**, 528–534 (1963)
19. H. Liebowitz, E.T. Moyer, Finite element methods in fracture mechanics. Comput. Struct. **31**, 1–9 (1989)
20. R.D. Henshell, K.G. Shaw, Crack tip finite elements are unnecessary. Int. J. Numer. Meth. Eng. **9**, 495–507 (1975)

21. R.S. Barsoum, On the use of isoparametric finite elements in linear fracture mechanics. Int. J. Numer. Meth. Eng. **10**, 25–37 (1976)
22. L.B. Sills, D. Sherman, On quarter-point three-dimensional finite elements in linear elastic fracture mechanics. Int. J. Fract. **41**, 177–196 (1989)
23. J.R. Rice, A path independent integral and the approximate analysis of strain concentration by notches and cracks. J. Appl. Mech. **35**, 379–386 (1968)
24. B. Abersek, J. Flasker, Numerical methods for evaluation of service life gear. Int. J. Numer. Meth. Eng. **38**, 2531–2545 (1995)
25. S. Pehan, T.K. Hellen, J. Flasker, Applying numerical methods for determining the service life of gears. Fatigue Fract. Eng. Mater. Struct. **18**, 971–979 (1995)
26. L.E. Spievak, P.A. Wawrzynek, A.R. Ingraffea, D.G. Lewicki, Simulating fatigue crack growth in spiral bevel gears. Eng. Fract. Mech. **68**, 53–76 (2001)
27. S. Glodez, M. Sraml, J. Kramberger, A computational model for determination of service life of gears. Int. J. Fatigue **24**, 1013–1020 (2002)
28. J.E. Collipriest, An experimentalist's view of the surface flaw problem, in *The Surface Crack: Physics Problems Compute Solutions, ASME* (1972), pp. 43–61
29. K. Inoue, M. Kato, N. Takatsu, Fracture mechanics based evaluation of strength of carburized gear teeth, in *Proceedings of the JSME International Conference on Motion and Power Transmissions* (1991), pp. 801–806
30. M. Guagliano, L. Vergani, Effect of crack closure on gear crack propagation. Int. J. Fatigue **23**, 65–73 (2001)
31. E. Wheeler, Spectrum loading and crack growth. J. Basic Eng. Trans. ASME **94**, 181–186 (1972)
32. S. Pehan, T.K. Hellen, J. Flakder, S. Glodez, Numerical methods for determining stress intensity factors vs crack depth in gear tooth roots. Int. J. Fatigue **19**, 677–685 (1997)
33. D.G. Lewicki, R. Ballarini, Gear crack propagation investigations. Tribotest J. **5**, 157–172 (1998)
34. D.G. Lewicki, R. Ballarini, Rim thickness effects on gear crack propagation life. Int. J. Fract. **87**, 59–86 (1997)
35. D.G. Lewicki, R.F. Handschuh, L.E. Spievak, P.A. Wawrzynek, A.R. Ingraffea, Consideration of moving tooth load in gear crack propagation predictions. Trans. ASME **123**, 118–124 (2001)
36. C. Li, H. Lee, Gear fatigue crack prognosis using embedded model, gear dynamic model and fracture mechanics. Mech. Syst. Signal Process. **19**, 836–846 (2005)
37. S. Choi, C.J. Li, Practical gear crack prognosis via gear condition index fusion, gear dynamic simulator, and fast crack growth model. J. Syst. Control Eng. **221**, 465–473 (2007)
38. C.J. Li, H. Lee, S.H. Choi, Estimating size of gear tooth root crack using embedded modeling. Mech. Syst. Signal Process. **16**, 841–852 (2002)
39. S. Choi, C.J. Li, Estimation of gear tooth transverse crack size from vibration by fusing selected gear condition indices. Meas. Sci. Technol. **17**, 2395–2400 (2006)
40. J.F. Archard, Contact and rubbing of flat surface. J. Appl. Phys. **24**, 981–988 (1953)
41. T.F.J. Quinn, Review of oxidation wear, Parts I and II. Tribol. Int. **16**, Part I 257–305 and Part II 305–315 (1983)
42. S. Wu, H.S. Cheng, A sliding wear model for partial-EHL contacts. ASME J. Tribol. **113**, 134–141 (1991)
43. S. Wu, H.S. Cheng, Sliding wear calculation in spur gears. J. Tribol. **115**, 493–500 (1993)
44. A. Flodin, S. Andersson, Simulation of mild wear in spur gears. Wear **207**, 16–23 (1997)
45. A. Flodin, S. Andersson, Simulation of wild wear in helical gears. Wear **241**, 123–128 (2000)
46. P. Bajpai, A. Kahraman, N.E. Anderson, A surface wear prediction methodology for parallel-axis gear pairs. J. Tribol. **126**, 597–605 (2004)
47. F. Zhao, Z. Tian, Y. Zeng, Integrated prognostics method for gear wear prediction. Finished
48. P. Tse, D. Atherton, Prediction of machine deterioration using vibration based fault trends and recurrent neural networks. J. Vib. Acoust. **121**, 355–362 (1999)

49. W.Q. Wang, M.F. Golnaraghi, F. Ismail, Prognosis of machine health condition using neuro-fuzzy systems. Mech. Syst. Signal Process. **18**, 813–831 (2004)
50. W. Wang, F. Ismail, F. Golnaraghi, Assessment of gear damage monitoring techniques using vibration measurements. Mech. Syst. Signal Process. **15**, 905–922 (2001)
51. W. Wang, An intelligent system for machinery condition monitoring. IEEE Trans. Fuzzy Syst. **16**, 110–122 (2008)
52. B. Samanta, C. Nataraj, Prognostics of machine condition using soft computing. Robot. Comput.-Integr. Manuf. **24**, 816–823 (2008)
53. Z. Tian, M.J. Zuo, Health condition prediction of gears using a recurrent neural network approach. IEEE Trans. Reliab. **59**, 700–705 (2010)
54. X. Zhang, L. Xiao, J. Kang, Degradation prediction model based on a neural network with dynamic windows. Sensors **15**, 6996–7015 (2015)
55. X. Zhang, J. Kang, E. Bechhoefer, J. Zhao, A new feature extraction method for gear fault diagnosis and prognosis. Eksploatacja i Niezawodnosc—Maint. Reliab. **16**, 295–300 (2014)
56. S. Hussain, H.A. Gabbar, Vibration analysis and time series prediction for wind turbine gearbox prognostics. Int. J. Progn. Health Manage. **4**, 69–79 (2013)
57. H. Link, W. LaCava, J. V. Dam, B. McNiff, S. Sheng, R. Wallen, W. McDade, S. Lambert, S. Butterfield, F. Oyague, Gearbox reliability collaborative project report: findings from phase 1 and phase 2 testing. Technical Report, NREL/TP-5000-51885 (2011)
58. W. Wang, Toward dynamic model-based prognostics for transmission gears, in *Proceedings of SPIE*, vol. 4733 (2002)
59. Z. Tian, M.J. Zuo, S. Wu, Crack propagation assessment for spur gears using model-based analysis and simulation. J. Intell. Manuf. **23**, 239–253 (2012)
60. P. Dempsey, R. Handschuh, A. Afjeh, Spiral bevel gear damage detection using decision fusion analysis, NASA/TM-2002-211814 (2002)
61. Y. Qu, D. He, J. Yoon, B.V. Hecke, J. Zhu, E. Bechhoefer, Gearbox tooth cut fault diagnostics using acoustic emission and vibration sensors—a comparative study. Sensors **14**, 1372–1393 (2014)
62. J. Zhu, J. Yoon, D. He, E. Bechhoefer, Online particle-contaminated lubrication oil condition monitoring and remaining useful life prediction for wind turbines. Wind Energy **18**, 1131–1149 (2015)
63. R.E. Kalman, A new approach to linear filtering and prediction problems. Trans. ASME J. Basic Eng. **82**, 35–45 (1960)
64. M.S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans. Signal Process. **50**, 174–188 (2002)
65. E. Bechhoefer, D. He, A process for data driven prognostics, in *Proceedings of the 2012 Conference of the Society for Machinery Failure Prevention Technology (MFPT)* (2012), pp. 193–212
66. D. He, E. Bechhoefer, J. Mam, J. Zhu, A particle filtering based approach for gear prognostics, in *Diagnostics and Prognostics of Engineering Systems: Methods and Techniques*, Chap 13 (2012)
67. M. Gasperin, D. Juricic, P. Boskoski, J. Vizintin, Model-based prognostics of gear health using stochastic dynamical models. Mech. Syst. Signal Process. **25**, 537–548 (2011)
68. D. Wang, Q. Miao, Q. Zhou, G. Zhou, An intelligent prognostic system for gear performance degradation assessment and remaining useful life estimation. J. Vib. Acoust. **137**, 021004-1–02100402100412 (2015)
69. G. Kacprzynski, A. Sarlashkar, M. Roemer, A. Hess, B. Hardman, Predicting remaining life by fusing the physics of failure modeling with diagnostics. J. Miner. Metals Mater. Soc. **56**, 29–35 (2004)
70. F. Zhao, Z. Tian, E. Bechhoefer, Y. Zeng, An integrated prognostics method under time-varying operating conditions. IEEE Trans. Reliab. **64**, 372–387 (2013)
71. F. Zhao, Z. Tian, Y. Zeng, A stochastic collocation approach for efficient integrated gear health prognosis. Mech. Syst. Signal Process. **39**, 372–387 (2013)

72. F. Zhao, Z. Tian, Y. Zeng, Uncertainty quantification in gear remaining useful life prediction through an integrated prognostics method. IEEE Trans. Reliab. **62**, 146–159 (2013)
73. E. Bechhoefer, D. He P. Dempsey, Gear health threshold setting based on a probability of false alarm, in *Annual Conference of the Prognostics and Health Management Society* (2011)
74. J. Qiu., C. Zhang, B. Seth, S.Y. Liang, Damage mechanics approach for bearing lifetime prognostics. Mech. Syst. Signal Process. **16**, 817–829 (2002)

Probabilistic Model-Based Prognostics Using Meshfree Modeling

Stephen Ekwaro-Osire, Haileyesus Belay Endeshaw, Fisseha M. Alemayehu and Ozhan Gecgel

Abstract Improved system reliability and reduced maintenance cost are guaranteed if the prediction of remaining useful life (RUL) is deemed to be accurate. Energy systems, like wind turbines, are the primary beneficiaries of this achievement as they tend to suffer from an unexpected early life failure of components that resulted in the loss of revenue and high maintenance costs. The issue of uncertainty in the prediction of a future state is yet a prevailing issue in prognostics and due attention is paramount. Hence, there is a need for establishing a comprehensive framework to quantify uncertainty in prognostics and this research addresses this issue by considering a research question that reads 'can uncertainty considerations improve the prediction of RUL?' The following specific aims were developed to answer the research question: (1) develop a meshfree cantilever beam with uncertainty in loading conditions, and (2) predict remaining useful life reliably. A probabilistic framework was developed that efficiently predicts remaining useful life of a component using a combination of meshfree model and degradation model. To account for prediction uncertainty, modeling and loading uncertainties are quantified and incorporated into the framework. As an example, the problem of a cantilever beam subjected to a fatigue loading was considered and local radial point interpolation method was used to find the stresses. The cyclic stresses and the damage model, constructed using the *S-N* equation, are implemented in the prognostics framework to predict the RUL. Uncertainties in the RUL were quantified in terms of probability density functions, cumulative distribution functions, and 98% confidence limit. The prognostics framework is flexible and can be used as a starting point for RUL prediction of other physical phenomena such as crack propagation, by incorporating more sources of uncertainties in order to make it comprehensive.

S. Ekwaro-Osire (✉) · H.B. Endeshaw · O. Gecgel
Department of Mechanical Engineering, Texas Tech University, Lubbock, TX, USA
e-mail: stephen.ekwaro-osire@ttu.edu

F.M. Alemayehu
School of Engineering, Computer Science and Mathematics,
West Texas A&M University, Canyon, TX, USA

S. Ekwaro-Osire et al. (eds.), *Probabilistic Prognostics and Health Management of Energy Systems*, DOI 10.1007/978-3-319-55852-3_5

Keywords Uncertainties · Prognostics and health management · Remaining useful life · Probabilistic · Meshfree modeling

1 Introduction

1.1 Prognostics and Health Management

Accurate failure predictions and health management could significantly reduce the operation and maintenance costs of an energy system. These costs account for a large amount of the total cost of the system [1]. While prognostics is the prediction of future states of a system or a component to predict the remaining useful life (RUL), health management refers to instantaneous health monitoring of a system [2–4]. System diagnosis results (measurements) will be used as the initial input data to prognostics [2]. Generally, prognostics follows two important steps: (i) state estimation using Bayesian tracking and (ii) future state prediction [5].

There are three prognostics methodologies known as model-based, statistical data driven, and hybrid methods [6, 7]. Data-driven methods build a relationship between measured data and the state of a system using machine learning and pattern recognition methodologies [8]. It has been shown that the results from a purely data-driven approach without the physical model results in a high uncertainty values and inconvenient for long term predictions [9]. An extensive review of statistical data driven approaches is provided by Si et al. [10]. Model-based approaches implement mathematical formulations to approximate the physics of the system for RUL prediction [8]. Hybrid methodologies use the combination of model-based and data-driven approaches [6]. Studies on data-driven approaches can be found in [8, 11]. Detail review of common model-based and data-driven algorithms and their advantages and disadvantages can be found in [12].

Since prognostics refers to predicting future state of a system, it is necessary to consider uncertainties to account for eminent variability [3, 13, 14]. There are various sources of uncertainty that should be considered in prognostics [3, 15]. Sources of uncertainty include sensor and measurement errors, state estimates, future loading conditions, and environmental conditions [16]. Of these sources of uncertainties, future loading is the most challenging in prognostics [5, 17]. Another classification of sources of uncertainty is present uncertainty, future uncertainty, modeling uncertainty and prediction method uncertainty [5, 16]. It should also be important to know that Remaining Useful Life (RUL) prediction of prognostics should be expressed as a distribution with a given confidence interval instead of a specific life estimate [18, 19]. It should be based on this confidence interval that business decisions should be made [18]. Preventive maintenance decisions would be more justifiable based on the confidence intervals and probability values of RULs.

1.2 Modeling

Finite element method (FEM) has been widely used for problems which do not have a closed form solution. Although FEM is a very useful numerical technique in finding approximate solutions, it has also some limitations. The first limitation arises from the necessity of re-meshing in crack propagation problems. Due to inherent nature of the cracks, crack propagation path is random and complex which causes misalignment with the edge of the finite elements. That results in the generation of discontinuous displacement fields within the elements [20–23]. Thus, re-meshing in FEM is required which makes finite element (FE) computation cumbersome and computationally expensive [20]. Moreover, since it is a mesh-based interpolation, FEM does not work well with distorted meshes which is another limitation [23]. In large deformations, from distorted or low-quality meshes, the accuracy of stresses at element interfaces get low which is caused by the assumption of continuous displacement field during FEM formulation [24]. To overcome some of the problems of FEM, the extended finite element method (X-FEM) was developed. Bordas et al. [25] presented a C++ open source framework for X-FEM, compiled by Visual Studio. In X-FEM the crack is modelled by adding enrichments to standard FEM in order to improve the approximation of crack propagation without the need of re-meshing. This achievement is done by describing the crack approximately by local signed distances of the nodes around the face of the crack [26]. Although X-FEM avoids the issue of re-meshing in FEM, lack of smoothness and inability of handling distorted meshes decreases the accuracy and creates limitation [23].

Meshfree methods (also known as meshless methods) were formulated to avoid some of the issues associated with FE approximations; i.e. issues caused by reliance on the mesh. Meshfree methods were applied in areas of solid mechanics, fluid dynamics, and astrophysics [23]. Unlike FE methods, meshfree methods do not depend on the predefined mesh to generate a system of algebraic equations for the problem domain [27]; instead, meshfree methods use nodes (also known as field nodes) for approximation [23, 24]. In other words, a meshless method is an approximation rather than an interpolation like FEM. This feature requires very careful treatments of essential boundary conditions, mirror symmetries and moving discontinuities [28]. The quality of higher order of continuity in meshfree methods becomes especially very beneficial in problems with discontinuities, such as cracks, and they can easily model problems with moving discontinuities such as crack propagation and phase transformation [23]. Some of the most common meshfree methods include smooth particle hydrodynamics method, element-free Galerkin (EFG) method [29], reproducing kernel particle method (RKPM), radial point interpolation method, and meshless local Petrov-Galerkin (MLPG) method [30]. Meshfree methods can be classified based on: formulation methods (local weak-forms, global weak-forms, and collocation techniques), function approximation methods (moving least squares, integral representation, point interpolation methods), and domain representation (domain-type and boundary-type methods)

[24]. EFG and RKPM are based on global weak-forms, whereas MLPG method is based on local weak-forms. Meshfree methods which are based on global weak-forms require background cells to perform integration, whereas those based on local weak-forms use quadrature domains for integration. Local meshfree methods use the so-called support domains for the sake of interpolation. Thus, interpolation will be based on the field nodes inside the support domain.

1.3 Filtering

Model parameter identification can be performed by various techniques, which are based on Bayesian inference to update model parameters using measured data. Filtering methods that are based on Bayesian inference include Kalman filter, particle filter [5, 18], and the extended Kalman filter, the Bayesian method [12]. Bayesian methods can be classified into two broad categories: non-linear process and linear process. Linear process Bayesian methods refer to Kalman filter. Kalman filter is applicable for linear systems with additive Gaussian noise. Kalman filter is not applicable to non-linear systems and/or non-Gaussian noise since these distributions require high dimensional integrals, which are not possible to evaluate analytically [31]. Some examples of Kalman filter applications are given in [32–34]. Non-linear process methods can be further categorized based on the type of noise added to the system as Gaussian additive noise (extended or unscented Kalman filter) and non-Gaussian additive noise (particle filter) [7, 18, 35, 36]. Non-linear filtering employs non-linear and non-Gaussian state-space model and estimates at least the first two states by using measured data [37].

Filtering approaches involve using an appropriate model function to estimate the state of the system or the component, $x(t)$, using the observed data, $y(t)$. This is because the state of the system may not be the parameter that is measured (observed) [17]. For instance, the quantity of oil debris can be used to estimate the level of damage to a gearbox. In this case, the quantity of oil debris is observed data, $y(t)$, and the level of damage of the gearbox is the state, $x(t)$, of the gearbox [38]. Similarly, battery life estimation involves the determination of the capacity, which is a function of the internal resistance. The internal resistance is directly measured to estimate the capacity of the battery. Therefore, the capacity of the battery is the state, $x(t)$, whereas the internal performance is $y(t)$. In other words, the capacity can be expressed as a function of the internal resistance, i.e. $x(t) = f(y(t))$ [7]. However, in some applications, observed data may be the same as the state of the system or component, i.e. $x(t) = y(t)$. A good example is the problem of crack propagation, where both the measured data, $y(t)$, and the state of the component, $x(t)$, is the crack size [7, 33].

Particle filter employs Monte Carlo sampling to implement a filtering method using Bayesian inference [9]. Particle filter has a capability of integrating measurement data from different sources systematically [31]. An et al. [7] and Arulampalam et al. [39] presented tutorials on methodology and implementation of the

particle filter. Particle filter has a greater process efficiency and is more suitable for selecting a variety of initial distributions [40]. It has also the advantage of straightforward implementation and the ability to control performance by the number of particles used [35]. Thus it can be directly applicable to fault detection and identification [41]. Orchard et al. [42] proposed a particle filter approach integrated with a correction algorithm and compared the results with those of Kalman filter; it was reported that the proposed approach is greater in accuracy and precision [37]. Particle filtering follows three important steps: prediction, updating, and resampling. In the prediction step, the prior probability density function is obtained. In the updating step, the likelihood of the prior distribution is obtained. The last step is resampling where measurement data is used to remove some of the samples and duplicate the others according to the weight that is given to the particles [7].

1.4 Degradation Models

The discretized form of Paris' law, presented in Eq. 1, has been selected as a degradation model to estimate future states and predict RUL of a system or a component based on available data. This model is used recursively to predict future crack sizes based on previous crack size estimate [7, 43], as given by

$$a_k = a_{k-1} + C(\Delta K)_{k-1}^{m}(\Delta \mathrm{N})_{k-1} \tag{1}$$

where a_k and a_{k-1} are the estimated and prior crack sizes, respectively, ΔN is the added number of cycles to the component, $(\Delta K)_{k-1}$ is the $(k–1)$th stress intensity range per cycle that depends on stress range $\Delta\sigma$ and empirical material constants C & m.

In addition, the life of a component (number of cycles-to-failure) under a cyclic loading can be determined from:

$$N = \left(\frac{\sigma_a}{a_f}\right)^{1/b_f} \tag{2}$$

where σ_{a} is a cyclic stress amplitude and a_{f} and b_{f} are the model parameters that are updated according to Manson's method of cumulative fatigue damage (CFD) [44].

1.5 Probabilistic Analysis

Uncertainty is ubiquitous in engineering systems. Not only is data collection usually uncertain, but also most engineering parameters are random in nature. This hinders the state of a system from being exactly determined. Hence, future state is expressed in terms of probability or reliability, which is the probability of obtaining

the desired performance [45]. Traditional deterministic design approaches are not practical in such aspect since they do not take uncertainties into consideration. On the other hand, the probabilistic analysis determines the reliability of an engineering system by quantifying uncertainties [46]. In the probabilistic analysis, random variables are carefully selected and their uncertainty is quantified and expressed in terms of probability density function (PDF) and cumulative distribution function (CDF). The random variables will then be used in a mathematical model to determine the uncertainty associated with the response variable. The role of probabilistic analysis in the reliability of wind turbine gearboxes is discussed in [19, 46, 47].

1.6 Remaining Useful Life Prediction

In model-based prognostics, RUL prediction involves filtering of observed data and state estimation using a degradation model. First, RUL prediction involves filtering of available data for state estimation. Once the available data have been used, future state prediction will be solely based on the physics of the problem, i.e. degradation model. The physics-based model performs RUL prediction until the criteria for the end of life is reached [7]. Sankararaman et al. [17] implemented first order reliability method to quantify uncertainties including loading uncertainty and state uncertainty in a lithium-ion battery. They quantified the uncertainty in the terms of variations in remaining battery life. Other researchers employed prognostics of fatigue crack propagation of a gear using gear dynamics model and FEM analysis to determine RUL [48]. Si et al. [10] outlined challenges regarding prediction of RUL. The first challenge is in building model-based prediction methods since model-based prediction is especially important for applications where measured data are unavailable. The second challenge is the issue of integrating multi-dimensional data from condition monitoring. The last challenge is the development of a model for RUL prediction of a system considering multiple modes of failure.

1.7 Motivation

The traditional schedule-based maintenance is not capable of preventing failure efficiently. This may result in system deterioration and an eminent failure causing unforeseen downtime [6]. This is especially apparent in wind turbine gearbox system, which often fails before its expected lifetime [49]. The degree of uncertainty is higher in prognostics than in diagnostics due to additional future state uncertainty. Thus, it is crucial to quantify uncertainty in prognostics [17]. Measurement, modeling, material property and future loading are some of the

uncertainties in prognostics [15, 16]. For this reason, uncertainties are considered in prognostics and a distribution of RUL is sought [6]. Despite the numerous research activities in prognostics and health management, the issue of accurate uncertainty quantification in prognostics and a reliable RUL prediction still prevails. Operation and maintenance costs of a wind turbine consist of a large portion of its total lifetime cost [50]. Hence, there is a need for a comprehensive framework that efficiently quantifies uncertainty in prognostics [5, 16]. Accurate prediction of RUL will improve reliability and reduce the maintenance cost [51]. There is a need for reliable physics-based models which can substitute for data-driven methods in case of insufficient data availability [10, 18]. Meshfree methods are reported to have better performance than FEM in modeling of discontinuous structures such as cracks [23, 24]. A meshfree method was selected in this study to investigate their advantages so that their use can be extended to discontinuous structures.

1.8 Research Question and Specific Aims

This research addresses the issue of uncertainty in RUL prediction by considering a research question that reads 'Can uncertainty considerations improve the prediction of RUL?' The following specific aims were developed to answer the research question: (1) develop a meshfree cantilever beam with uncertainty in loading conditions, (2) predict remaining useful life using probabilistic methods.

2 Methodology

Prediction of RUL of a cantilever beam under fluctuating and varying fluctuating (constant and variable amplitude) fatigue loading was considered. Results from meshfree modeling show the behavior of the cantilever beam under deterministic loading condition. Besides, the framework that was developed to predict RUL is presented. Results of the RUL prediction using a deterministic approach are discussed first [44], followed by a probabilistic RUL prediction. Deterministic results of the RUL from the current state up to failure are presented both for constant amplitude fatigue loading as well as varying fluctuating loading conditions [19]. Changes in the path of the RUL as a result of CFD were also discussed. Probabilistic RUL prediction also considers the two case: fluctuating fatigue loading with constant amplitude and variable fluctuating loading. Results are presented as PDFs and CDFs of the RUL at selected future cycles. The range of possible RUL values and the implication of the reduction in the range of RUL values with an increase in the number of cycles are explained.

2.1 Efficient Modeling

An efficient meshfree method called local radial point interpolation method (LRPIM) that employs augmented basis functions of radial and polynomial basis functions [24]; and presented in detail in authors' previous publication [19], is implemented to calculate the stress values near the fixed end of the cantilever beam. The LRPIM programming flowchart is also presented in the same publication [19]. For ease of study, a 2D cantilever beam, shown in Fig. 1, subjected to a completely reversed random cyclic loading is considered. The following model parameters given in Table 1, as indicated in [19], are used to model the system.

2.2 RUL Prediction

A MATLAB framework, shown in Fig. 2, was developed for RUL estimation of the cantilever beam subjected to cyclic loading. The framework consists of subroutines for LRPIM modeling of the cantilever beam and material degradation modeling for RUL prediction. The model is capable of performing both deterministic and probabilistic RUL estimations. The first set of input parameters are related to the characteristics of the cantilever beam such as the Young's modulus, E, the width, B, height, H, poison's ratio, v, the ultimate strength of the material, S_u, and the traction, P. The second set of inputs, which are model parameters for the meshfree modeling, is listed in Table 1.

The framework first uses the meshfree method subroutine to compute the maximum normal stress on the beam. In the case of deterministic analysis, the maximum stress is directly entered into the degradation model subroutine together

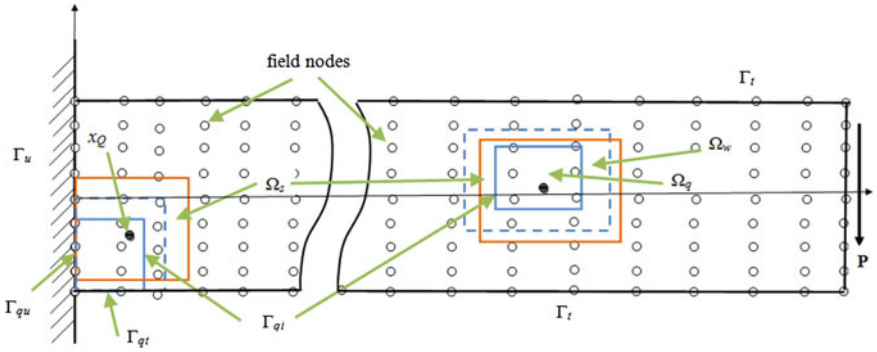


Fig. 1 Cantilever beam depicting nodes, Gauss quadrature points, x_Q, quadrature domains, Ω_q, support domains, Ω_s, weight domain, Ω_w, local boundaries (Γ_{qt}, Γ_{qu}, and Γ_{qi}), and global boundaries (Γ_u and Γ_t)

Table 1 LRPIM model parameters

Parameter	Value	Parameter	Value	Parameter	Value
α_s	3.0	α_c	4.0	*ndx*	2
α_q	1.7	*q*	1.03	*ndy*	2
α_w	3.0	*dcx*	*L*/20	*p*	3
Radial basis function	Multi-quadrics	*dcy*	*H*/8		

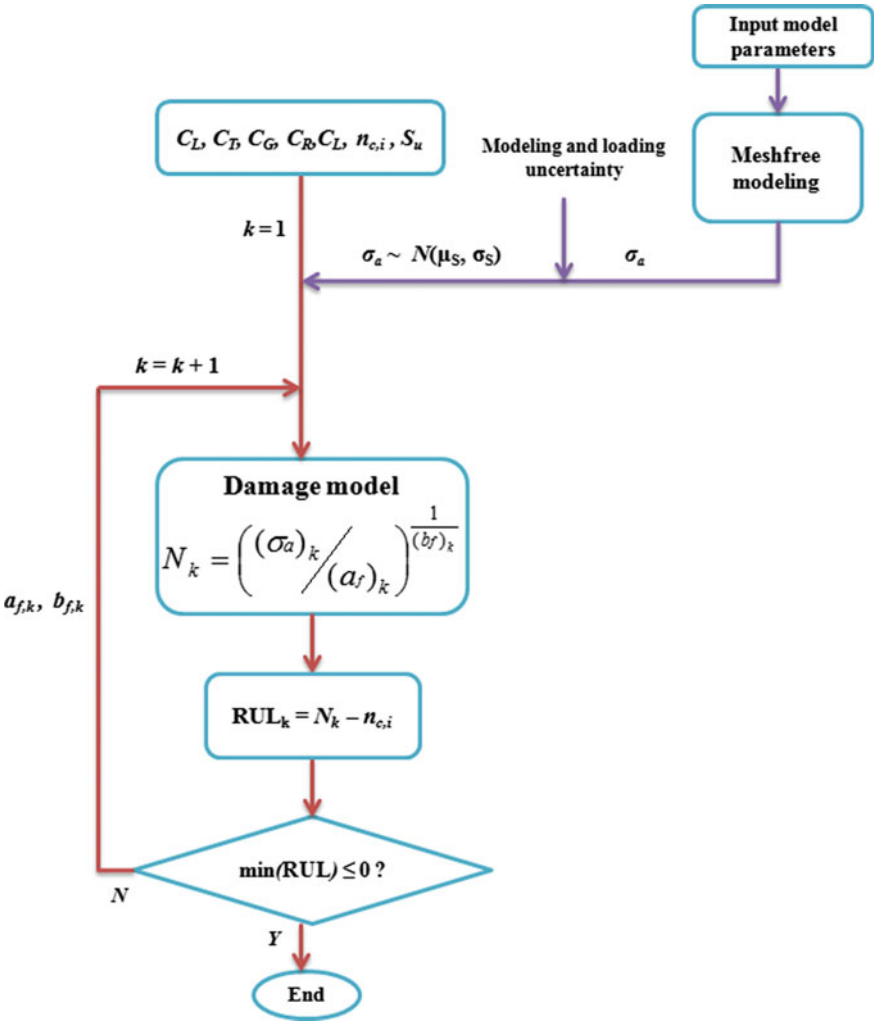


Fig. 2 Flowchart of a framework for prognostics of a cantilever beam under cyclic loading [19]

with the ultimate strength, S_u. The relationship between dynamic loading, σ_a, and fatigue life, N, was presented in [19] as,

$$\sigma_a = a_f N^{b_f} \tag{3}$$

where

$$b_f = -\frac{1}{3}\log(S_m/S_e) \tag{4}$$

and

$$a_f = 10^{(\log(S_m) - 3b_f)}. \tag{5}$$

S_m is the fatigue strength at 10^3 given by 0.9 S_u and S_e is the endurance limit of the beam given by:

$$S_e = \frac{1}{2} S_u C_L C_G C_S C_T C_R \tag{6}$$

where C_L, C_G, C_S, C_T, and C_R are correction factors due to loading, size, surface, temperature, and reliability, respectively. From Eq. 3, one can find that $N = \left(\frac{\sigma_a}{a_f}\right)^{1/b_f}$. Hence, the degradation model to predict RUL is given by a piecewise function for r number of loading and unloading events as

$$RUL = \begin{cases} \left(\frac{\sigma_{a,1}}{a_{f,1}}\right)^{\frac{1}{b_{f,1}}} - n_c & n_c = [0, n_{c,1}] \\ \left(\frac{\sigma_{a,2}}{a_{f,2}}\right)^{\frac{1}{b_{f,2}}} - n_c & n_c = [n_{c,1}, n_{c,2}] \\ \vdots & \\ \left(\frac{\sigma_{a,r}}{a_{f,r}}\right)^{\frac{1}{b_{f,r}}} - n_c & n_c = [n_{cr-1}, n_{c,r}] \end{cases} \tag{7}$$

where n_c is the number of cycles. Equation (7) is used to predict the RUL of a component subjected to σ_a for n_c cycles r number of times. Note that parameters a_f and b_f are updated in every loading case due to CFD according to Manson's rule [43].

2.2.1 Deterministic RUL Prediction

In deterministic RUL prediction, the degradation model subroutine returns the expected life of the beam undergoing the given stress after a given n_c number of cycles. The framework then obtains the corresponding RUL by employing Eq. 7. In the case of multiple loading scenarios, the framework follows a recursive algorithm to compute the RUL given by Eq. 7 multiple times. It accumulates an array of RUL values that correspond to each loading case and plots RUL versus number of cycles

to show the path of the RUL until failure. The deterministic analysis section of this study includes two loading cases. The first one employs multiple applications of single amplitude loads, whereas the second case employs multiple applications of multiple amplitude loads. In each case, the loads are consecutively applied until failure.

2.2.2 Probabilistic RUL Prediction

Probabilistic RUL prediction employed continuous (uninterrupted) loading and consecutive (interrupted) loading conditions. In both loading conditions, the framework takes the maximum stress value obtained from the meshfree subroutine as a mean to generate a PDF of the stress. Monte Carlo method was then used in the material degradation subroutine to compute the RUL using Eq. 7.

In the case of uninterrupted loading condition, 10% of standard deviation (SD) was employed to obtain the stress PDF. The PDF of initial life, N, of the beam was obtained from Eq. 7 and the RUL versus the number of cycles plot was generated by subtracting the number of cycles, n_c, from N. Note that parameters a_f and b_f are not updated (i.e. CFD is not incorporated) since Eq. 7 is used only at the initial life prediction. Single mean and multiple mean values were considered in the RUL prediction of uninterrupted loading condition.

The interrupted loading application was categorized as a fluctuating (single mean amplitude) and varying fluctuating (multiple mean amplitude loading) applications. In single mean application, 10% SD was employed to the mean value of 200 MPa at each loading application. In other words, similar loading conditions were employed consecutively until failure. In the case of multiple mean application, however, the mean value was changed in every loading application keeping the SD constant. The degradation model parameters a_f and b_f of each sample stress were updated with every level of load application. Due to the randomness nature of the stress, the RUL was a PDF with each level of load application. The median RUL and the RUL at the bounds of the 98% confidence level at each level of load application was selected. After the application of the last loading, the framework plots these RUL values. Finally, the above steps were repeated by using a SD of 5 and 15%, resulting in three sets of results corresponding to 5, 10, and 15% SD of the stress and results were plotted together with the rest of the corresponding analyses for comparison.

3 Results and Discussion

The results presented and discussed hereunder are based on the meshfree LRPIM model of a cantilever beam and a degradation model. In a previous publication [19] of the authors, stress and deflection results of the meshfree model were verified with exact solutions and have shown good agreement. Detail discussion about

verification of the meshfree model and its computational efficiency are provided in this publication.

3.1 RUL Prediction

3.1.1 Deterministic RUL Prediction

The deterministic RUL results for fluctuating loadings of single as well as varying amplitude is published in [19]. In this publication, it has been demonstrated that the RUL is expected to drop down at the end of each interval since CFD will reduce the endurance limit of the material after the beam has run for certain amount of cycles in each interval.

3.1.2 Probabilistic RUL Prediction

Similar to the loading conditions of the deterministic RUL estimation published previously [19], multiple applications of cyclic loads with constant and varying amplitudes were employed until failure conditions are met [19]. Nevertheless, the amplitudes of the loading conditions were, at each level, considered as random variables with a standard PDF, mean and standard deviations. Results in this section are categorized into two major section. First, results in the case of loading interruption (hence considering CFD) are presented in sections 'Loadings of Single Mean Value" and "Loadings of Multiple Mean Values", for a single and multiple mean values, respectively. Next, results of RUL in the case of uninterrupted loading conditions (i.e. no CFD consideration) are presented in sections "Single Uninterrupted Loading Application" and "Multiple Uninterrupted Loading Application".

Loadings of Single Mean Value

For the interrupted but single mean value case, three Gaussian stress PDFs were considered as an input for RUL estimation. These PDFs used the deterministic stress value result of the meshfree method as their mean value and three cases of coefficients of variation (COV) of 0.05, 0.10 and 0.15, i.e., SDs of 5, 10, and 15% of the mean value. From each Gaussian (normal) PDF, one thousand random samples were generated as shown in the histograms of Fig. 3 for each COV values.

Here, the effect of CFD due to consecutive interrupted loading applications as well as the effect of various degrees of loading uncertainties were studied. Applying the three random stress cases into the degradation model, the corresponding three RUL (remaining useful cyclic life) CDFs and histograms, during the initial cyclic life, were obtained for each loading applications, as shown in Figs. 4a, b, respectively. It can be seen from these figures that the range of stress values is directly

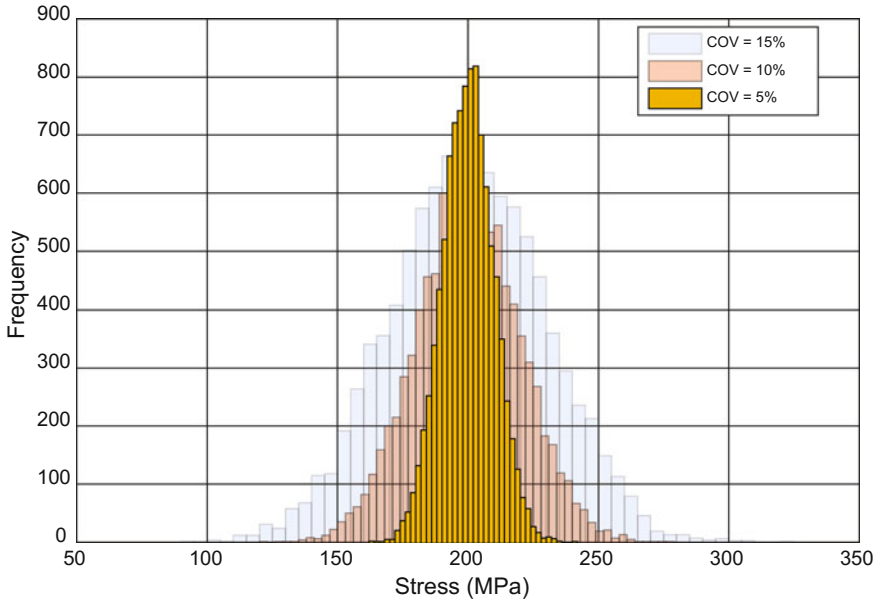


Fig. 3 Histograms of stress with a mean value of 200 MPa and COVs of 5, 10, and 15%

proportional to the magnitude of the COV, which depicts uncertainty, i.e., the wider the uncertainty in the loading condition, the more uncertain the prediction of RUL will be. It can be observed that the stress with COV = 15% results in RUL CDF that has a wider range. On the other hand, 5% COV causes the RUL CDF (depicted as COV = 5%) to have the least range. The CDFs also show the probability of getting a prespecified RUL value or less. For instance, from Fig. 4a, it can be depicted that the probability of getting RUL = 2.5×10^5 or less, according to the CDF marked COV = 5%, is about 80%. The probability of getting RUL = 3 10^5 or less, on the other hand, shows 100% using the same COV = 5% CDF, whereas it shows 88% according to the COV = 15% CDF. This shows that the variation on the input stress PDF has a significant effect on the output RUL. Figure 4b shows histograms of the RUL for three stress with different COV values. The increase in variation of the RUL histogram can be clearly seen with increase in COV of the stress.

Similarly, the histograms after the third and fifth cyclic loading depict gradual decrease in the mean value of the RUL as shown in Fig. 5 and Fig. 6, respectively. However, it can be seen that the effect of the stress due to its varying COV still remains.

RUL values of the median and 98% confidence limits versus number of cycles are depicted in Fig. 7. The figure shows three pairs of bounds of RUL that correspond to three stress PDFs with COVs of 5, 10, and 15%. Stress SD of 15% of the mean (COV of 15%) resulted in the widest bounds of RUL, whereas 5% stress COV

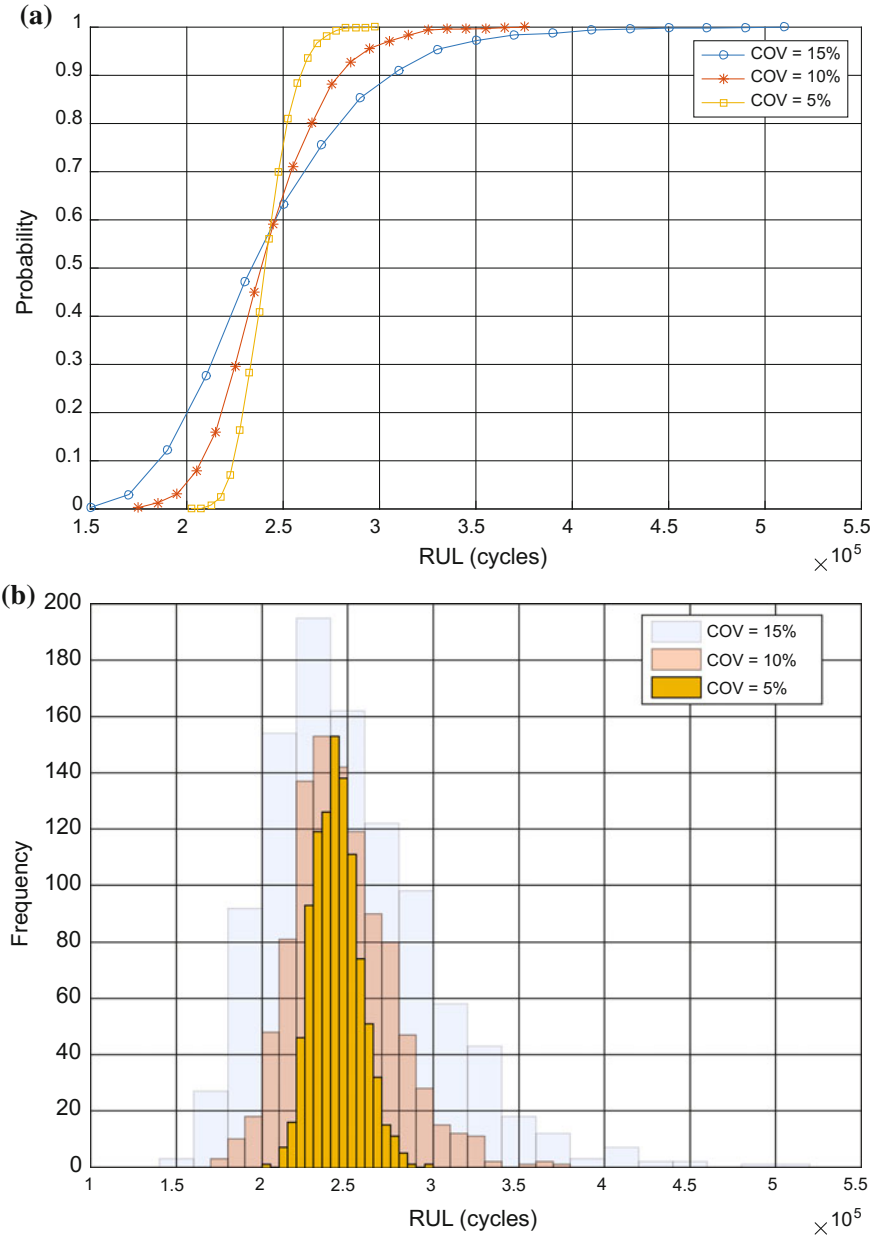


Fig. 4 Cyclic lives generated by using stresses with different COVs. **a** CDFs of RUL **b** Histograms of RUL

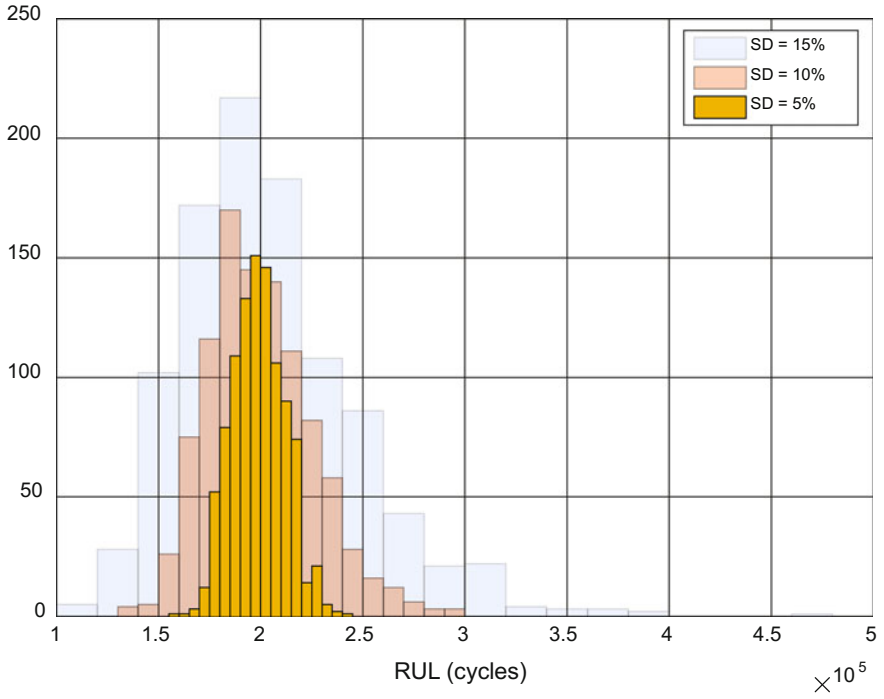


Fig. 5 Histograms of RUL at the third load application

resulted in the lowest RUL bounds. Thus, the higher the stress variation, the higher the uncertainty will be in the RUL prediction. Besides, it is interesting to note that the level of RUL uncertainty is also affected by the cycle at which the prediction is made. As can be seen in Fig. 7, the width of the bounds of all COV values, which imply the level of uncertainty, decrease towards the end of life.

Loadings of Multiple Mean Values

In this case, the meshfree deterministic stress outputs of varying, fluctuating (seven different amplitudes) loading conditions were used as mean values to generate seven consecutive random loading cases. To compare the influence of the degree of uncertainty of stress on RUL, three Gaussian PDFs of each loading case were generated using the aforementioned mean values and three COVs of 5, 10, and 15% for each case. Each random and fluctuating (cyclic) loading case was applied for 1.5×10^4 cycles.

Figure 8 shows the CDFs and histograms of the RUL for three predictions of initial life using the first mean value (i.e. 200 MPa) and three COVs (5, 10, and 15%). Note that since the stress parameters used are the same, results are similar to

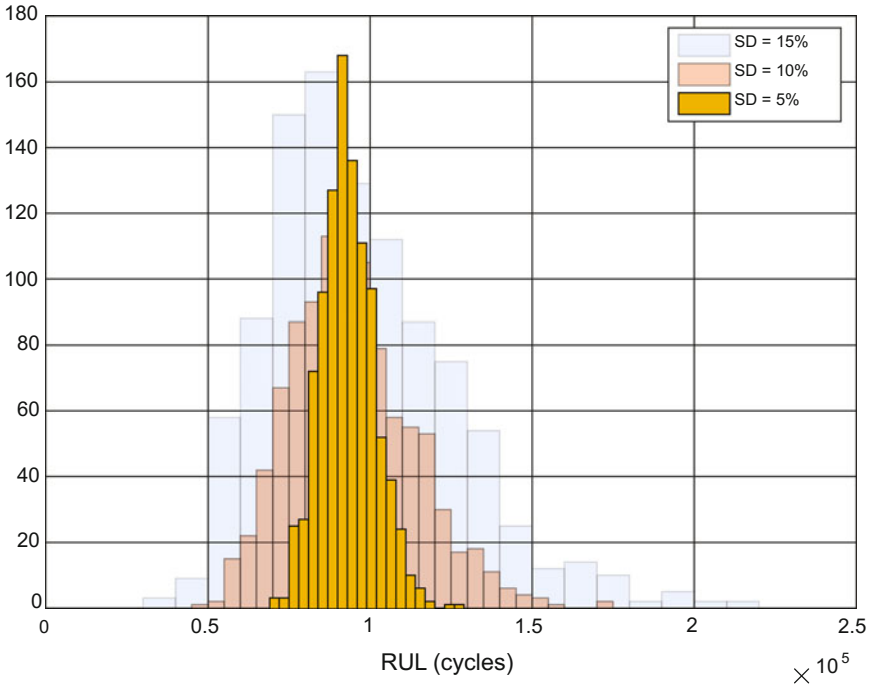


Fig. 6 Histograms of RUL at the fifth load application

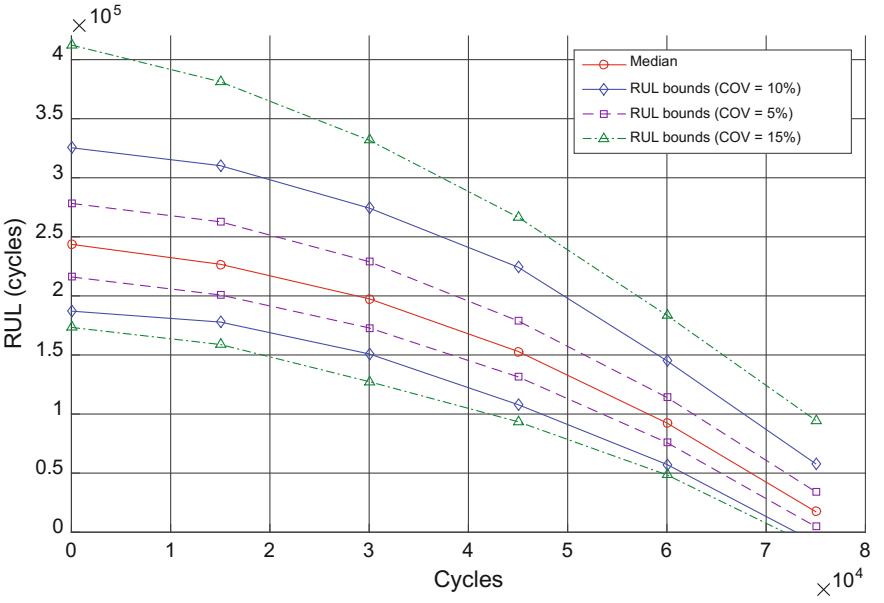


Fig. 7 RUL values showing the median and 98% bounds of confidence levels for input stresses with different COVs

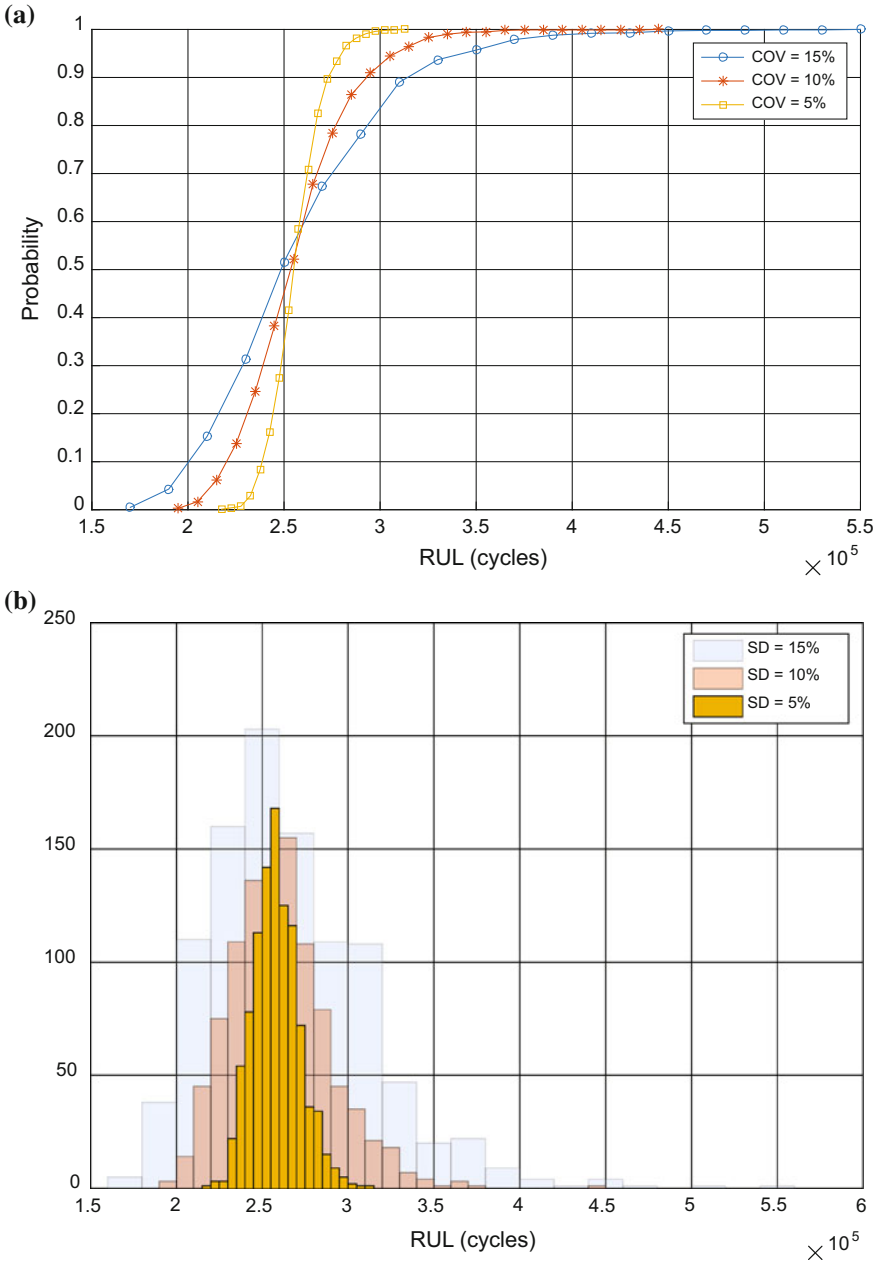


Fig. 8 RUL generated by using stresses with different COVs. **a** CDFs of RUL **b** Histograms of RUL

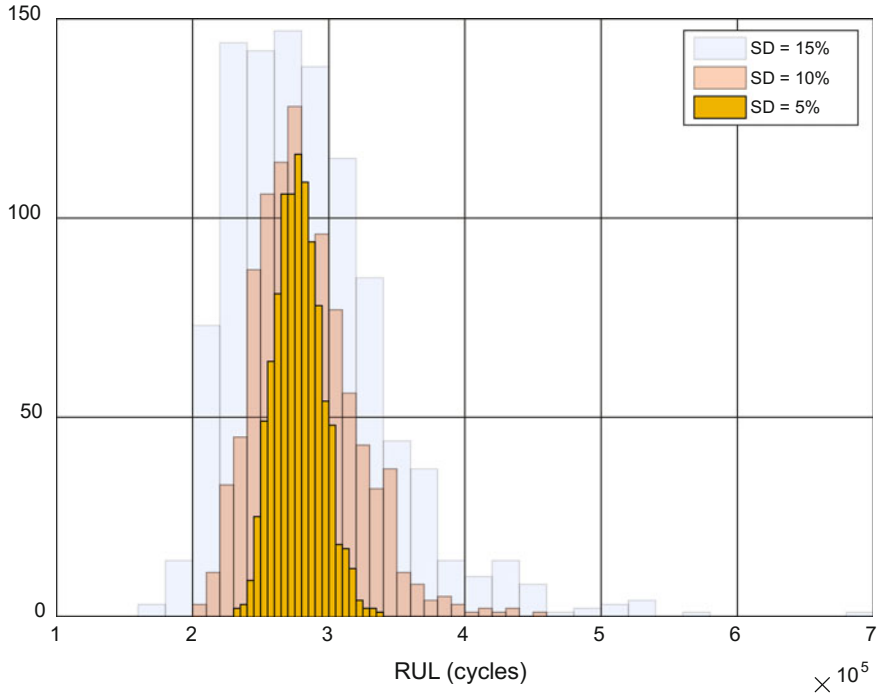


Fig. 9 Histograms of RUL at the third load application

the initial life for single mean value application shown in Fig. 4. Figures 9 and 10 show the RUL histograms of the third and sixth loading cases.

Figure 11 shows the median RUL and its three 98% bounds of confidence level corresponding to 5, 10, and 15% stress COVs. The second mean stress value used was 250 MPa. Note that the RUL is inversely related to the mean of the stresses. Therefore the reduction in the new RUL prediction emanates not only from the previous 1.5×10^4 cycles of 200 MPa stress but also from the increase in the mean value of the new stress. The third loading is reduced to 150 MPa. This makes the RUL prediction to rise despite a decrease due to the second cyclic loading. A rise in the 4th loading application is also the result of a further decrease in the mean of the stress to 70 MPa.

From the presented results above, it can be deduced that probabilistic analysis provides a better RUL prediction showing a range of possible outcomes and the probability of occurrence under seven discrete mean values of variable cyclic loading cases each applied for 1.5×10^4 cycles. Probabilistic prognostics accounts for uncertainty, and hence, encompasses a range of possible RUL values that enable the engineer to predict the status of a component with a quantified level of confidence.

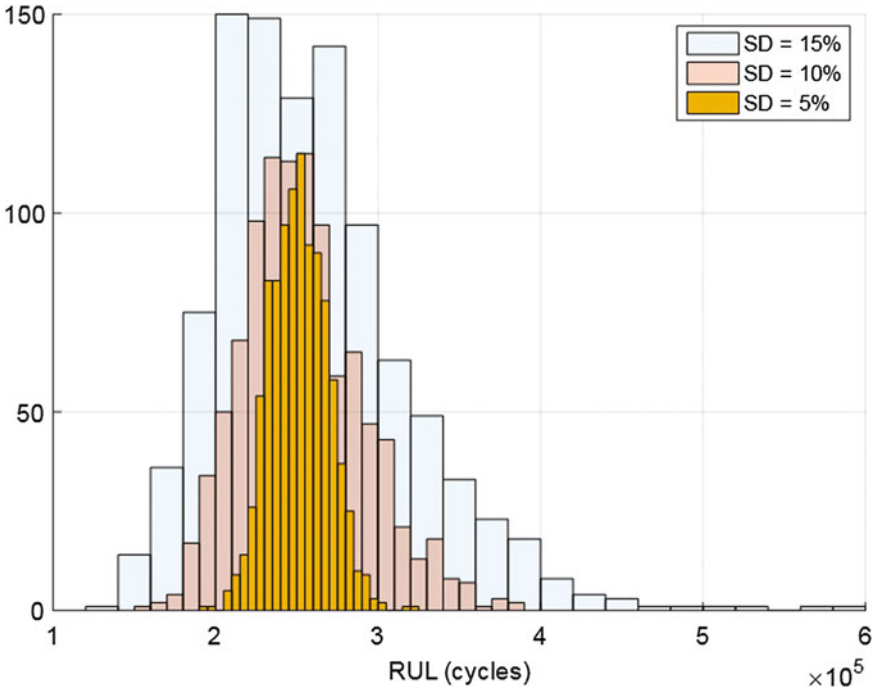


Fig. 10 Histograms of RUL at the sixth load application

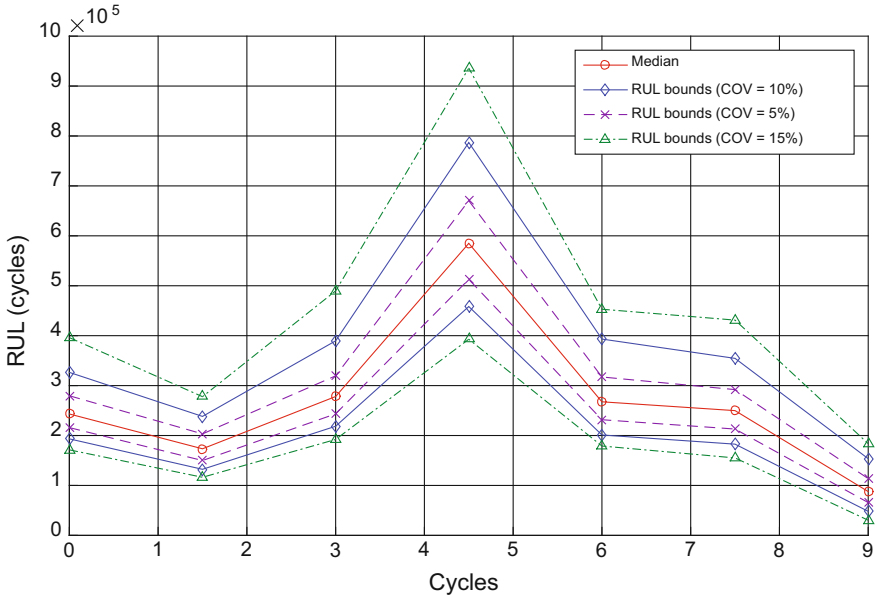


Fig. 11 Trajectory of RUL lines showing 98% confidence limits

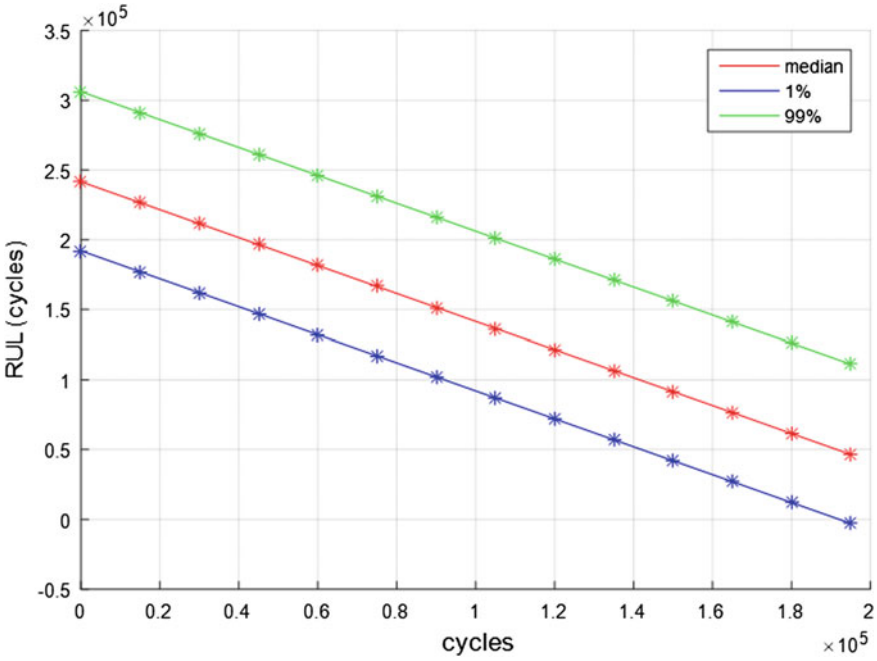


Fig. 12 RUL values showing the median and 98% bounds of confidence interval

Single Uninterrupted Loading Application

The cyclic stress was considered as a random variable with $\sigma_a \sim N(200, 20)$ MPa. In this case, CFD was not considered since the loading was uninterrupted until failure.

Figure 12 shows the RUL of the component as a function of the number of cycles with 98% bounds of RUL trajectory once the initial life estimate was obtained from the S-N plot. In the case of a single uninterrupted loading condition, i.e. without CFD, and considering only one random variable, i.e. the stress, the final end of life is the same as what was initially predicted before loading. The RUL trajectories, therefore, will have a slope of –1. Another observation from Fig. 12 is that unlike interrupted loading conditions, the variation of the RUL values does not decrease with increase in n_c. This is because, when CFD is considered, the S-N lines are continuously updated to give a better approximation. With every updated S-N line, the interval between prediction and end of life becomes less and less. Reduced interval, on the other hand, means reduced uncertainties associated with the interval of prediction up to end of life. However, in this case, one uses the original S-N line with no updates. This makes the interval between the initial RUL prediction and end of life to be large. Hence, there will not be a reduction in the uncertainties of RUL prediction as the system approaches its end of life, i.e. the 98% confidence bounds stay parallel.

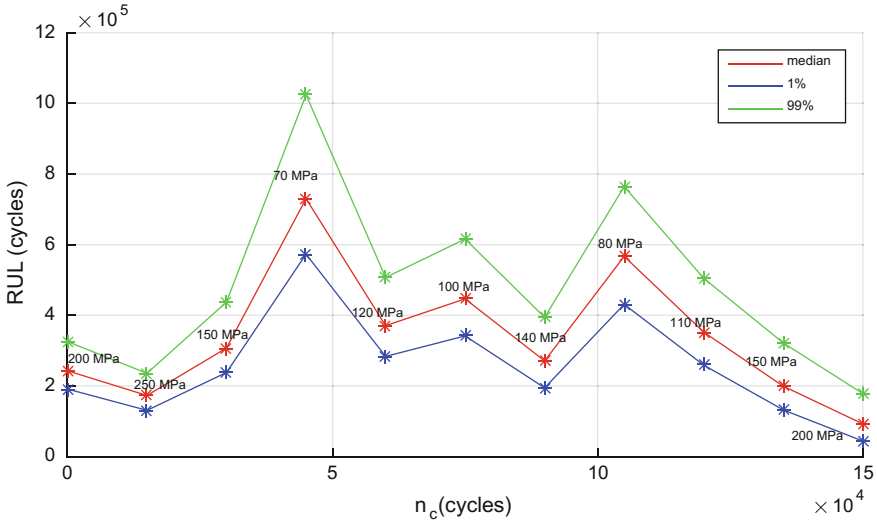


Fig. 13 RUL values showing the median and 98% bounds of confidence interval

Multiple Uninterrupted Loading Application

The case of uninterrupted but variable loading application is considered in this section. Several load stress values of the various mean were applied consecutively to the system without interruption. This means that CFD does not play a role in the life of the system.

Figure 13 shows variation of RUL values with number of cycles, depicting the median and 98% confidence limits. It can be observed that the bounds of the confidence interval do not converge. This implies that the uncertainties associated with RUL prediction do not reduce since the S-N lines used are not updated, as discussed in section "Single Uninterrupted Loading Application".

3.1.3 Conclusions

Meshfree modeling provides an efficient way of representing a physical problem that could easily be integrated with a degradation model in prognostics. The program for meshfree modeling was efficient and convenient as it is easy to alter values of parameters for model performance and accuracy.

A framework for computing the RUL of a component was developed. The framework is capable of predicting RUL both deterministically and probabilistically. The framework was used to perform RUL prediction of a cantilever beam subjected to fatigue loading. Deterministic analyses of a cantilever beam using the framework under constant and variable loading depict that CFD plays a significant role in RUL predictions of a component undergoing cyclic loadings. It was also

shown that the amplitude of the stress can change the trajectory of the RUL and the component may fail in less number of cycles than what was predicted during initial loading. Therefore, it is crucial to continuously estimate the magnitude of future loading and to quantify its uncertainty to accurately determine the path of the RUL lines. Future loading uncertainty is especially manifested in wind turbine gearboxes, which are subjected to dynamic loading. Here, future loading is uncertain since future wind speed is uncertain. Probabilistic prognostics is, thus, essential to show possible RUL values so that a decision will be made to deter failure and avoid downtime. Loading uncertainties were considered and quantified to provide reliable RUL prediction. Probabilistic analysis of the cantilever beam was performed by quantifying loading uncertainties to provide reliable RUL prediction. Results show the bounds of possible RUL for 98% interval of different degrees of uncertainties of the input stress. It was shown that the variation in RUL is highly affected by the uncertainty in the input stress. Therefore, it is important to accurately quantify uncertainties of random variables for reliable RUL prediction.

The framework integrates efficient meshfree modeling and damage estimation. By capturing these uncertainties, the framework provides results that depict probabilities of RUL in the desired future time. Future loading conditions could be continuously updated when they are available for improved RUL prediction. The computational framework aids in decision making and fault mitigation.

Future work includes construction of a reliable and comprehensive prognostics framework for remaining useful life prediction using particle filter considering various uncertainties such as loading, modeling, measurement, and material parameter uncertainties. Crack propagations in a gear teeth of a wind turbine gearbox will be studied using enhanced prognostics framework.

References

1. A.K. Garga, K.T. McClintic, R.L. Campbell, C.-C. Yang, M.S. Lebold, T.A. Hay, C.S. Byington, Hybrid reasoning for prognostic learning in CBM systems, in *Proceedings of the 2001 IEEE Aerospace Conference*, **6**, 2957–2969 (2001)
2. G.W. Bartram, System health diagnosis and prognosis using dynamic bayesian networks (2013)
3. S. Sankararaman, Significance, interpretation, and quantification of uncertainty in prognostics and remaining useful life prediction. Mech. Syst. Signal Process. **52–53**(1), 228–247 (2015)
4. M.G. Pecht, Prognostics and health management, in *Solid State Lighting Reliability: Components to Systems*, ed. by W.D. van Driel, X.J. Fan (Springer New York, New York, NY, 2013), pp. 373–393
5. S. Sankararaman, K. Goebel, An uncertainty quantification framework for prognostics and condition-based monitoring, in *16th AIAA Non-Deterministic Approaches Conference* (2014), pp. 1–9
6. G. Bartram, S. Mahadevan, Probabilistic prognosis with dynamic bayesian networks. Int. J. Progn. Health Manag. **6**(SP4), 1–23 (2015)
7. D. An, J.H. Choi, N.H. Kim, Prognostics 101: a tutorial for particle filter-based prognostics algorithm using Matlab, Reliab. Eng. Syst. Saf. **115**, 161–169 (2013)

8. J. Liu, A. Saxena, K. Goebel, B. Saha, W. Wang, *An Adaptive Recurrent Neural Network for Remaining Useful Life Prediction of Lithium-ion Batteries* (2010), pp. 0–9
9. B. Saha, K. Goebel, J. Christophersen, Comparison of prognostic algorithms for estimating remaining useful life of batteries. Trans. Inst. Meas. Control **31**(3–4), 293–308 (2009)
10. X.S. Si, W. Wang, C.H. Hu, D.H. Zhou, Remaining useful life estimation—a review on the statistical data driven approaches. Eur. J. Oper. Res. **213**(1), 1–14 (2011)
11. D. An, N.H. Kim, J.H. Choi, Statistical aspects in neural network for the purpose of prognostics. J. Mech. Sci. Technol. **29**(4), 1369–1375 (2015)
12. D. An, N.H. Kim, J.H. Choi, Practical options for selecting data-driven or physics-based prognostics algorithms with reviews. Reliab. Eng. Syst. Saf. **133**, 223–236 (2015)
13. F. Zhao, Z. Tian, Y. Zeng, A stochastic collocation approach for efficient integrated gear health prognosis. Mech. Syst. Signal Process. **39**(1–2), 372–387 (2013)
14. M. Daigle, A. Saxena, and K. Goebel, An efficient deterministic approach to model-based prediction uncertainty estimation, in *Annual conference of the prognostics*, 2012, 1–10
15. F. Zhao, Z. Tian, Y. Zeng, Uncertainty quantification in gear remaining useful life prediction through an integrated prognostics method. IEEE Trans. Reliab. **62**(1), 146–159 (2013)
16. S. Sankararaman, K. Goebel, Why is the remaining useful life prediction uncertain ? in *Annual Conference of the Prognostics and Health Management Society 2013* (2013), pp. 1–13
17. S. Sankararaman, M.J. Daigle, K. Goebel, Uncertainty quantification in remaining useful life prediction using first-order reliability methods. IEEE Trans. Reliab. **63**(2), 1–17 (2014)
18. J.Z. Sikorska, M. Hodkiewicz, L. Ma, Prognostic modelling options for remaining useful life estimation by industry. Mech. Syst. Signal Process. **25**(5), 1803–1836 (2011)
19. H.B. Endeshaw, F.M. Alemayehu, S. Ekwaro-Osire, J.P. Dias, A probabilistic model-based prognostics using meshfree modeling: a case study on fatigue life of a cantilever beam, in *Proceedings of the ASME 2016 International Mechanical Engineering Congress and Exposition* (2016), pp. 1–13
20. F. Chaari, T. Fakhfakh, M. Haddar, Analytical modelling of spur gear tooth crack and influence on gearmesh stiffness. Eur. J. Mech. A/Solids **28**(3), 461–468 (2009)
21. B.N. Rao, S. Rahman, An efficient meshless method for fracture analysis of cracks. Comput. Mech. **26**(4), 398–408 (2000)
22. F. Liu, R.I. Borja, A contact algorithm for frictional crack propagation with the extended finite element method. Int. J. Numer. Methods Eng. **76**, 1489–1512 (2008)
23. V.P. Nguyen, T. Rabczuk, S. Bordas, M. Duflot, Meshless methods: a review and computer implementation aspects. Math. Comput. Simul. **79** (3), 763–813 (2008)
24. G.-R. Liu, Y.-T. Gu, *An Introduction to Meshfree Methods and their Programming* (Springer, Dordrecht, The Netherlands, 2005)
25. S. Bordas, P.V. Nguyen, C. Dunant, H. Nguyen-dang, A. Guidoum, An extended finite element library. Int. J. Numer. Methods Eng. **41**, 1–33 (2006)
26. S. Yixiu, L. Yazhi, A simple and efficient X-FEM approach for non-planar fatigue crack propagation. Procedia Struct. Integr. **2**, 2550–2557 (2016)
27. G.R. Liu, *Meshfree Methods: Moving Beyond the Finite Element Method*, 2nd edn. (CRC Press, Boca Raton, 2010)
28. Y.P. Chen, A. Eskandarian, M. Oskard, J.D. Lee, Meshless simulation of crack propagation in multiphase materials. Theor. Appl. Fract. Mech. **45**(1), 13–17 (2006)
29. T. Belytschko, L. Gu, Y.Y. Lu, Fracture and crack growth by element free Galerkin methods. Model. Simul. Mater. Sci. Eng. **2**(3A), 519–534 (1999)
30. S.N. Atluri, T. Zhu, A new Meshless Local Petrov-Galerkin (MLPG) approach in computational mechanics. Comput. Mech. **22**(2), 117–127 (1998)
31. E. Zio, G. Peloni, Particle filtering prognostic estimation of the remaining useful life of nonlinear components. Reliab. Eng. Syst. Saf. **96**(3), 403–409 (2011)
32. E. Phelps, P. Willett, T. Kirubarajan, C. Brideau, Predicting time to failure using the IMM and excitable tests. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **37**(5), 630–642 (2007)
33. A. Ray, S. Tangirala, Stochastic modeling of fatigue crack dynamics for on-line failure prognostics. IEEE Trans. Control Syst. Technol. **4**(4), 443–451 (1996)

34. D.C. Swanson, J. Michael Spencer, S.H. Arzoumanian, Prognostic modelling of crack growth in a tensioned steel band. Mech. Syst. Signal Process. **14**(5), 789–803 (2000)
35. M.J. Daigle, K. Goebel, A model-based prognostics approach applied to pneumatic valves. Int. J. Progn. Health Manag. **2**, 1–16 (2011)
36. M. Daigle, K. Goebel, A comparison of filter-based approaches for model-based prognostics (2012)
37. M. Orchard, G. Vachtsevanos, A particle filtering approach for on-line fault diagnosis and failure prognosis. Meas. Control **31**(3–4), 1–18 (2007)
38. R. Dupuis, Application of oil debris monitoring for wind turbine gearbox prognostics and health management, in *Annual Conference of the Prognostics and Health Management Society* (2010)
39. M.S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans. Signal Process. **50**(2), 174–188 (2002)
40. S.J. Lee, G. Zi, S. Mun, J.S. Kong, J.H. Choi, Probabilistic prognosis of fatigue crack growth for asphalt concretes. Eng. Fract. Mech. **141**, 212–229 (2015)
41. M.E. Orchard, G.J. Vachtsevanos, A particle-filtering approach for on-line fault diagnosis and failure prognosis. Trans. Inst. Meas. Control **31**(3–4), 221–246 (2009)
42. M. Orchard, G. Kacprzynski, K. Goebel, B. Saha, G. Vachtsevanos, Advances in uncertainty representation and management for particle filtering applied to prognostics, in *2008 International Conference on Prognostics and Health Management, PHM 2008* (2008)
43. R. Budynas, K. Nisbett, *Shigley's Mechanical Engineering Design*, 10th edn. (McGraw-Hill, New York, NY, 2015)
44. S. Ekwaro-Osire, H. B. Endeshaw, D. H. Pham, and F. M. Alemayehu, Uncertainty in remaining useful life prediction, in *23rd ABCM International Congress of Mechanical Engineering*, 2015
45. A. Haldar, S. Mahadevan, *Probability, Reliability and Statistical Methods in Engineering Design*.(John Wiley & Sons, Inc., 2000)
46. F.M. Alemayehu, S. Ekwaro-Osire, Uncertainty considerations in the dynamic loading and failure of spur gear pairs. J. Mech. Des. **135**(8), 84501-1–7 (2013)
47. F.M. Alemayehu, S. Ekwaro-Osire, Loading and design parameter uncertainty in the dynamics and performance of high-speed-parallel-helical stage of a wind turbine gearbox. J. Mech. Des., **136**(9), 91002–1–91002–13 (2014)
48. C.J. Li, H. Lee, Gear fatigue crack prognosis using embedded model, gear dynamic model and fracture mechanics. Mech. Syst. Signal Process. **19**(4), 836–846 (2005)
49. F.M. Alemayehu, S. Ekwaro-Osire, Probabilistic performance of helical compound planetary system in wind turbine. J. Comput. Nonlinear Dyn. **10**(4), 41003 (2015)
50. Z. Tian, T. Jin, B. Wu, F. Ding, Condition based maintenance optimization for wind power generation systems under continuous monitoring. Renew. Energy **36**(5), 1502–1509 (2011)
51. Z. Tian, An artificial neural network method for remaining useful life prediction of equipment subject to condition monitoring. J. Intell. Manuf. **23**(2), 227–237 (2009)

Cognitive Architectures for Prognostic Health Management

James A. Crowder and John N. Carbone

Abstract In the real-time battlefield arena, situational awareness becomes critical to making the right decisions and achieving the overall goals for the system. The key to Situational Awareness is not simply collecting and disseminating data, but it is actually getting the right information to the right users at the right time. In ground processing systems, various sensors, spacecraft, and other data sources gather and generate data different relevant contexts. What is required is an Integrated System Health Management (ISHM) processing architecture that allows users to turn the data into meaningful information and to reason about that information in a context relative to the user at that time, and to update the information real-time as the situation changes. In short, it is imperative that the information processing environment be efficient, timely, and accurate. This chapter will describe an Intelligent Information Agent processing environment which allows data to be processed into relevant and actionable knowledge using high-fidelity knowledge relativity threads. Based on the technologies described above, the situational management and recombinant knowledge assimilation process built on top of a multi-dimensional high-fidelity knowledge relationship store is one of the most innovative components of this Prognostic Health Management (PHM) system. Utilizing the Artificial Cognitive Neural Framework (ACNF), it can provide real-time processing and display of dynamic, situational awareness information.

Keywords Prognostic health management · Cognitive systems · Intelligent agents · Integrated system health management

J.A. Crowder (✉) · J.N. Carbone
Raytheon Intelligence and Information Systems Division, 16800 E. Centretech Parkway, Aurora, CO 80011, USA
e-mail: jacrowder@raytheon.com

S. Ekwaro-Osire et al. (eds.), *Probabilistic Prognostics and Health Management of Energy Systems*, DOI 10.1007/978-3-319-55852-3_6

1 Introduction

Even in modern architectures, true Integrated System Health Management (ISHM) and Prognostic Health Management (PHM) that drives situational awareness is difficult because the enterprise system has to become more aware, more flexible, and more agile than ever before. Information gathering, processing, and analyzing must be done continually to keep track of current trends in the context of the current situations, both local and overall, and to provide timely and accurate knowledge to allow the users to anticipate and respond to what is happening in a changing environment. To achieve the combination of awareness, flexibility, and agility means supporting dynamic and flexible processes that adapt as situations change. This is possible with learning and evolving Intelligent Information Agents, such as those described here.

Data Steward Agents will support growing volumes of data and allow Reasoner Agents to produce accurate and relevant metrics about past, current, and future situations (prognostics). Through inter-agent communication, they provide control and visibility into the entire ground processing enterprise. This is made possible by integrating the processing environment into the flexible, distributed, Service Oriented Architecture (SOA) that enables secure collaboration, advanced information management, dynamic system updates, and customer, rule-based processes (Advisor Agents).

The inter-agent communication allows shared awareness which, in turn, enables faster operations and more effective information analysis and transfer, providing users with an enhanced visualization of the overall constellation and situational awareness across the ground processing system's Enterprise Infrastructure. This Intelligent Agent-based system can deal with massive amounts of information to levels of accuracy, timeliness, and quality never before possible.

Even applications that deal with object-oriented technologies fail to achieve the goals of awareness, flexibility, and agility because their processes are hard coded into the applications. The flexible, learning, and adapting Intelligent Software Agents can adapt, collaborate, and provide the increased flexibility required in a growing and changing signal/source environment [1].

2 Integrated System Health Management

Following the evolution of diagnostic systems, prognostic initiatives started to be introduced in order to try to take advantage of the maintenance planning and logistics benefits. However, the early prognostic initiatives often were driven by in-field failures that resulted in critical safety or high-cost failures, and thus retrofitted technology was hard to implement and costly to develop. Hence, diagnostic and prognostic system developers found the need to analyze and describe the benefits associated with reducing in-field failures and their positive impact on

safety, reliability, and overall lifecycle-cost reduction. This has led to many cost-benefit analyses and ensuing discussions and presentations to engineering management about why the diagnostic and prognostic technologies need to be included in the design process of the system and not simply an afterthought once field failures occur. This had lead us to the point where many complex vehicle/system designs, like DD(X), GPS, and various weapon systems are now developing ''designed in'' health management technologies that can be implemented within the Integrated Maintenance & Logistics and supports the system throughout its lifetime. This ''designed in'' approach to health management is performed with the hardware/software design itself and also acts as the process for system validation and managing inevitable changes from in-field experiences and evaluating system design tradeoffs, as shown in Fig. 1 [2].

Realizing such an approach involves synergistic deployments of component health monitoring technologies as well as integrated reasoning capabilities for the interpretation of fault-detect outputs. Further, it will involve the introduction of learning technologies to support the continuous improvement of the knowledge enabling these reasoning capabilities. Finally, it will involve organizing these elements into a maintenance and logistics architecture that governs integration and interoperation within the system, between its on-board elements and their ground-based support functions, and between the health management system and external maintenance and operation functions. Here we present and discuss the required prognostic functions of an Integrated Health Management System that, if applied correctly, can directly affect the operations and maintenance of the equipment and positively affect the lifecycle costs.

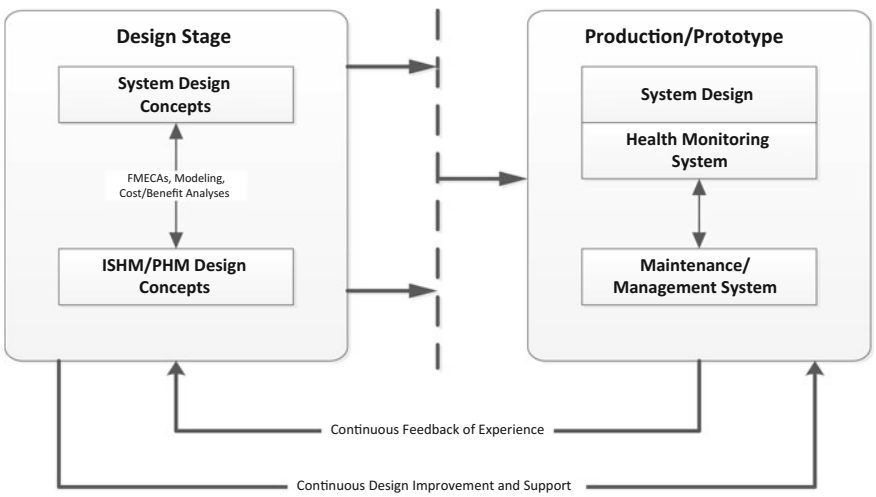


Fig. 1 The "Design in" approach to system health management

2.1 Prognostics and Diagnostics

A comprehensive health management system philosophy integrates the results from the monitoring sensors all the way through to the reasoning software that provides decision support for optimal use of maintenance resources. A core component of this strategy is based on the ability to (1) accurately predict the onset of impending faults/failures or remaining useful life (RUL) of critical components and (2) quickly and efficiently isolate the root cause of failures once failure effects have been observed. In this sense, if fault/failure predictions can be made, the allocation of replacement parts or refurbishment actions can be scheduled in an optimal fashion to reduce the overall operational and maintenance logistic footprints. From the fault isolation perspective, maximizing system availability and minimizing downtime through more efficient troubleshooting efforts is the primary objective.

In addition, the diagnostic and prognostic technologies require an integrated maturation environment for assessing and validating PHM system accuracy at all levels in the system hierarchy. Developing and maintaining such an environment will allow for inaccuracies to be quantified at every level in the system hierarchy and then be assessed automatically up through the health management system architecture. The final results reported from the system-level reasoners and decision support is a direct result of the individual results reported from these various levels when propagated throughout the process. Hence, an approach for assessing the overall PHM system accuracy is to quantify the associated uncertainties at each of the individual levels, as illustrated in Fig. 2, and build up the accumulated inaccuracies as information is passed up the system architecture. This type of hierarchical verification and validation (V&V) and maturation process will be able to provide the capability to assess diagnostic and prognostic technologies in terms of their ability to detect subsystem faults, to diagnose the root cause of the faults, to predict the RUL of the faulty component, and to assess the decision-support reasoner algorithms. Specific metrics include accuracy, false-alarm rates, reliability, sensitivity, stability, economic cost/benefit, and robustness, just to name a few. Cost-effective implementation of a diagnostic or prognostic system will vary depending on the design maturity and operational/logistics environment of the monitored equipment. However, one common element to successful implementation is feedback. As components or Line Replaceable Units (LRUs) are removed from service, disassembly inspections must be performed to assess the accuracy of the diagnostic and prognostic system decisions [3]. Based on this feedback, system software and warning/alarm limits should be optimized until desired system accuracy and warning intervals are achieved. In addition, selected examples of degraded component parts should be retained for testing that can better define failure progression intervals.

A systems-oriented approach to prognostics requires that the failure detection and inspection-based methods be augmented with forecasting of parts degradation, mission criticality and decision support. Such prognostics must deal not only with the condition of individual components, but also the impact of this condition on the

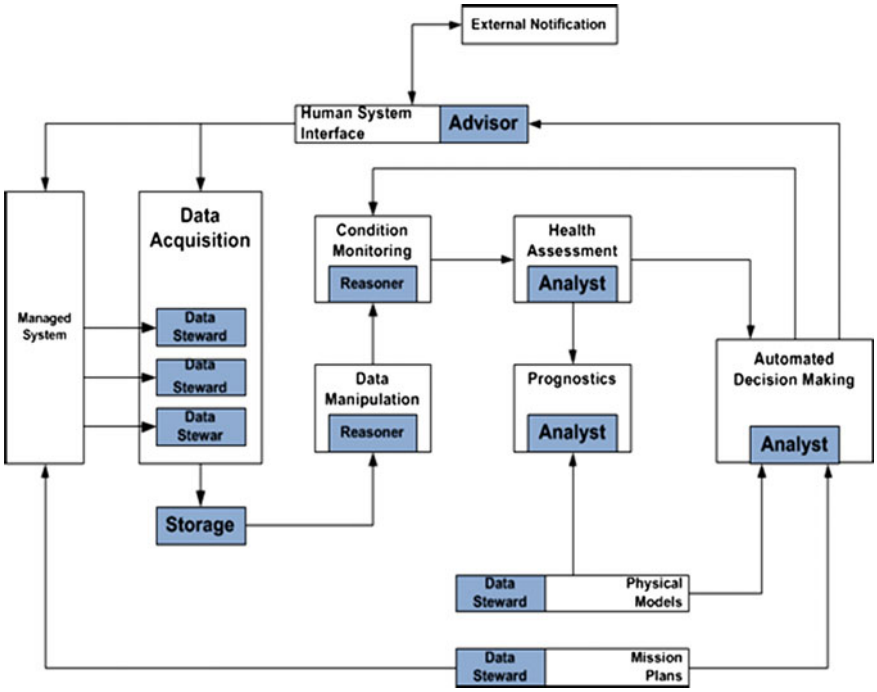


Fig. 2 Functional layers in an ISHM system

mission-readiness and the ability to take appropriate actions [4]. However, such a continuous health management system must be carefully engineered at every stage of a system design, operation and maintenance. Figure 2 illustrates the overall ISHM/PHM process which includes modeling, sensing, diagnosis, inference and prediction (prognostics), learning, and updating. The two most important steps in this process are (1) fault detection and diagnosis and (2) prognostic reasoning (prediction):

3 Fault Detection and Diagnostic Reasoning

This determines if a component/subsystem/system has moved away (degraded) from nominal operating parameters, along a known path, to a point where component performance may be compromised. Novelty detection determines if the component has moved away from what is considered acceptable nominal operations and away from all known fault health (diagnostics as defined above) propagation paths [1].

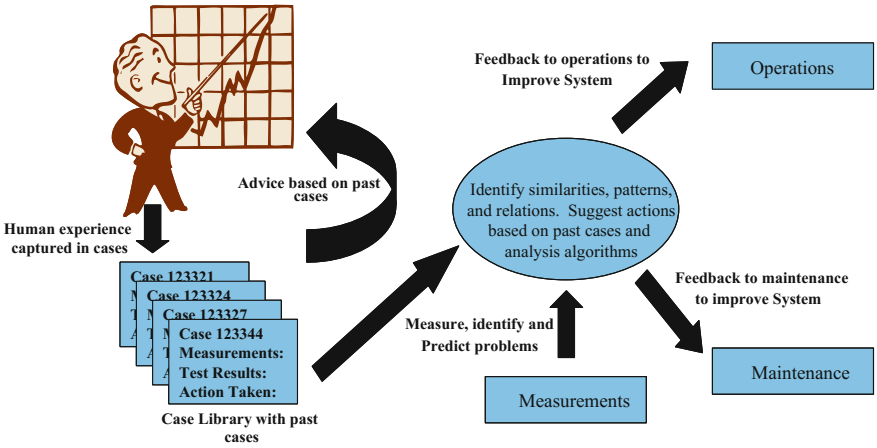


Fig. 3 Prognostic process utilizing I^2As

4 Prognostic Reasoners

The purpose of reasoners is the assessment of the component's current health and a prediction of the component's future health, or RUL. There are two variations of the prediction problem. The first prediction type may have just a short horizon time—is the component good to fly the next mission? The second type is to predict how much time we have before a particular fault will occur and, by extension, how much time we have before we should replace it, or it may be even longer term—tell me when to schedule removal of an engine for overhaul.

Accurate prognosis is a requirement for implementing Prognostic Health Management (PHM). The creation of a prognostic algorithm is a challenging problem. There are several areas that must be addressed in order to develop a prognostic reasoner that achieves a given level of performance. Figure 3 illustrates the prognostic process utilizing Intelligent Information Agents (I^2As).

5 The Prognostic Process

The prognostics component (utilizing Analyst Agents) provides specific information to the Advisor Agents about the system's state of health, status, RUL, confidence and recommendations. A graphical representation of the inputs and outputs to the Prognostics Analyst Agent is illustrated in Fig. 4. The description of the inputs and outputs is given below in Fig. 5.

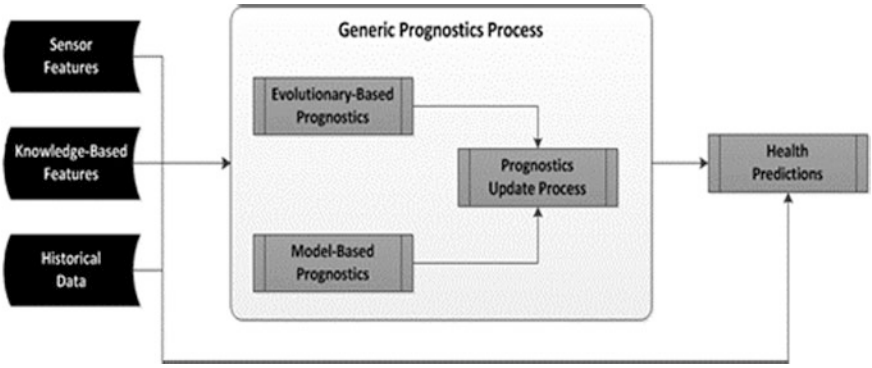


Fig. 4 Prognostic analyst agent processing

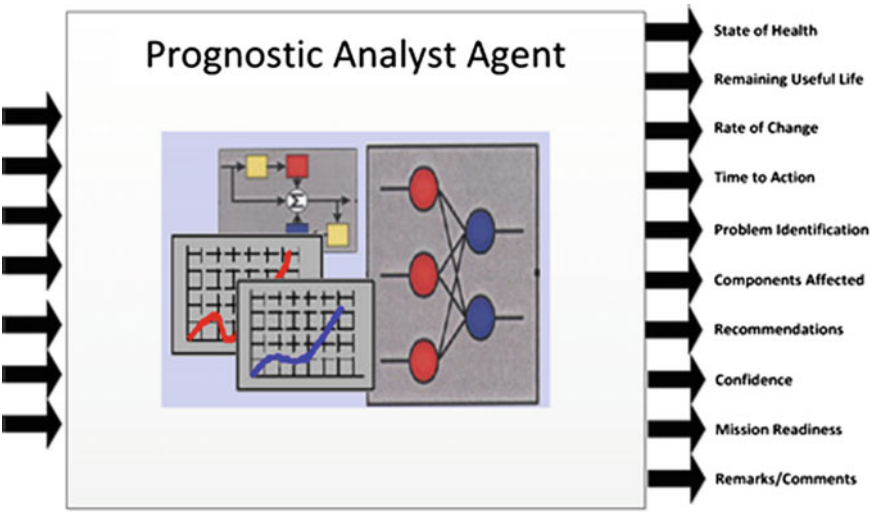


Fig. 5 Prognostic analysis inputs and outputs

6 Automated Decision Making

The Automated Decision Making component utilizes Advisor Agents that acquire data primarily from Diagnostic and Prognostic Analyst Agents. The primary function of the Automated Decision Making Advisor Agents is to provide recommended actions and alternatives and the implications of each recommended action. Recommendations may include maintenance action schedules, modifying the operational configuration of assets and equipment in order to accomplish mission objectives, or modifying mission profiles to allow mission completion. The Automated Decision Making Advisor Agents take into account operational history

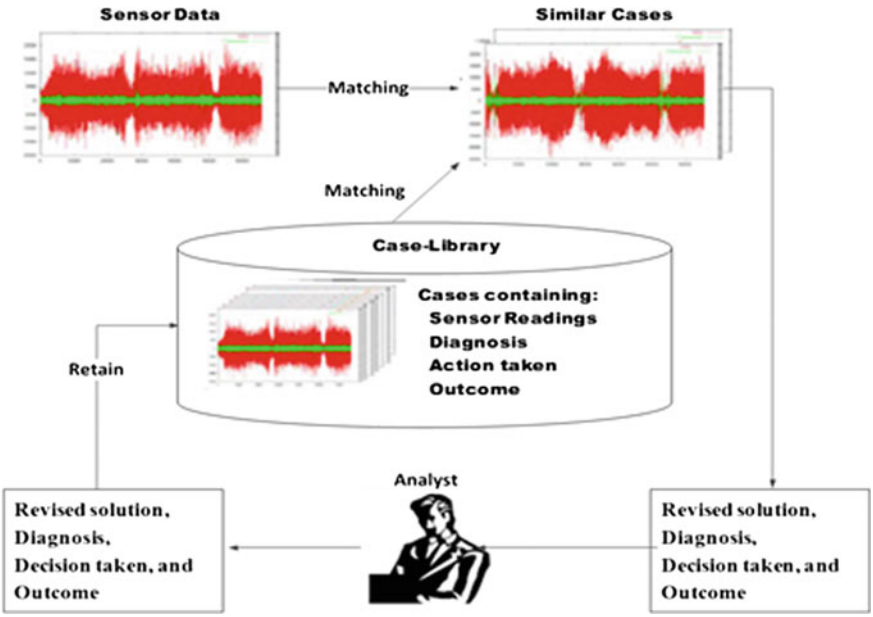


Fig. 6 PHM decision making process

(including usage and maintenance), current and future mission profiles, high-level unit objectives, and resource constraints. This is always a Human-in-the-Loop to assess the correctness of major decisions and adjust the decision process. Figure 6 illustrates the PHM Decision Making Process.

7 Prognostic Learning Using High-Fidelity Relationships

A major problem with optimization of system prognostics is that systems are not built to remember what they do so; it makes it difficult for a system to learn and improve. This is related to capacity constraints which drive the practical applications which, in turn, drives an inability to think about how systems could be constructed so that they could heal themselves. A key attribute to high fidelity prognostic system management, learning and prediction includes the implementation of Recombinant Knowledge Assimilation (RNA) frameworks [2] and the utilization of memory threads created using high-fidelity relationship mappings. Biomedical and health care systems must access vast stores of research and clinical information in their attempts to gather information about a particular topic/disease/condition, and often searches yield thousands of possible sources, most of which are not relevant to the context that the medical professional is seeking. Over a period of time, this recursive refinement of knowledge and context

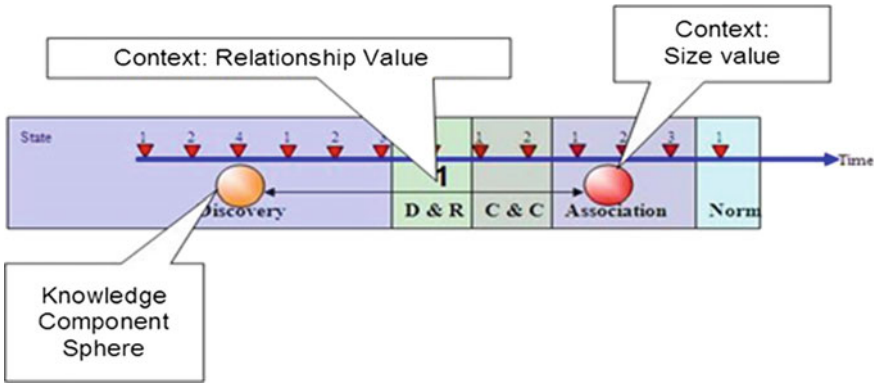


Fig. 7 Knowledge relativity thread

occurs as user cognitive system interaction where the granularity of information content results are analyzed, followed by the formation of relationships and related dependencies. Ultimately, the knowledge attained from assimilating the information content reaches a threshold of decreased ambiguity and level of understanding which acts as a catalyst for decision-making, subsequently followed by actionable activity or the realization that a research objective has been attained. Therefore, a system employing knowledge threads holding relationship mappings are critical for prognostic evaluation and proper resolution of system issues of all types (e.g., power optimization, processing improvement etc.). Figure 7 illustrates the Knowledge Relativity Thread (KRT) concept.

In its simplest form, a KRT provides a state defining relationship value context, just as in the brain, but with the ability to capture and represent relationships multi-dimensionally and to a specific context.

An example of relationship derivation is the KRT's use of an abstraction of Newton's Law of Gravitation as an analogy for representing relationships between two objects of knowledge using context, is written as Eq. (1) shown below, which describes the components of the formula for representing relationships between two objects of knowledge using context:

$$A = B\frac{(I_1 I_2)}{c^2} \tag{1}$$

where

A is the magnitude of the attractive force between the two objects of knowledge,
B is a balance variable,
I_1 is the importance measure of the first object of knowledge,
I_2 is the importance measure of the second object of knowledge, and
c is the closeness between the two objects of knowledge

Other mathematical constructs can represent value relationships (e.g. Shannon and Renyi entropy). For optimizing systems based upon prognostics, capturing some form of value relationship construct is required.

8 Prognostic Technologies: Intelligent Information Agents (I^2As)

The I^2A architecture is a framework for constructing a hybrid system of Intelligent Information Software Agents. This provides a productivity toolkit for adding intelligent software agent functions to applications and modern architectural frameworks. This provides the constructs for building multi-agent intelligent autonomic systems. This includes the framework for providing business rules and policies for run-time systems, including an autonomic computing core technology within a multi-agent infrastructure. Figure 8 illustrates the Intelligent Information Agents for the I^2A framework.

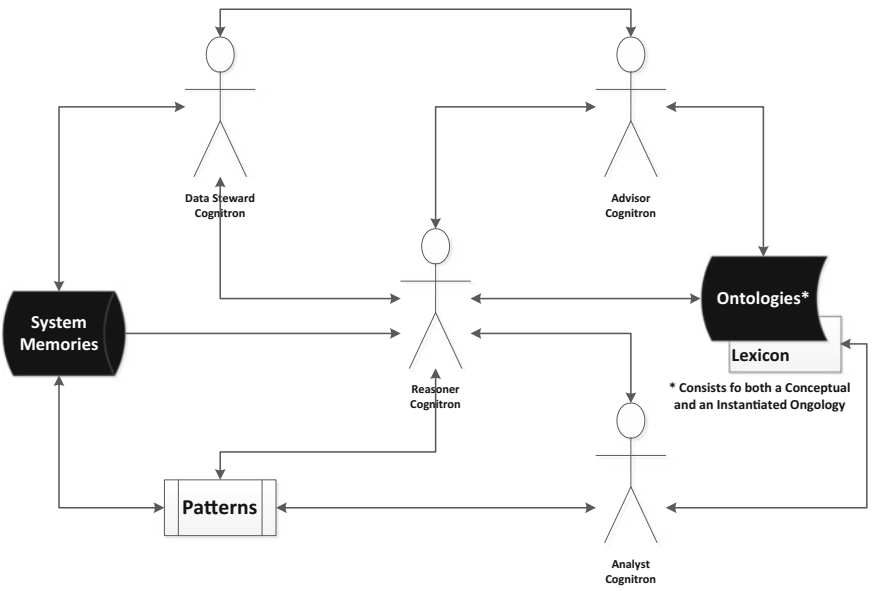


Fig. 8 The I^2A framework

9 The I²A Framework

The I²A hybrid computing architecture uses genetic, neural-network and fuzzy logic that are used to integrate diverse sources of information, associate events in the data and make observations [5]. When combined with a dialectic search, the application of hybrid computing promises to revolutionize information processing. The dialectic search seeks answers to questions that require interplay between doubt and belief, where our knowledge is understood to be fallible. This 'playfulness' is key to hunting in information and is explained in more detail in the section that address the Dialectic Argument Structure. Figures 9, 10, and 11 further explain this. The dialectic search avoids the problems associated with analytic methods and word searches. In its place, information is used to develop and assess hypotheses seeded by a domain expert. This is achieved using I²A that augments human reason by learning from the expert how to argue and develop a hypothesis. Using Franklin and Graesser's definition for a software agent, we would define the I²A as: an autonomous agent situated in and part of the information ecosystem, comprehending its environment and acting upon it over time, in pursuit of its own agenda, so as to affect what it comprehends in the future. The I²A have certain abilities that distinguish it from software objects and programs and provide it with the intelligence it needs to mimic human reasoning [5].

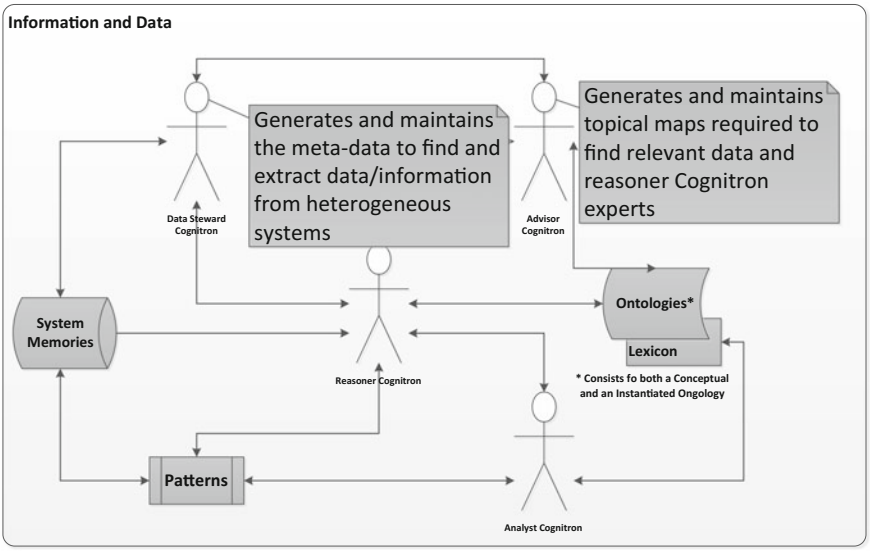


Fig. 9 The data steward and advisor agents

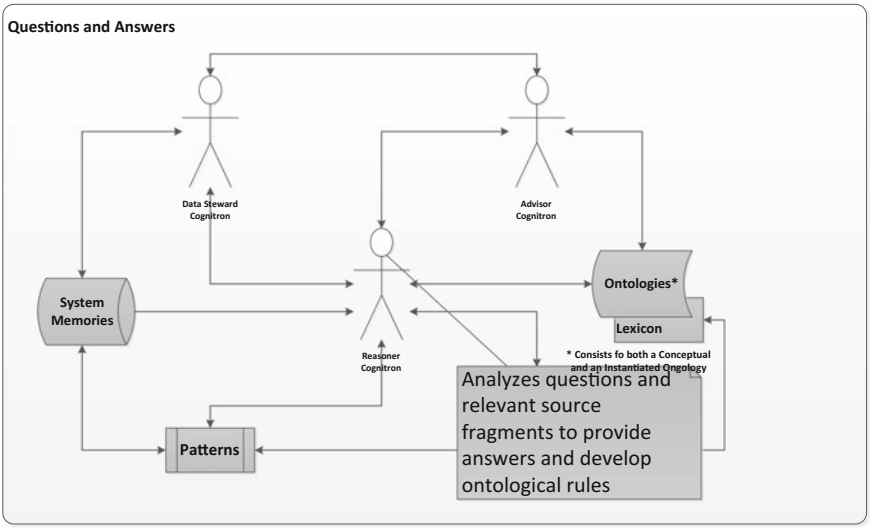


Fig. 10 The reasoner agent

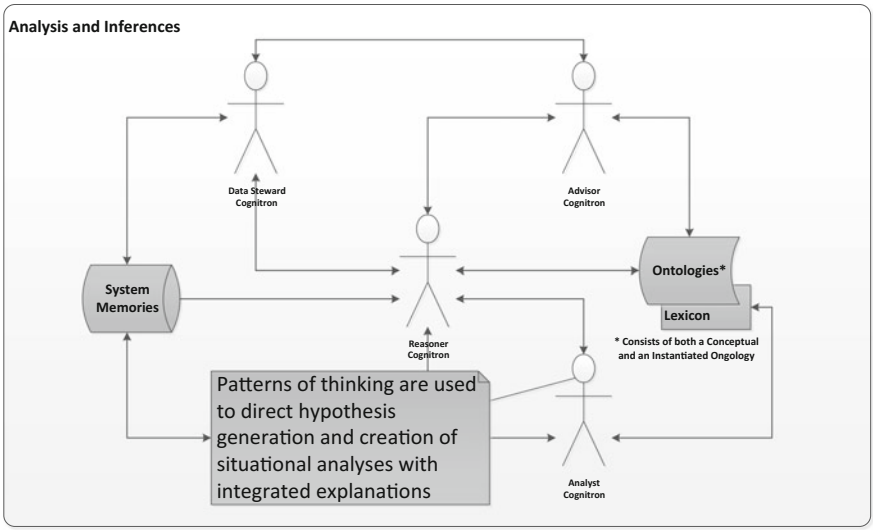


Fig. 11 The analyst agent

This process includes Search Information Agents that mine through multiple sources to provide data/information to other Intelligent Information Agents throughout the PHM processing environment. This is called the Federated Search

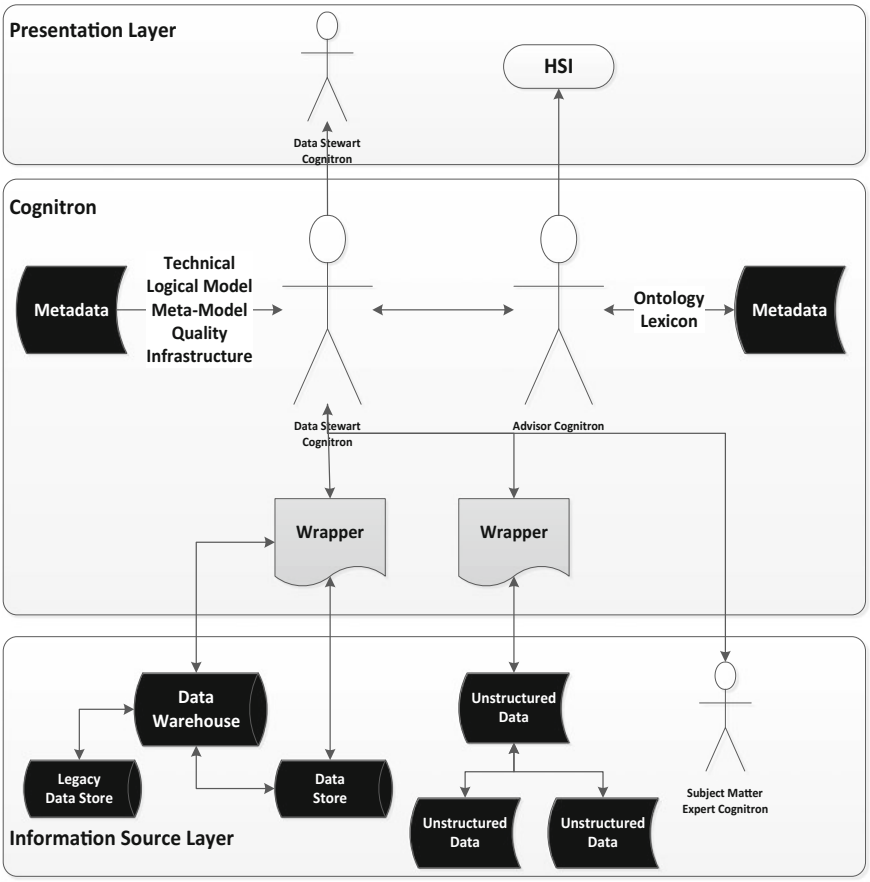


Fig. 12 Federated search within PHM

and is shown in Fig. 10 [6]. Notice that this process includes utilizing Subject Matter Experts (SMEs) to provide initial information to PHM. The system cannot just spontaneously generate initial knowledge, it must be fed information to learn from (not just train as in traditional neural network systems, but learn the information). This includes a learning based question and answer processing architecture that allows the ISHM/PHM processing environment to ask questions, based on contextual understanding of the information it is processing, and extract answers, either from its own inference engines, its own memories, other information contained in its storage systems, or outside information from other information sources, or SMEs. This process is illustrated in Fig. 12 [7].

This allows the modern PHM architecture to comprise a host of functional capabilities [8]:

1. Sensing and data acquisition,
2. Signal processing, conditioning and health assessment diagnostics and prognostics, and
3. Decision reasoning.

In addition, an intelligent Human System Interface (HSI) is required to provide the user with relevant, context-sensitive information about system condition. Utilizing the Intelligent Information Agent Architecture described here, an ISHM could provide a complete range of functionality from data collection through recommendations for specific actions. The key functions that an I^2A ISHM system could facilitate include:

1. Sensing and data acquisition (Data Steward Agents)
2. Signal Processing and feature extraction (Reasoner Agents)
3. Production of alarms or alerts (Advisor Agents)
4. Failure or fault diagnosis and health assessment (Analyst Agents)
5. Prognostics: projection of health profiles to future health or estimation of Remaining Useful Life (RUL) (Analyst and Advisor Agents)
6. Decision reasoning: recommendations or evaluation of asset readiness for a particular operational scenario (Advisor Agents)
7. Management and control of data flows and/or test sequences (Data Steward Agents)
8. Management of historical data storage and historical data access (Data Steward Agents)
9. System configuration management (Data Steward Agents)
10. Human System Interface (Interface Agents—Advisor Agents)

The use of Intelligent Information Agents allows both granular approaches (individual agents implementing individual functions) and integrated approaches (individual agents collaborating together to integrate a number of functions). The PHM architecture would take into account data flow requirements to control flexibility and performance across the PHM system. This allows the I^2A PHM system to support the full range of data flow requirements through both real-time and event-based data reporting and processing. Time-based reporting is further categorized as periodic or aperiodic. The event-based reporting and processing is based upon the occurrence of events (e.g., exceeding limits, state changes, etc.).

10 The Dialectic Argument Search

The Dialectic Argument Search (DAS) uses the Toulmin Argument Structure to find and relate information that develops a larger argument, or intelligence lead. The Question and Answer flow for the Dialectic Search is shown in Fig. 13. The DAS, illustrated in Fig. 14, serves two distinct purposes. First, it provides an effective basis for mimicking human reason. Second, it provides a means to glean relevant information from the Topic Map and transform it into actionable intelligence (practical knowledge.) These two purposes work together to provide an intelligent system that captures the capability of the ISHM operator to sort through diverse information and find clues [9].

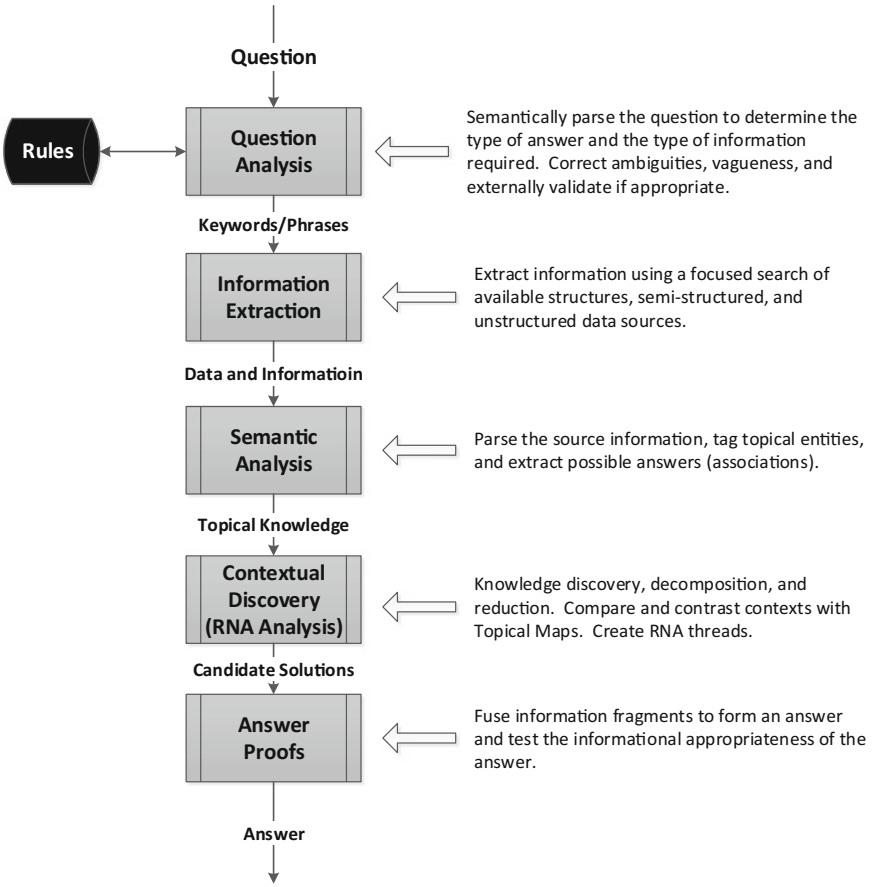


Fig. 13 Question and answer flow for DAS

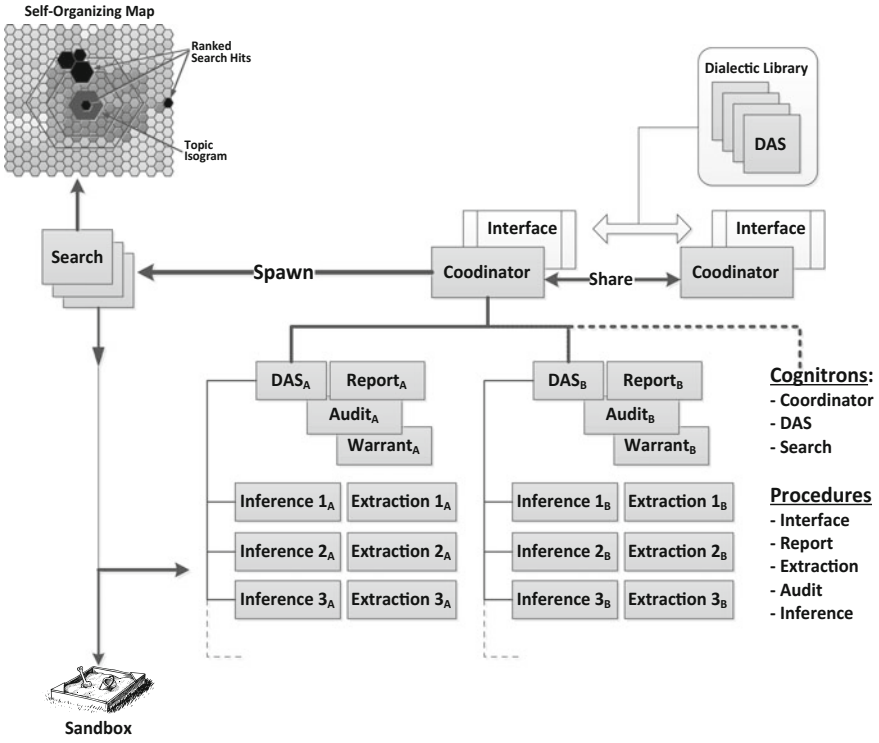


Fig. 14 The dialectic argument structure

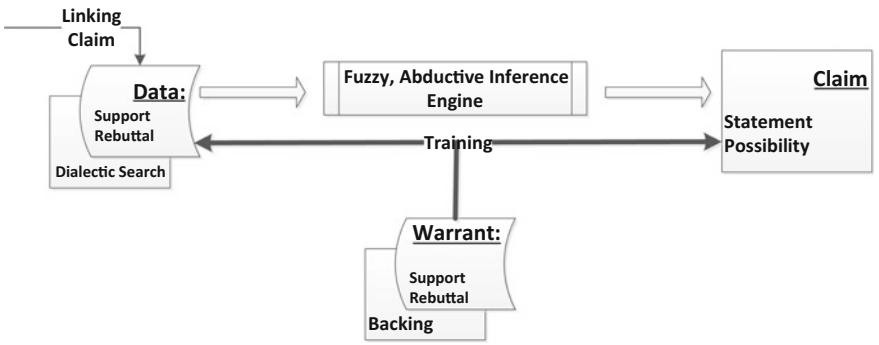


Fig. 15 The intelligent DAS software agency

Figure 15 illustrates a possible Intelligent Software Agent Architecture that could be used to implement the DAS: three different agents, the Coordinator, the DAS and the Search, all working together, each having its own learning objectives [10].

11 Conclusion and Discussion

Inter-agent communication allows shared awareness which in turn, enables faster operations and more effective information analysis and transfer providing users with an enhanced visualization of the overall constellation and situational awareness across a PHM infrastructure. The Intelligent Agent-based PHM can deal with massive amounts of information to levels of accuracy, timeliness, and quality never before possible. The knowledge relationship value store provides the underlying mechanism for high fidelity prognostic evaluation. The Data Steward Agents will support growing volumes of data and allow applications that deal with object-oriented technologies to achieve the goals of awareness, flexibility, and agility. The flexible, learning, and adapting Intelligent Software Agents of the PHM system can adapt, collaborate, and provide the increased flexibility required in a growing, changing environment [11].

References

1. J.A. Crowder, Intelligent agents for integrated system health management: modern diagnostic and prognostic techniques, in *Proceedings of the Air Force 3rd Annual Integrated Health Management Systems Conference, Cincinnati, OH*, 2010
2. J. Carbone, *A Framework for Enhancing Transdisciplinary Research Knowledge* (Texas Tech University, 2010)
3. J.A. Crowder, Multiple information agents for real-time ISHM: architectures for real-time warfighter support, in *Proceedings of the International Conference on Artificial Intelligence, Las Vegas*, 2010
4. J. Crowder, J. Carbone, Fuzzy methodologies for multi-sensor information fusion with applications to precision PNT, in *Proceedings of the 17th International Conference on Artificial Intelligence, Las Vegas, NV*, 2015
5. J.A. Crowder, Fuzzy possibilistic data model algebras for a new generation of databases, in *Proceedings of the Air Force 5th Annual Integrated Health Management Systems Conference, Cincinnati, OH*, 2008
6. J. Crowder, J. Carbone, Fuzzy Methodologies for multi-sensor information fusion with applications to precision PNT, in *Proceedings of the 17th International Conference on Artificial Intelligence, Las Vegas, NV*, 2015
7. J. Crowder, Cognitive architectures for prognostic health management, in *Proceedings of the 2013 IEEE International Conference on Prognostics and System Health Monitoring, Gaithersburg, MD*, 2013
8. J. Crowder, Artificial neural diagnostics and prognostics: Self-soothing in cognitive systems, in *Proceedings of the 2013 IEEE International Conference on Prognostics and System Health Monitoring, Gaithersburg, MD*, 2013
9. J. Crowder, The advanced learning, abductive network (ALAN), in *Proceedings of the AIAA Space 2013 Conference, San Diego, CA*, 2013
10. J. Crowder, S. Friess, J. Carbone, *Artificial Cognition Architectures* (Springer Publishing, New York, 2014)
11. J. Crowder, Probabilistic/Possibilistic Abductive Neural Networks (P^2ANNS) for decision support in autonomous systems, in *Proceedings of the AIAA Space 2013 Conference, San Diego, CA*, 2013

Part II
Modeling and Uncertainty Quantification

A Review of Crack Propagation Modeling Using Peridynamics

João Paulo Dias, Márcio Antonio Bazani, Amarildo Tabone Paschoalini and Luciano Barbanti

Abstract Improvements on prognostics and health management (PHM) techniques are extremely important in order to prevent system failure and reduce costs with maintenance and machine downtime. In the particular case of system components subjected to fracture failure, such improvements are closely related to the effect of crack propagation mechanisms on the quantification of the system remaining useful life (RUL). This chapter presents a review of the state-of-the-art of crack propagation modeling techniques and discusses the current limitations of finite elements methods (FEM) to model structures with cracks. The chapter also gives special attention to peridynamics (PD), a continuum non-local approach that has been considered to be a promising method to model structures with crack discontinuities. Therefore, the purpose of this chapter is to answer the following research question: "Can PD be a potential alternative to FEM on modeling of crack propagation problems in predicting RUL?" In order to answer this question, a literature review of the most relevant works on crack modeling field is presented and discussed. An application that involves a classical 2D crack propagation problem in a pre-notched glass plate is also included, in which comparisons between numerical predictions and experimental observations were performed. It was shown that PD produces more accurate predictions than FEM based-methods from both qualitative and quantitative perspectives.

Keywords Prognostics and health management · Crack propagation modeling · Peridynamics

J.P. Dias (✉)
Department of Mechanical Engineering, Texas Tech University, Lubbock, TX, USA
e-mail: joao-paulo.dias@ttu.edu

M.A. Bazani · A.T. Paschoalini
Department of Mechanical Engineering, São Paulo State University (UNESP), Ilha Solteira, SP, Brazil

L. Barbanti
Department of Mathematics, São Paulo State University (UNESP), Ilha Solteira, SP, Brazil

S. Ekwaro-Osire et al. (eds.), *Probabilistic Prognostics and Health Management of Energy Systems*, DOI 10.1007/978-3-319-55852-3_7

1 Introduction

1.1 Prognostics and Health Management

Complex energy systems (e.g. wind turbines) are often exposed to extreme and uncertain operation conditions that require accurate health monitoring techniques in order to preserve the system's reliability during its designed lifetime [1]. Additionally, the reduction on costs of operation and maintenance, which constitutes one of the main goals of industry, can be achieved through the incorporation of enhanced health monitoring techniques [2]. Prognostics and health management (PHM) is a well-established tool to predict the future condition of a system (or its components) taking into account the available information of the past usage and the current health state (diagnostics) of the system [3].

Prognostics-based frameworks are able to determine the future condition of the system by quantifying the remaining useful life (RUL), which measures the remaining amount of time or cycles of the system before failure based on a convenient failure criterion [1, 4]. The knowledge of this information is extremely useful for engineers to plan and schedule maintenance tasks before the failure of a system or a specific component occur. For this reason, the consolidation of a robust method to estimate RUL is fundamentally important for activities involving PHM [5, 6].

In order to predict RUL properly, previous knowledge on the degradation behavior of the system is also required. In general, such degradation behavior can be represented by mathematical models based on the physical description of the system damage (physics-based models) or, in other hand, based on test data measurements (data-driven models) [7]. Between these models, physics-based models are always preferred when test data are not available and there is a physical model of the damage whose unknown parameters can be estimated, such as models for battery degradation or structural damage caused by crack propagation [8].

Fracture of mechanical component due to crack propagation is the main cause of system failure when subjected to cyclic loads. Crack propagation in structures depends on some load operation conditions such as amplitude, stress ratio and frequency. Furthermore, the stochastic nature of such conditions makes crack parameters difficult to predict [9]. It is well established, for instance, that fatigue crack damage on gear tooth reduces its structural stiffness, which can affect the dynamic behavior (increase of vibration, noise and wear) of gearboxes operating under certain critical conditions, [10–12]. In this context, it is clear that modeling of fatigue crack growth and propagation plays a crucial role on RUL prediction of dynamics systems, [13, 14]. Due to the uncertainties involved in the crack propagation and branching phenomena, some authors [15, 16] have proposed a probabilistic PHM to estimate RUL as shown in Fig. 1. This figure illustrates the flowchart of the PHM procedure in which the damage model to predict the future state of the system damage is the focal point of the process. In other words, Fig. 1 shows that the selection of the damage model inputs (physical measurement,

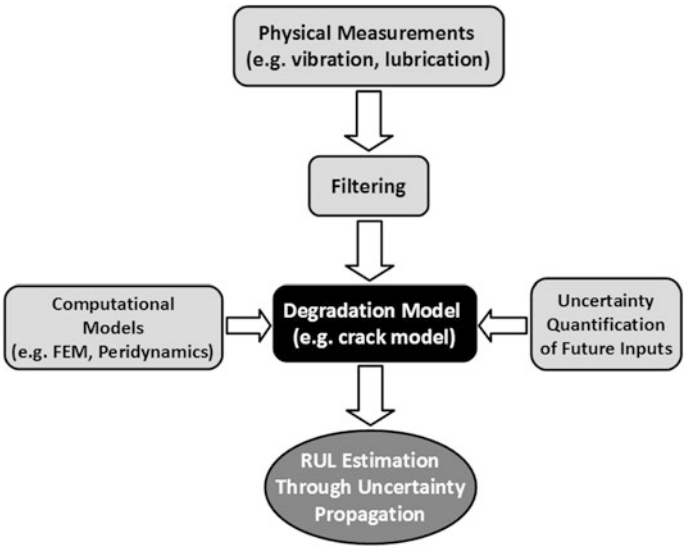


Fig. 1 Probabilistic PHM process for a crack propagation problem (adapted from [15] and [16])

filtering process, uncertainty quantification process and computational model) will have a direct impact on the estimation of probability distribution of RUL.

Focusing on crack damage models, the selection of the computational model plays a crucial role on the proper characterization of some crack features such as, propagation speed, direction and branching. Crack propagation is a complex phenomenon and different models has been proposed by several authors [17–21]. The majority of these models are finite element method (FEM) based-models which, despite the progress brought on the understanding of crack physics, still have serious limitations regarding mathematical formulation and computational effort [20, 22]. On the other hand, a new method called peridynamics (PD) recently developed by Silling [23] uses an integral formulation that has shown to be more appropriate to predict some crack parameters, such as, crack onset and crack branching according to experimental data [21, 24].

1.2 Motivation

As discussed in the previous section, the motivations of this chapter is based on:

- the need to improve PHM techniques in order to reduce costs with unnecessary maintenance procedures and machine downtime [1, 2];
- the need to quantify uncertainties on RUL estimation in order to predict RUL more accurately and prevent system/components failure events [15, 16];

- the need to understand the physics of crack propagation better and study its effects on quantification of RUL of system/components subjected to fracture failure [18, 19]; and
- the need to develop new numerical methods capable to overcome FEM mathematical and computational limitations to model discontinuities such as cracks [20, 22].

Based on the motivations outlined above, the following research question was formulated: "Can PD be a potential alternative to FEM on modeling of crack propagation problems in predicting RUL?" In order to answer this question, the chapter is organized as follows: Sect. 2 presents a review on the FEM applications to model cracks. Section 3 introduces PD as an alternative method to overcome some of the difficulties of FEM to deal with crack onset and branching. In Sect. 4, an application of 2D crack modeling in a pre-notched glass plate from literature is presented and discussed, showing the potential of PD to predict crack propagation in prognostics models and finally, the conclusions of the chapter are presented in Sect. 5.

2 Crack Propagation Modeling Using Finite Elements Methods

FEM is the most applied numerical method on both research and industrial fields to model problems related to structural damage [25]. The method applies two different forms to solve a problem: the strong form and the weak form. The strong form consists of the governing partial differential equations and the boundary conditions for a physical system whereas the weak form is an integral form of these equations [26].

A large variety of structural failure problems involving fractures due to fatigue loads can be modelled using the FEM approach. In such problems, cracks are common features and are mainly responsible for the structural failure. However, conventional FEM does not often account for the stress singularities around the crack tips, which consequently makes FEM not accurate enough to model crack problems [14, 27]. The main mathematical weakness of conventional FEM lies on the assumption that a body remains continuous as it deforms [28]. Consequently, this assumption leads to the difficulty of constructing mathematical formulations for the region around a singularity or a discontinuity such as a crack. It has been shown that the spatial derivatives for the partial differential equations cannot be constructed around a crack tip or surface [29].

In the rare situation when the crack surface coincides with the edge of the finite elements, FEM may handle crack tip asymptotic stresses [13]. However, such a situation is very unlikely as the crack propagates "randomly" through the material. A common characteristic of a crack is its unpredictable growth or propagation. The modeling of crack propagation using conventional FEM is computationally

intensive due to the need of the mesh to conform the crack contour [13, 18] at each time step as crack evolves.

Besides the needed mesh modifications to track the changes in geometrical and topological characteristics of the crack as it propagates, another drawback of conventional FEM is the necessity of the local refinements of the mesh around the crack surface [18]. This results in dense local meshes that often increases considerably the computational time processing [14].

To overcome such issues raised by the FEM formulation and its meshing process, some modified FEM's have been proposed to address the stress singularities problem [14]. One of the most widely used methods is an approach known as extended finite element method (XFEM). This method is based on the partition of unity property of the elements [30] allowing the crack not to be constrained to element boundaries, i.e. the crack can pass through the elements [31], which completely avoids the need of re-meshing the domain as the crack propagates [13]. Since XFEM permits the incorporation of local enrichment functions, modeling of moving cracks without changes in the mesh domain is possible due to the evolution of these enrichment functions with the crack interface geometry [18].

Despite that important progress has been made by using XFEM on crack propagation modeling, elements subdivision can bring out extra complexity and increase computational cost to the numerical integration of the method [20]. Additionally, due to the difficulties in setting up proper enrichment functions, the method still has problems in matching accurate predictions on crack propagation speed and crack branching angles according to observed experimental data [21, 24].

There have been recently also other methods proposed to address FEM deficiencies to deal with fatigue crack propagation and branching modeling. Some of these methods include cohesive zone modeling (CZM) [27], XFEM improvements based on analytical solutions to describe crack tip field enrichment functions [14], cell-based strain smoothing approach [18], some meshfree-based models [32], and an extention of XFEM for crack growth modeling to materials under creep condition [30]. A detailed discussion about these methods is out of the scope of this chapter and the aforementioned references are recommended to interested readers.

3 Peridynamics

3.1 Background

As already shown in the previous section, the mathematical formulation of the classical continuum theory constitutes the major difficulty in using techniques based on FEM to model crack behavior in solids [22]. Experimental observations have shown that crack growth and propagation occurs mostly due to microstructural irregularities of the material, in a scale size not captured by continuum models. For this reason, FEM's are limited to predict crack growth accurately [24]. Despite the

fact that XFEM has had some success in describing single discontinuities, it still requires auxiliary equations to predict nucleation and the propagation of a crack [33], which in turn makes the prediction of crack branching patterns more difficult [22].

Recently, peridynamics (PD) was proposed to address the shortcomings of FEM in dealing successfully with the nucleation and propagation of cracks in solids. PD is, in essence, a non-local reformulation of classical continuum mechanics [31, 34] which totally suppresses the hypothesis that a body remains continuous as it deforms [22]. As opposed to methods based on local approaches, PD is based on an integral formulation of the constitutive equations of motion which does not include spatial derivatives of the displacements [29, 35]. PD theory employs displacements rather than displacement derivatives in its formulation since spatial derivatives are not valid at the discontinuities generated by cracks [22]. The integral-based formulations are naturally able to deal with the presence of discontinuities in the material [27, 33]. This feature allows cracks to emerge and propagate spontaneously in multiple locations, along arbitrary paths without evoking additional mathematical relationships and/or crack growth criteria [27, 31].

PD has been shown to be a promising method to describe crack initiation, growth and propagation on fracture related applications [35]. It has already been applied successfully to model damage problems [34] considering its good accurate predictions on the shape of the crack paths, branching patterns and propagation speed [20, 21].

3.2 Problem Formulation

The key feature of PD theory is that material points are not allowed to interact only with their nearest neighbors but also with the points inside a given region, in PD defined by a horizon δ [20]. Such an interaction between a point $\boldsymbol{x}$ and a neighbor point $\boldsymbol{x}'$, both represented by vector Cartesian coordinates, is called bond [33] and is illustrated schematically in Fig. 2. First, we define a region H_x around the point $\boldsymbol{x}$ limited by a radius δ (horizon), within which any other material points $\boldsymbol{x}'$ can

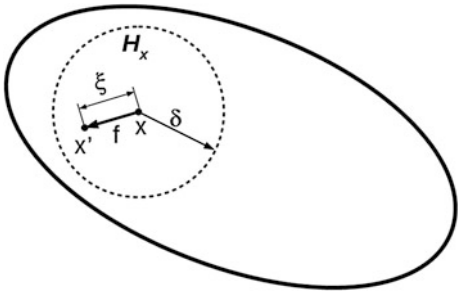


Fig. 2 Volume correction for the collocation points inside the horizon

interact with $\boldsymbol{x}$. The force that the material point $\boldsymbol{x}'$ exerts on the point $\boldsymbol{x}$ is denoted by the vector $\boldsymbol{f}$. Thus, the following relationship for the region $H_{\boldsymbol{x}}$ is valid [21]:

$$\xi = \left\|\boldsymbol{x}' - \boldsymbol{x}\right\| < \delta, \quad \forall \boldsymbol{x}' \in H_x, \tag{1}$$

in which $\xi = \left\|\boldsymbol{x}' - \boldsymbol{x}\right\|$ is the relative position between the two interacting point. This means that the state of any material point is determined by its pairwise interaction with the points that are located within the region $H_{\boldsymbol{x}}$. In other words, a pair of interacting material points can be formed only if the distance between them is less than the horizon radius δ.

Starting from the assumption that the force interactions between any pair of material points inside the domain obey Newton's Second Law of motion, the PD equation of motion can be formulated as [22]:

$$\rho \frac{d^2}{dt^2}\boldsymbol{u}(\boldsymbol{x}, t) = \int_{H_x} f(\boldsymbol{\eta}, \xi) dV_{x'} + \boldsymbol{b}(\boldsymbol{x}, t), \tag{2}$$

where ρ is the material density, $\boldsymbol{u}(\boldsymbol{x}, t)$ is the displacement vector field, t is the time, $dV_{x'}$ is the material point volume, $\boldsymbol{b}(\boldsymbol{x}, t)$ designates a prescribed body-force density field and $\boldsymbol{\eta} = \left\|\boldsymbol{u}\left(\boldsymbol{x}', t\right) - \boldsymbol{u}(\boldsymbol{x}, t)\right\|$ is the relative displacement between two interacting points. The vector $f(\boldsymbol{\eta}, \xi)$ is denoted as the pairwise force function, [20, 21] which represents the force per unit of volume exerted between the material points $\boldsymbol{x}$ and $\boldsymbol{x}'$. It is important to notice that Eq. (2) is an integral-differential equation, in which its only derivative appears at the inertia term (time derivative).

In order to model the pairwise force function $f(\boldsymbol{\eta}, \xi)$, a procedure based on linear micro-elasticity theory is followed. Basically, $f(\boldsymbol{\eta}, \xi)$ can be derived from the micro-elastic potential energy function, $\omega(\boldsymbol{\eta}, \xi)$, that connects a pair of material points as:

$$f(\boldsymbol{\eta}, \xi) = \frac{\partial}{\partial \boldsymbol{\eta}} \omega(\boldsymbol{\eta}, \xi). \tag{3}$$

If we consider that the stretch (or elongation) of the material varies linearly with the pairwise force between the two points, the micro-elastic potential, $\omega(\boldsymbol{\eta}, \xi)$ can be defined as [20, 21, 33]:

$$\omega(\boldsymbol{\eta}, \xi) = \frac{c(\xi) s^2 \xi}{2}, \tag{4}$$

where $c(\xi)$ is the micro-modulus that represents the elastic stiffness of the bond and s is the stretch of a bond given by,

$$s = \frac{\|\boldsymbol{\eta} + \boldsymbol{\xi}\| - \boldsymbol{\xi}}{\boldsymbol{\xi}}. \tag{5}$$

The PD elastic strain energy density W at a point $\boldsymbol{x}$ is obtained by integrating the micro-elastic potential (Eq. 4) over the horizon region [21],

$$W(\boldsymbol{x}) = \frac{1}{2}\int_{H_x} \omega(\boldsymbol{\eta}, \boldsymbol{\xi}) d\boldsymbol{x}' = \frac{1}{2}\int_{-\delta}^{\delta} \left[\frac{c(\xi)s^2 r}{2}\right] 2\pi r dr = \frac{\pi}{6} c(\xi) s^2 \delta^3. \tag{6}$$

Assuming a constant value for the micro-modulus $c(\xi) = c_0$ [36], it can be obtained by equating Eq. (6) to the classical strain energy density which results in [20, 21]:

$$c_0 = \frac{6E}{\pi\delta^3(1-\nu)}, \tag{7}$$

in which E and ν are the Young's modulus and the Poisson's ration of the material, respectively.

Structural failure in PD occurs when all bounds within a horizon break down, i.e. if they are stretched beyond a critical value, s_0 [20]. After this event, the contact force between the points is ceased and they no longer interact with each other. This particular event points out the initiation of a crack whose direction and velocity of propagation will be determined as other bounds in the domain are broken. As suggested by Silling and Askari [36], the critical stretch for bound failure, s_0, is a function of the critical fracture energy release, G_0, the material Young's modulus, E, and the horizon radius δ. For 2D cases, this expression is given by [20, 33]:

$$s_0 = \sqrt{\frac{4\pi G_0}{9E\delta}}. \tag{8}$$

Therefore, using Eqs. (4) and (5) on Eq. (3) and using the failure criteria expressed in Eq. (8), the pairwise force function is defined as:

$$f(\boldsymbol{\eta}, \boldsymbol{\xi}) = \begin{cases} c(\xi) s \frac{\boldsymbol{\xi} + \boldsymbol{\eta}}{\|\boldsymbol{\xi} + \boldsymbol{\eta}\|}, & if\ s < s_0 \\ 0, & if\ s \geq s_0 \end{cases}. \tag{9}$$

3.3 Solution

In order to obtain the numerical solution of Eq. (2), consider a 2D domain discretization as illustrated in Fig. 3. Each discrete material point of the domain is equally spaced each other by a distance Δ. A horizon radius $\delta = 3\Delta$ is defined

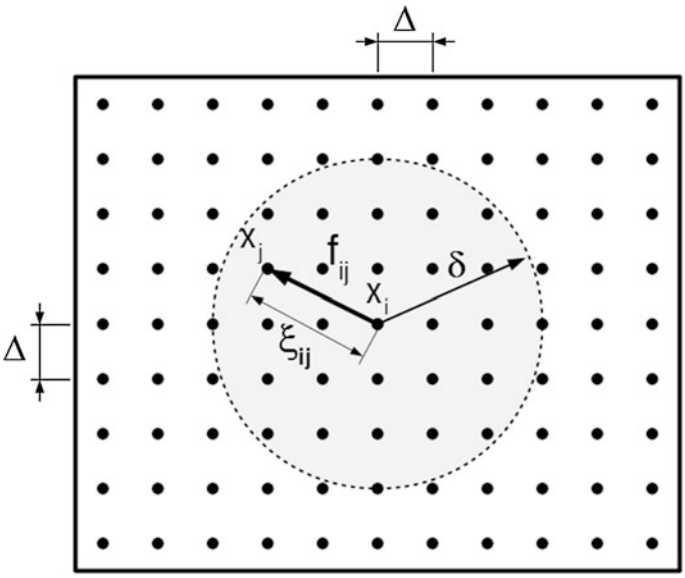


Fig. 3 PD discretization scheme of the domain

around a material point $\boldsymbol{x_j}$. All $\boldsymbol{x_j}$ points whose relative positions ξ_{ij} to $\boldsymbol{x_i}$ are smaller than the horizon radius will interact with the point $\boldsymbol{x_i}$. Therefore, the discretized PD equation of motion can be written as [22]:

$$\ddot{\rho}\boldsymbol{u}_i^n = \sum_j f\left(\boldsymbol{\eta}_{ij}, \xi_{ij}\right) V_j + \boldsymbol{b}_i^n, \tag{10}$$

in which the subscript n denotes the time step, V_j is the volume of the sub-domain that is represented by the collocation point located at $\boldsymbol{x}_j$ and the relative position and relative displacement are respectively written as:

$$\boldsymbol{\xi}_{ij} = \boldsymbol{x}_j^n - \boldsymbol{x}_i^n, \text{ and} \tag{11}$$

$$\boldsymbol{\eta}_{ij} = \boldsymbol{u}_j^n - \boldsymbol{u}_i^n. \tag{12}$$

For the discretization of the second derivative of the displacement with respect to the time (the inertial term at the left side of Eq. 10), a central differences scheme can be used for a discrete time interval Δt [22]. Thus,

$$\ddot{\boldsymbol{u}}_i^n = \frac{\boldsymbol{u}_i^{n+1} - 2\boldsymbol{u}_i^n + \boldsymbol{u}_i^{n-1}}{\Delta t^2}. \tag{13}$$

4 Crack Propagation Modeling Using Peridynamics

This section presents and discusses an application of crack propagation modeling using PD available in the open literature [20, 22]. The main purpose is to illustrate PD's capability of describing crack propagation problems satisfactorily. The application involves a classic 2D problem of crack growth in a pre-notched glass plate studied by two different research groups: Ha and Bobaru [20] and Agwai et al. [22]. Their PD predictions were compared against FEM-based model predictions and experimental observations made by other authors [37–40].

4.1 Problem Description

The problem presented by Ha and Bobaru [20] and Agwai et al. [22] considered a rectangular glass plate (40 mm high vs. 100 mm length) as depicted in Fig. 4. A notch with 50 mm length is placed at the center of left edge of the plate in order to increase the stress distribution at this region and facilitate the crack propagation, in which a uniform dynamic load is applied perpendicularly to the horizontal edges of the plate. Specific material properties, load conditions and model parameters used in PD simulations can be found in [20] and [22]. Predictions of some crack growth parameters obtained by these authors, such as crack patterns and propagation speeds are discussed in the next sub-sections.

4.2 Crack Pattern and Branching Predictions

PD qualitative predictions of the crack propagation pattern and branching features were performed by Agwai et al. [22] and their results were compared against

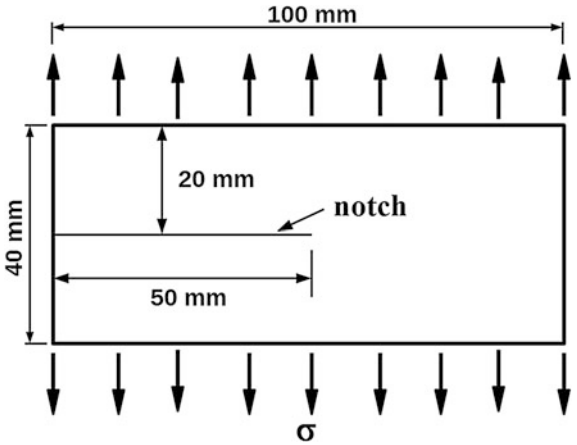


Fig. 4 Illustration of a pre-notched glass plate submitted to a dynamic stretching load

experimental data obtained by Ramulu and Kobayashi [37]. Moreover, Agwai et al. [22] also compared their predictions with two FEM-based approaches from other authors, which modeled the same problem: cohesive zone model (CZM) [38] and a modified version of XFEM to enhance crack modeling, named crack node method (CNM) [39]. The experimental study [37] showed that the crack propagates linearly, with very small branches forming up and stopping suddenly, from the tip of the notch through a considerably length inside the glass plate. From this point on, the crack splits into two branches that keep growing independently each other symmetrically separated by an angle of about 40º with no small branches welling up [22].

Agwai et al. [22] results showed that PD was able to predict the experimental behavior observed by Ramulu and Kobayashi [37] exceptionally well. The main points of agreement between PD prediction and experimental observations were: (i) the proper representation of the crack breakage into two main branches, (ii) the capture of the formation of small branches, which start growing and terminates subsequently, before the crack split up into two, and (iii) the non-existence of small branches along the path of the two main branches.

Crack propagation predictions using CZM method were performed considering different levels of mesh refinement for both structured and non-structured meshes [38]. In general terms, their simulations were also able to capture the phenomenon of crack breakage into two main branches as observed in Ramulu and Kobayashi [37] experiments. Nevertheless, their results were shown to be highly dependent on the mesh parameters (element size, shape and orientation), once this method allows the crack only to contour the elements boundaries [22], having a direct impact on the results for the direction of crack propagation. CZM method was not able to predict properly the small branches during the propagation path from the notch tip to the main branching initiation observed experimentally in [37]. Furthermore, Agwai et al. [22] observed that only the unstructured grid produced an irregular phenomenon slightly similar to the small branches. Moreover, the symmetric propagation pattern observed in the experiment after the crack splits up into the two main branches was affected by the mesh parameters: only refined meshes could capture this symmetry; coarser grids predicted highly unsymmetrical crack branching patterns.

While CZM method could not predict properly the incidence of small branches during the crack propagation, the results obtained through XFEM-CNM [39] showed to follow the opposite direction. Similar to PD and CZM methods, XFEM-CNM predicted the occurrence of the crack breakage into two main branches according to the experimental observations from Ramulu and Kobayashi [37]. However, XFEM-CNM captured a disproportionally large number of small branches welling up even after the crack splits up into two. This phenomenon was not observed experimentally, in which no small branches were reported along the two main crack branches. Agwai et al. [22] suggests that this phenomenon seems to be caused by a model mechanism that cannot discriminate properly when a small branch should be introduced or not. Other important feature observed on XFEM-CNM predictions was that, similarly to CZM model, the symmetry between

the two crack main branches was difficult to capture properly. Agwai et al. [22] also reasonably explains that the mesh-dependency nature of FEMs requires sophisticated mesh refinement schemes to overcome this problem, which could result on considerable increase on computational power demand.

4.3 Crack Propagation Speed Predictions

Additionally to the qualitative crack path and branching comparisons, Agwai et al. [22] also quantified crack propagation speeds and compared their results with the experimental observations from Ramulu and Kobayashi [37] and the results of FEM approaches. Essentially, they reported that regarding the time elapsed to initiate the crack breakage into the two main branches, both PD and CZM predictions fitted better to experiments than XFEM-CNM. According to the authors, both methods allowed the crack initially to propagate for a long time before it splits up. Nevertheless, the crack propagation velocities calculated using PD and XFEM-CNM quantitatively showed very similar behavior after the branching.

Ha and Bobaru [20] also used a PD model to calculate crack propagation speed for the same problem and compared their predictions against the experiments from Bowden et al. [40] for a soda-lime glass plate. The differences between Bowden et al. [40] and Ramulu and Kobayashi [37] experiments are on the type of glass used (which slightly differs each other on the mechanical properties) and on the loading conditions. Ha and Bobaru [20] calculated the crack propagation speed by computing the difference on the position of the crack tip between two consecutive time steps. Their results showed that the maximum crack propagation speed using PD model is in good agreement with the experimental measurements from Bowden et al. [40] (over-predicting experimental results about only 6%). However, the authors highlighted that, even considering that the experiment were performed under different type of loading (quasi-static in the experiment and dynamic in their simulation), the agreement between both can be considered good.

Lastly, a summary table of the comparisons between the crack propagation models and the experimental works discussed along this section is outlined on Table 1. Each line of the table shows the main qualitative (second to forth columns) and quantitative (last column) results for each model. The results of Table 1 show clearly that PD predictions fitted better to the experiment observations than the FEM models from both qualitative and quantitative perspectives. Thus, considering the discussions presented in this chapter and summarized in Table 1, the answer of the research question outlined in the Sect. 1 is “yes”: PD can be considered a potential alternative to finite element on modeling of crack propagation problems.

Table 1 Summary table of the comparison between crack models and experiments analyzed on this chapter

Crack model	Occurrence of two main branches	Occurrence of small branches	Symmetry between the main branches	Crack propagation speed
Peridynamics[a] [20]	Not presented	Not presented	Not presented	Good agreement with experimental maximum crack propagation speed (only 6% difference) even considering different loading conditions
Peridynamics[b] [22]	Captured	Well captured	Well captured	Good agreement with experimental branching initiation time; good agreement with XFEM-CNM [39] propagation speed
CZM structured coarse mesh[b] [38]	Captured	Under-estimated	Not captured	Good agreement with experimental branching initiation time; under-estimated propagation speed
CZM structured refined mesh[b] [38]	Captured	Under-estimated	Well captured	
CZM unstructured refined mesh[b] [38]	Captured	Under-estimated	Well captured	
XFEM-CNM[b] [39]	Captured	Over-estimated	Not captured	Under-estimated experimental branching initiation time; good agreement with PD [22] propagation speed

[a]Comparisons performed with experiments from Bowden et al. [40]
[b]Comparisons performed with experiments from Ramulu and Kobayashi [37]

5 Conclusions

This chapter presented a review of some of the most relevant works dedicated to model fracture problems related to crack growth and propagation, which is a fundamental concern of prognostics and health management (PHM) practices. Currently, such problems are mostly solved using continuum local-based models such as finite element methods (FEM). However, some limitations of these methods, particularly the ones concerning the mathematical difficulties on dealing with crack discontinuities were pointed out. In an attempt to mitigate FEM shortcomings, special attention was given to peridynamics (PD), a continuum non-local approach, that, due to its integral formulation, has been considered a promising method to

model crack discontinuities. This is due to PD's ability to work with discontinuous domains without the need of any supplemental relationship that dictates the direction of crack propagation.

An application of a classical crack modeling in a 2D pre-notched glass plate using PD, which was selected from the literature, was presented and discussed. The results of the reviewed PD works were compared with FEM predictions and with experimental observations, both obtained from other authors. It was clearly verified that PD predicted more accurately the experimental observations than FEM models from both qualitative perspective (crack pattern description and branching features) and quantitative perspective (crack propagation speed). For these reasons, one can conclude that the answer of the research question outlined in the Sect. 1 of the chapter is "yes": PD can be considered as a potential alternative to FEM on modeling of crack propagation problems in PHM. Despite the need of more advances and refinements on PD technique, mainly to deal with more complexes geometries, it has been proved that PD is an extremely useful tool on crack modeling, which can be soon incorporated to the PHM routines to determine the remaining useful life (RUL) of systems/components subjected to fatigue loading.

Acknowledgements Dr. João Paulo Dias (corresponding author) would like to thank Professor Stephen Ekwaro-Osire (corresponding editor of this book) for the fruitful discussions during the writing process of this chapter.

References

1. G. Bartram, S. Mahadevan, Probabilistic prognosis with dynamic bayesian networks. Int. J. Progn. Health Manage. **2**, 2153–2648 (2015)
2. A.K. Garga, K.T. McClintic, R.L. Campbell, C.-C. Yang, M.S. Lebold, T.A. Hay, C.S. Byington, Hybrid reasoning for prognostic learning in CBM systems, in *2001 IEEE Aerospace Conference Proceedings*, vol. 6 (2001), pp. 2957–2969
3. B. Saha, K. Goebel, S. Poll, J. Christophersen, Prognostics methods for battery health monitoring using a bayesian framework. IEEE Trans. Instrum. Meas. **58**(2), 291–296 (2009)
4. D. An, N.H. Kim, J.H. Choi, Practical options for selecting data-driven or physics-based prognostics algorithms with reviews. Reliab. Eng. Syst. Saf. **133**, 223–236 (2015)
5. X.S. Si, W. Wang, C.H. Hu, D.H. Zhou, Remaining useful life estimation—a review on the statistical data driven approaches. Eur. J. Oper. Res. **213**(1), 1–14 (2011)
6. S. Sankararaman, K. Goebel, Why is the remaining useful life prediction uncertain?, in *Annual Conference of the Prognostics and Health Management Society* (2013), pp. 1–13
7. S. Sankararaman, M.J. Daigle, K. Goebel, Uncertainty quantification in remaining useful life prediction using first-order reliability methods. IEEE Trans. Reliab. **63**(2), 1–17 (2014)
8. D. An, J.-H. Choi, N.H. Kim, Prognostics 101: A tutorial for particle filter-based prognostics algorithm using Matlab. Reliab. Eng. Syst. Saf. **115**, 161–169 (2013)
9. H. Xiaoping, T. Moan, C. Weicheng, An engineering model of fatigue crack growth under variable amplitude loading. Int. J. Fatigue **30**(1), 2–10 (2008)
10. Z. Chen, Y. Shao, Dynamic simulation of spur gear with tooth root crack propagating along tooth width and crack depth. Eng. Fail. Anal. **18**(8), 2149–2164 (2011)
11. F. Chaari, T. Fakhfakh, M. Haddar, Analytical modelling of spur gear tooth crack and influence on gearmesh stiffness. Eur. J. Mech.-A/Solids **28**(3), 461–468 (2009)

12. Y. Pandya, A. Parey, Simulation of crack propagation in spur gear tooth for different gear parameter and its influence on mesh stiffness. Eng. Fail. Anal. **30**, 124–137 (2013)
13. I.V. Singh, B.K. Mishra, S. Bhattacharya, R.U. Patil, The numerical simulation of fatigue crack growth using extended finite element method. Int. J. Fatigue **36**(1), 109–119 (2012)
14. X.F. Hu, W.A. Yao, A new enriched finite element for fatigue crack growth. Int. J. Fatigue **48**, 247–256 (2013)
15. F.M. Alemayehu, S. Ekwaro-Osire, Probabilistic Model-Based Prognostics using Meshfree Modeling, in *Probabilistic Prognostics and Health Management of Energy Systems*, ed. By S. Ekwaro-Osire, A.C. Gonçalves, F.M. Alemayehu (Springer, New York, Chapter 1, 2017). ISBN: 978-3-319-55851-6
16. S. Ekwaro-Osire, H. B. Endeshaw, F.M. Alemayehu, O. Geçgel, Probabilistic Model-Based Prognostics using Meshfree Modeling, in *Probabilistic Prognostics and Health Management of Energy Systems*, ed. By S. Ekwaro-Osire, A.C. Gonçalves, F.M. Alemayehu (Springer, New York, Chapter 5, 2017). ISBN: 978-3-319-55851-6
17. N. Sukumar, D.L. Chopp, E. Béchet, N. Moes, Three-dimensional non-planar crack growth by a coupled extended finite element and fast marching method. Int. J. Numer. Meth. Eng. **76** (5), 727–748 (2008)
18. L. Chen, T. Rabczuk, S.P.A. Bordas, G.R. Liu, K.Y. Zeng, P. Kerfriden, Extended finite element method with edge-based strain smoothing (ESm-XFEM) for linear elastic crack growth. Comput. Methods Appl. Mech. Eng. **209**, 250–265 (2012)
19. M.-H. Gozin, M. Aghaie-Khafri, Quarter elliptical crack growth using three dimensional finite element method and crack closure technique. J. Mech. Sci. Technol. **28**(6), 2141–2151 (2014)
20. Y.D. Ha, F. Bobaru, Studies of dynamic crack propagation and crack branching with peridynamics. Int. J. Fract. **162**(1-2), 229–244 (2010)
21. Y.D. Ha, F. Bobaru, Characteristics of dynamic brittle fracture captured with peridynamics. Eng. Fract. Mech. **78**(6), 1156–1168 (2011)
22. A. Agwai, I. Guven, E. Madenci, Predicting crack propagation with peridynamics: a comparative study. Int. J. Fract. **171**(1), 65–78 (2011)
23. S.A. Silling, Reformulation of elasticity theory for discontinuities and long-range forces. J. Mech. Phys. Solids **48**(1), 175–209 (2000)
24. R. Beckmann, R. Mella, M.R. Wenman, Mesh and timestep sensitivity of fracture from thermal strains using peridynamics implemented in Abaqus. Comput. Methods Appl. Mech. Eng. **263**, 71–80 (2013)
25. W. He, J. Liu, D. Xie, Probabilistic life assessment on fatigue crack growth in mixed-mode by coupling of Kriging model and finite element analysis. Eng. Fract. Mech. **139**, 56–77 (2015)
26. J. Fish and T. Belytschko, *A First Course in Finite Elements*, Wiley, 2007
27. M. Taylor, D.J. Steigmann, A two-dimensional peridynamic model for thin plates. Math. Mech. Solids **20**(8), 998–1010 (2013)
28. E. Madenci, E. Oterkus, *Peridynamic Theory and Its Applications* (Springer, New York, 2014)
29. R.W. Macek, S.A. Silling, Peridynamics via finite element analysis. Finite Elem. Anal. Des. **43**(15), 1169–1178 (2007)
30. Q. Meng, Z. Wang, Extended finite element method for power-law creep crack growth. Eng. Fract. Mech. **127**, 148–160 (2014)
31. W. Liu, J.W. Hong, A coupling approach of discretized peridynamics with finite element method. Comput. Methods Appl. Mech. Eng. **245**, 163–175 (2012)
32. T. Rabczuk, T. Belytschko, Cracking particles: a simplified meshfree method for arbitrary evolving cracks. Int. J. Numer. Meth. Eng. **61**, 2316–2343 (2004)
33. D. Dipasquale, M. Zaccariotto, U. Galvanetto, Crack propagation with adaptive grid refinement in 2D peridynamics. Int. J. Fract. **190**(1-2), 1–22 (2014)
34. W. Hu, Y.D. Ha, F. Bobaru, S.A. Silling, The formulation and computation of the nonlocal J-integral in bond-based peridynamics. Int. J. Fract. **176**(2), 195–206 (2012)

35. X. Chen and M. Gunzburger, Continuous and discontinuous finite element methods for a peridynamics model of mechanics. Comput. Methods Appl. Mech. Eng. **200**(9), 1237-1250 (2011)
36. S.A. Silling, E. Askari, A meshfree method based on the peridynamic model of solid mechanics. Comput. Struct. **83**(17), 1526–1535 (2005)
37. M. Ramulu, A.S. Kobayashi, Mechanics of crack curving and branching—a dynamic fracture analysis. Int. J. Fract. **273**(4), 187–200 (1985)
38. J.-H. Song, H. Wang, T. Belytschko, A comparative study on finite element methods for dynamic fracture. Comput. Mech. **42**(2), 239–250 (2008)
39. J.-H. Song, T. Belytschko, Cracking node method for dynamic fracture with finite elements. Int. J. Numer. Meth. Eng. **77**(3), 360–385 (2009)
40. F.P. Bowden, J.H. Brunton, J.E. Field, A.D. Heyes, Controlled fracture of brittle solids and interruption of electrical current. Nature **216**, 38–42 (1967)

Modeling and Quantification of Physical Systems Uncertainties in a Probabilistic Framework

Americo Cunha Jr.

Abstract Uncertainty quantification (UQ) is a multidisciplinary area, that deals with quantitative characterization and reduction of uncertainties in applications. It is essential to certify the quality of numerical and experimental analyses of physical systems. The present manuscript aims to provide the reader with an introductory view about modeling and quantification of uncertainties in physical systems. In this sense, the text presents some fundamental concepts in UQ, a brief review of probability basics notions, discusses, through a simplistic example, the fundamental aspects of probabilistic modeling of uncertainties in a physical system, and explains what is the uncertainty propagation problem.

Keywords Uncertainty quantification · Stochastic modeling of uncertainties · Probabilistic approach

1 An Introductory Overview on UQ

Typically, highly complex engineering projects use both numerical simulations and experimental tests on prototypes to specify a certain system or component with desired characteristics. These two tools are used in a similar way by scientists to investigate physical phenomena of interest. However, none of these approaches provides a response that is an exact reproduction of the physical system behaviour, because computational model and test rig are subject to uncertainties, which are intrinsic to modeling process (lack of knowledge on the physics) and model parameters (measurement inaccuracies, manufacturing variabilities, etc.).

In order to improve the reliability level of numerical results and experimental data, it is necessary to quantify the underlying uncertainties. The cautious experimentalists have been doing this for many decades, leading to a high level competence in what concerns the specification of the level of uncertainty in an experiment. It is worth

A. Cunha Jr. (✉)
NUMERICO - Nucleus of Modeling and Experimentation with Computers,
Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brazil
e-mail: americo@ime.uerj.br

S. Ekwaro-Osire et al. (eds.), *Probabilistic Prognostics and Health Management of Energy Systems*, DOI 10.1007/978-3-319-55852-3_8

remembering that an experiment that does not specify the level of uncertainty is not well seen by the technical/scientific community. On the other hand, just recently the numerical community has begun to pay attention to the need of specifing the level of confidence for computer simulations.

Uncertainty quantification (UQ) is a multidisciplinary area that deals with quantitative characterization and the reduction of uncertainties in applications. One reason that UQ has gained such popularity over the last years, in numerical world, is due to several books on the subject have recently emerged [1–12]. To motivate its study, we present three important scenarios where UQ is an essential tool:

Decision making: Some risk decisions, which negative result can cause catastrophic failure or huge financial costs, need to be well analysed before a final opinion by the responsible party. The possible variabilities that generate uncertain scenarios need to be taken into account in the analysis. The evaluation of these uncertain scenarios has the task of assisting the responsible party to minimize the chances of a wrong decision. Briefly, and in this context, UQ is essential to provide the necessary certification for a risk decision.

Model validation: Experimental data are widely used to check the accuracy of a computational model which is used to emulate a real system. Although this procedure is already being used by scientists and engineers for many decades, there is still no universally accepted criteria to ensure the model quality. However, it is known that any robust criteria of model validation must take into account the simulation and experiment uncertainties.

Robust design: An increasingly frequent requirement in several projects is the robust design of a component which consists make a specific device low sensitive to variation on its properties. This requires the quantification of model and parameters uncertainties.

In a very simplistic way, we can summarize UQ objectives as (i) *add error bars to experiments and simulations*, and (ii) *define a precise notion of the validated model*.

The first objective is illustrated in Fig. 1a, which shows the comparison between a simulation result with experimental data, and in Fig. 1b, that presents the previous graph with the inclusion of an envelope of reliability around the simulation. As careful experimentalists, which use error bars for a long time, UQ mainly focuses on "error bars for simulations".

Moreover, a possible notion of validated model is illustrated in Fig. 2, where experiment and simulation are compared, and the computational model is considered acceptable if the admissible range for the experimental value (defined by the point and its error bar) is contained within the reliability envelope around the simulation.

This chapter is organised into six sections. Besides this introduction, there is a presentation of some fundamental concepts of UQ in Sect. 2; a brief review on probability theory basics in Sect. 3; an exposure of the fundamental aspects of probabilistic modeling of uncertainties, through a simplistic example, in Sect. 4; the presentation of the uncertainty propagation problem in Sect. 5; and the final remarks in Sect. 6.

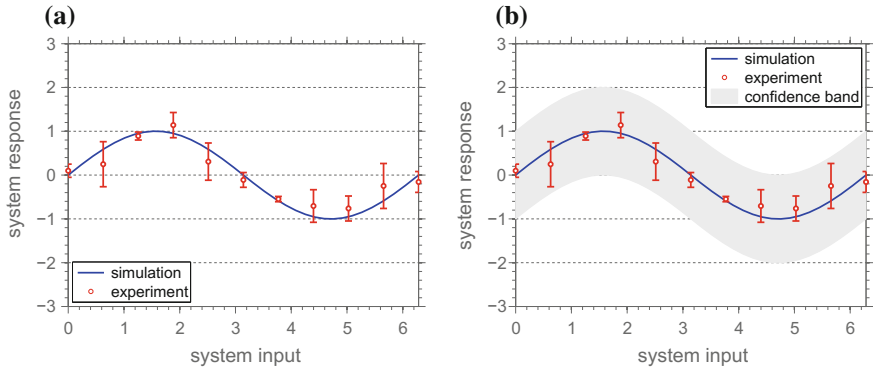


Fig. 1 **a** Comparison between simulation and experimental data, without an envelope of reliability for the simulation, and **b** including this envelope

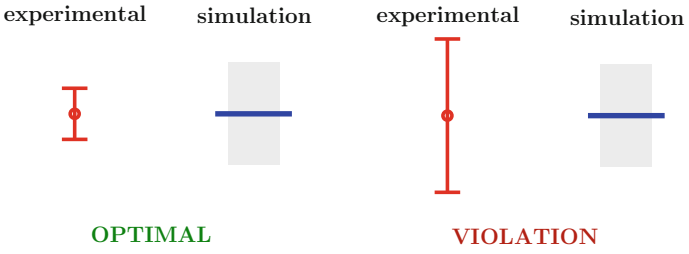


Fig. 2 Illustration of a possible notion of validated model

It is noteworthy that many of the ideas that are presented in this manuscript are very influenced by courses taught by the author's doctoral supervisor, Prof. Christian Soize [13–15]. Lectures of Prof. Gianluca Iaccarino, Prof. Alireza Doostan, and collaborators were also very inspiring [16–18].

2 Some Fundamental Concepts on UQ

This section introduce some fundamental notions in the context of UQ.

2.1 Errors and Uncertainties

Unfortunately, until the present date, there is still no consensus in UQ literature about the notions of errors and uncertainties. This manuscript presents the definitions we think make more sense, introduced by [19].

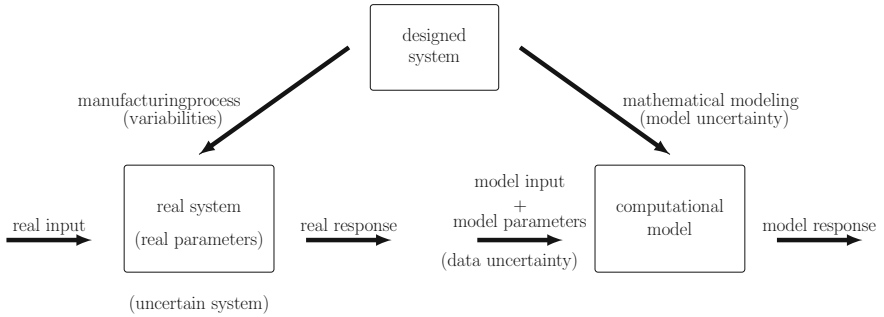


Fig. 3 Schematic representation of the relationship between the designed system, the real system and the computational model [19]

Let's start with three conceptual ideas that will be relevant to the stochastic modeling of physical systems: *designed system*, *real system* and *computational model*. A schematic illustration of these concepts is shown in Fig. 3.

Designed system: The designed system consists of an idealized project for a physical system. It is defined by the shape and geometric dimensions, material properties, connection types between components (boundary conditions), and many other parameters. This ideal system can be as simple as a beam or as complex as an aircraft [19].

Real system: The real system is constructed through a manufacturing process taking the designed system as reference. In contrast to the designed system, the real system is never known exactly, as the manufacturing process introduces some variabilities in the system geometric dimensions, on its materials properties, etc. No matter how controlled the construction process is, these deviations from the conceptual project are impossible to eliminate, since any manufacturing process is subjected to finite accuracy. Thus, the real system is uncertain with respect to the designed system [19].

Computational model: In order to analyze the real system behaviour, a computational model should be used as predictive tool. The construction of this computational model initially performs a physical analysis of the designed system, identifies the associated physical phenomena and makes hypotheses and simplifications about its behaviour. The identified physical phenomena are then translated into equations in a mathematical formulation stage. Using the appropriate numerical methods, the model equations are then discretized and the resulting discrete system of equations is solved, providing an approximation to the computational model response. This approximate response is then used to predict the real system behaviour [19].

Numerical errors: The response obtained with the computational model is, in fact, an approximation to the model equation's true solution. Inaccuracies, intrinsic to the discretization process, are introduced in this step giving rise to *numerical errors* [19]. Other source of errors are: (i) the finite precision arithmetic that is used to perform the calculations, and (ii) possible bugs in the computer code implementation of the computational model.

Uncertainties on the data: The computational model is supplied with model input and parameters, which are (not exact) emulations of the real system input and parameters, respectively. Thus, it is uncertain with respect to the real system. The discrepancy between the real system and computational model supplied information is called *data uncertainties* [4, 19].

Uncertainties on the model: In the conception of the computational model, considerations made may or may not be in agreement with reality, which should introduce additional inaccuracies known as model uncertainties. This source of uncertainty is essentially due to lack of knowledge about the phenomenon of interest and, usually, is the largest source of inaccuracy in computational model response [4, 19].

Naturally, uncertainties affect the response of a computational model, but they should not be considered errors because they are physical in nature. Errors are purely mathematical in nature and can be controlled and reduced to a negligible level if the numerical methods and algorithms used are well known by the analyst [4, 19]. This differentiation is summarized in Fig. 4.

2.2 Verification and Validation

Today *verification* and *validation*, also called V&V, are two concepts of fundamental importance for any carefully done work in UQ. Early works advocating in favor of these ideas, and showing their importance, date back to the late 1990s and early 2000s [20–23]. The impact on the numerical simulation community was not immediate, but has been continuously growing over the years, conquering a prominent space in the last ten years, especially after the publication of Oberkampf and Roy's book [24].

These notions are well characterized in terms of two questions:

Verification:

Are we solving the equation right?

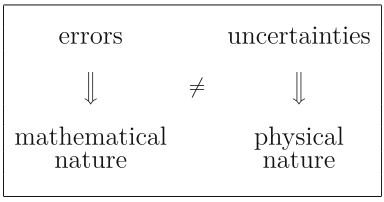


Fig. 4 The difference between errors and uncertainties

Validation:

Are we solving the right equation?

Although extremely simplistic, the above "definitions" communicate, directly and objectively, the key ideas behind the two concepts. Verification is a task whose goal is to make sure that the model equation's solution is being calculated correctly. In other words, it is to check if the computational implementation has no critical bug and the numerical method works well. It is an exercise in mathematics. Meanwhile, validation is a task which aims to check if the model equations provide an adequate representation of the physical phenomenon/system of interest. The proper way to do this "validation check up" is through a direct comparison of the model responses with experimental data carefully obtained from the real system. It is an exercise in physics. In Fig. 5 the reader can see a schematic representation of the difference between the two notions.

An example in V&V: A skydiver jumps vertically in free fall, from a helicopter that is stopped in flight, from a height of $y_0 = 2000\,\mathrm{m}$ with velocity $v_0 = 0\,\mathrm{m/s}$. Such situation is illustrated in Fig. 6. Imagine we want to know the skydiver height in every moment of the fall. To do this we develop a (toy) model where the falling man is idealized as point mass $m = 70\,\mathrm{kg}$, under the action of gravity $g = 9.81\,\mathrm{m/s^2}$. The height at time t is denoted by $y(t)$.

The skydiver's height at time t can be determined through the following initial value problem (IVP)

$$\begin{aligned} m\,\ddot{y}(t) + m\,g &= 0, \\ \dot{y}(0) &= v_0, \\ y(0) &= y_0, \end{aligned} \tag{1}$$

where the upper dot is an abbreviation for a time derivative, i.e., $\dot{\square} := d\square/dt$. This toy model is obtained from Newton's 2nd law of motion and considers the weigh as the only force acting on the skydiver body.

Imagine that we have developed a computer code to integrate this IVP using a standard 4th order Runge-Kutta method [25]. The model response obtained with this computer code is shown in Fig. 7.

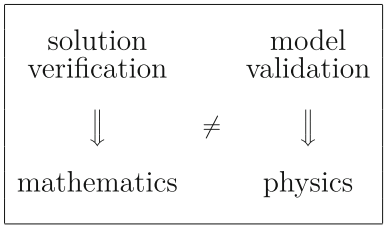


Fig. 5 The difference between verification and validation

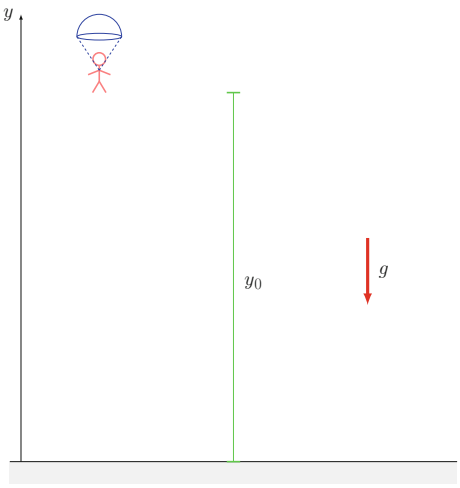


Fig. 6 V&V example: a skydiver in free fall from an initial height y_0

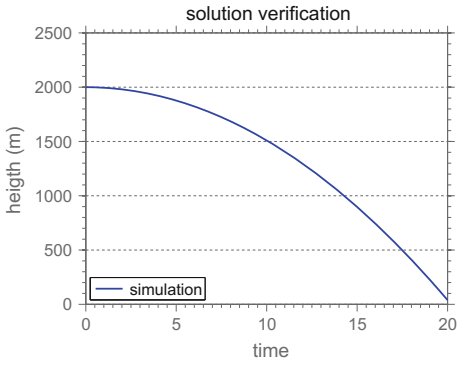


Fig. 7 Response obtained with the toy model

To check accuracy of the numerical method and its implementation we have at our disposal the analytical (reference) solution of the IVP, given by

$$y(t) = -\frac{1}{2} g\, t^2 + v_0\, t + y_0. \tag{2}$$

In Fig. 8a we can see the comparison between toy model response (solid blue curve —) and the reference solution (dashed red curve - - -). We note that both curves are in excellent agreement, but if we look at Fig. 8b, which shows the diference between numerical and analytical solutions, it is evident the effectiveness of the numerical method and the robustness of its implementation become ever clearer.

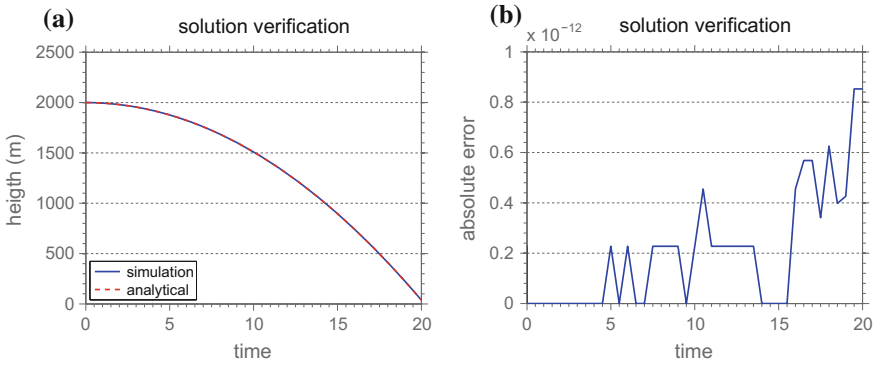


Fig. 8 **a** Solution verification: comparison between toy model response and reference solution; **b** absolute error of Runge-Kutta method approximation

Here the verification was made taking as reference the real solution of the model equation. In the most frequent case, the model equations solution is not known. In such a situation, the verification task can be performed, for instance, using the *method of manufactured solutions* [24, 26–28].

Now let's turn our attention to model validation, and compare simulation results with experimental data, such as shown in Fig. 9a. We note that the simulation is completely in disagreement with the experimental observations. In other words, the model does not provide an adequate representation of the real system behaviour.

The toy model above take into account the gravitational force which attracts the skydiver toward the ground, but neglects air resistance effects. This is the major reason for the observed discrepancy, the model deficiency (model uncertainty). If the air drag force effects are included, the improved model below is obtained

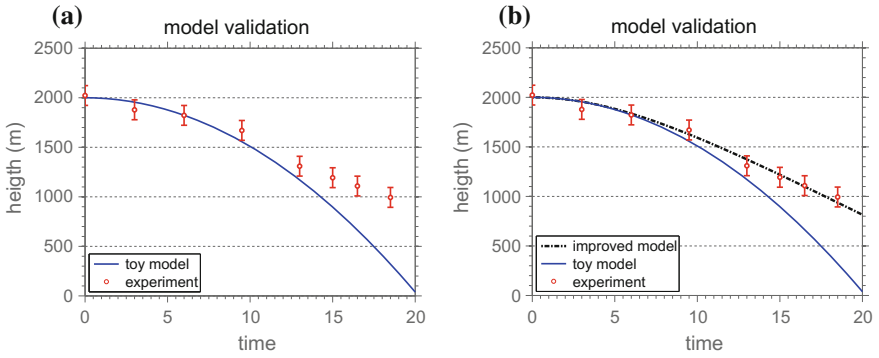


Fig. 9 **a** Model validation: comparison between experimental data and the toy model, **b** comparison between experimental data, the toy model, and the improved model

$$m\,\ddot{y}(t) + m\,g - \frac{1}{2}\rho\,A\,C_D\,(\dot{y}(t))^2 = 0, \tag{3}$$
$$\dot{y}(0) = v_0,$$
$$y(0) = y_0,$$

where ρ is the air mass density, A is the cross-sectional area of the falling body, and C_D is the (dimensionless) drag coefficient.

With this new model, a better agreement between simulation and experiment is expected. In Fig. 9b the reader can see the comparison between experimental data and the responses of both models, where we note that the improved model provides more plausible results.

An important message, implicit in this example, is that epistemic uncertainties can be reduced by increasing the actual knowledge about the phenomenon/system of interest [22, 24].

2.3 Two Approaches to Model Uncertainties

Being uncertainties in physical system the focus of stochastic modeling, two approaches are found in the scientific literature to deal with then: *probabilistic*, and *non-probabilistic*.

Probabilistic approach: This approach uses probability theory to model the physical system uncertainties as random mathematical objects. This approach is well-developed and very consistent from the mathematical foundations point of view for this reason, there is a consensus among the experts that it is preferable whenever possible to use it [4].

Non-probabilistic approach: This approach uses techniques such as interval analysis, fuzzy finite element, imprecise probabilities, evidence theory, probability bounds analysis, fuzzy probabilities, etc. In general these techniques are less suitable for problems in high stochastic dimension. Usually they are applied only when the probabilistic approach can not be used [4].

Because of their aleatory nature, data uncertainties are, quite naturally, well represented in a probabilistic environment. Thus, the *parametric probabilistic approach* is an appropriate method to describe this class of uncertainties. This procedure consists in describe the computational model random parameters as random objects (random variables, random vectors, random processes and/or random fields) and then consistently construct their joint probability distribution. Consequently, the model response becomes aleatory, and starts to be modeled by another random object, depending on the nature of the model equations. The model response is calculated using a stochastic solver. For further details, we recommend [4, 19, 29–31].

When model uncertainties are the focus of analysis, the non-probabilistic techniques receive more attention. Since the origin of this type of uncertainty is epistemic (lack of knowledge), it is not naturally described in a probabilistic setting. More details on non-probabilistic techniques can be seen in [32–34]. However, the use of probability theory for model uncertainties is still possible through a methodology called *nonparametric probabilistic approach*. This method, which also take into account the data uncertainty, was proposed in [35], and describes the mathematical operators in the computational model (not the parameters) as random objects. The probability distribution of these objects must be constructed in a consistent way, using the Principle of Maximum Entropy. The methodology lumps the model level of uncertainty into a single parameter, which can be identified by solving a parameter identification problem when (enough) experimental data is available. An overview of this technique can be seen in [19, 31].

A *generalized probabilistic approach* describing model and data uncertainties on different probability spaces, with some advantages, is presented in [36, 37].

3 A Brief on Probability Theory

This section presents a brief review of probability basic concepts. Such exposition is elementary, being insufficient for a solid understanding of the theory. Our objective is only to equip the reader with basic probabilistic vocabulary necessary to understand UQ scientific literature. For deeper studies on probability theory, we recommend the references [38–41].

3.1 Probability Space

The mathematical framework in which a random experiment is described consists of a triplet $(\Omega, \mathbb{\Sigma}, \mathbb{P})$, where Ω is called *sample space*, $\mathbb{\Sigma}$ is a σ-*algebra* over Ω, and $\mathbb{P}$ is a *probability measure*. The trio $(\Omega, \mathbb{\Sigma}, \mathbb{P})$ is called *probability space*.

Sample space: The set which contains all possible outcomes (events) for a certain random experiment is called sample space, being represented by Ω. An elementary event in Ω is denoted by ω. Sample spaces may contain a number of events that is finite, denumerable (countable infinite) or non-denumerable (non-countable infinity). The following three examples, respectively, illustrate the three situations:

Example 3.1 (finite sample space) Rolling a given cube-shaped fare die, where the faces are numbered from 1 through 6, we have $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Example 3.2 (denumerable sample space) Choosing randomly an integer even number, we have $\Omega = \{\ldots, -8, -6, -4, -2, 0, 2, 4, 6, 8, \ldots\}$.

Example 3.3 (non-denumerable sample space) Measuring the temperature (in Kelvin) at Rio de Janeiro city during the summer, we have $\Omega = [a, b] \subset [0, +\infty)$.

σ-algebra: In general, not all of the outcomes in Ω are of interest so that, in a probabilistic context, we need to pay attention only to the relevant events. Intuitively, the σ-algebra $\mathbb{\Sigma}$ is the set of relevant outcomes for a random experiment. Formally, $\mathbb{\Sigma}$ is σ-algebra if:

- $\phi \in \mathbb{\Sigma}$ (contains the empty set);
- for any $\mathcal{A} \in \mathbb{\Sigma}$ we also have $\mathcal{A}^c \in \mathbb{\Sigma}$ (closed under complementation);
- for any countable collections of $\mathcal{A}_i \in \mathbb{\Sigma}$, it is true that $\bigcup_{i=1}^{\infty} \mathcal{A}_i \in \mathbb{\Sigma}$ (closed under denumerable unions).

Example 3.4 Consider the experiment of rolling a die with sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ where we are interested in knowing if the result is odd or even. In this case, a suitable σ-algebra is $\mathbb{\Sigma} = \{\Omega, \{1, 3, 5\}, \{2, 4, 6\}, \phi\}$. On the other hand, if we are interested in knowing the upper face value after rolling, an adequate 2^{ω} (set of all subsets of ω). Different σ-algebras generate distinct probability spaces.

Probability measure: The probability measure is a function $\mathbb{P} : \mathbb{\Sigma} \to [0, 1] \subset \mathbb{R}$ which indicates the level of expectation that a certain event in $\mathbb{\Sigma}$ occurs. In technical language, $\mathbb{P}$ has the following properties:

- $\mathbb{P}\{\mathcal{A}\} \geq 0$ for any $\mathcal{A} \in \mathbb{\Sigma}$ (probability is nonnegative);
- $\mathbb{P}\{\Omega\} = 1$ (entire space has probability one);
- for any denumerable collection of mutually disjoint events $\mathcal{A}_i$, it is true that $\mathbb{P}\left\{\bigcup_{i=1}^{\infty} \mathcal{A}_i\right\} = \sum_{i=1}^{\infty} \mathbb{P}\left\{\mathcal{A}_i\right\}$.

Note that $\mathbb{P}\{\phi\} = 0$ (empty set has probability zero).

3.2 Random Variables

A mapping $\mathbb{X} : \Omega \to \mathbb{R}$ is called a *random variable* if the preimage of every real number under $\mathbb{X}$ is a relevant event, i.e.,

$$\mathbb{X}^{-1}(x) = \{\omega \in \Omega : \mathbb{X}(\omega) \leq x\} \in \mathbb{\Sigma}, \qquad \text{for every } x \in \mathbb{R}. \tag{4}$$

We denote a realization of $\mathbb{X}$ by $\mathbb{X}(\omega)$.

Random variables provide numerical characteristics of interesting events, in such a way that we can forget the sample space. In practice, when working with a probabilistic model, we are concerned only with the possible values of $\mathbb{X}$.

Example 3.5 The random experiment is now toss a two fare dice, then $\Omega = \{(d_1, d_2) : 1 \leq d_1 \leq 6 \text{ and } 1 \leq d_2 \leq 6\}$. Define the random variables $\mathbb{X}_1$ and $\mathbb{X}_2$ in such way that $\mathbb{X}_1(\omega) = d_1 + d_2$ and $\mathbb{X}_2(\omega) = d_1\, d_2$. The former is a numerical indicator of the sum of dice upper faces values, while the latter characterizes the product of these numbers.

3.3 Probability Distribution

The *probability distribution* of $\mathbb{X}$, denoted by $P_{\mathbb{X}}$, is defined as the probability of the elementary event $\{\mathbb{X} \leq x\}$, i.e.,

$$P_{\mathbb{X}}(x) = \mathbb{P}\{\mathbb{X} \leq x\}. \tag{5}$$

$P_{\mathbb{X}}$ has the following properties:

- $0 \leq P_{\mathbb{X}}(x) \leq 1$ (it is a probability);
- $P_{\mathbb{X}}$ is non-decreasing, and right-continuous;
- $\lim_{x\to-\infty} P_{\mathbb{X}}(x) = 0$, and $\lim_{x\to+\infty} P_{\mathbb{X}}(x) = 1$;

so that

$$P_{\mathbb{X}}(x) = \int_{\xi=-\infty}^{x} dP_{\mathbb{X}}(\xi), \tag{6}$$

and

$$\int_{\mathbb{R}} dP_{\mathbb{X}}(x) = 1. \tag{7}$$

$P_{\mathbb{X}}$ is also known as *cumulative distribution function* (CDF).

3.4 Probability Density Function

If the function $P_{\mathbb{X}}$ is differentiable, then we call its derivative the *probability density function* (PDF) of $\mathbb{X}$, using the notation $p_{\mathbb{X}}$.

Given that $p_{\mathbb{X}} = dP_{\mathbb{X}}/dx$, we have $dP_{\mathbb{X}}(x) = p_{\mathbb{X}}(x)\, dx$, and then

$$P_{\mathbb{X}}(x) = \int_{\xi=-\infty}^{x} p_{\mathbb{X}}(\xi)\, d\xi. \tag{8}$$

The PDF is a function $p_{\mathbb{X}} : \mathbb{R} \to [0, +\infty)$ such that

$$\int_{\mathbb{R}} p_{\mathbb{X}}(x)\, dx = 1. \tag{9}$$

3.5 Mathematical Expectation Operator

Given a function $g : \mathbb{R} \to \mathbb{R}$, the composition of g with the random variable $\mathbb{X}$ is also a random variable $g(\mathbb{X})$.

The *mathematical expectation* of $g(\mathbb{X})$ is defined by

$$\mathbb{E}\{g(\mathbb{X})\} = \int_{\mathbb{R}} g(x)\, p_{\mathbb{X}}(x)\, dx. \tag{10}$$

With the aid of this operator, we define

$$\begin{aligned} m_{\mathbb{X}} &= \mathbb{E}\{\mathbb{X}\} \\ &= \int_{\mathbb{R}} x\, p_{\mathbb{X}}(x)\, dx, \end{aligned} \tag{11}$$

$$\begin{aligned} \sigma_{\mathbb{X}}^2 &= \mathbb{E}\left\{\left(\mathbb{X} - m_{\mathbb{X}}\right)^2\right\} \\ &= \int_{\mathbb{R}} (x - m_{\mathbb{X}})^2\, p_{\mathbb{X}}(x)\, dx, \end{aligned} \tag{12}$$

and

$$\sigma_{\mathbb{X}} = \sqrt{\sigma_{\mathbb{X}}^2}, \tag{13}$$

which are the *mean value*, *variance*, and *standard deviation* of $\mathbb{X}$, respectively. Note further that

$$\sigma_{\mathbb{X}}^2 = \mathbb{E}\left\{\mathbb{X}^2\right\} - m_{\mathbb{X}}^2. \tag{14}$$

The ratio between standard deviation and mean value is called *coefficient of variation* of $\mathbb{X}$

$$\delta_{\mathbb{X}} = \frac{\sigma_{\mathbb{X}}}{m_{\mathbb{X}}}, \quad m_{\mathbb{X}} \neq 0. \tag{15}$$

These scalar values are indicators of the random variable behaviour. Specifically, the mean value $m_{\mathbb{X}}$ is a central tendency indicator, while variance $\sigma_{\mathbb{X}}^2$ and standard deviation $\sigma_{\mathbb{X}}$ are measures of dispersion around the mean. The difference in these dispersion measures is that $\sigma_{\mathbb{X}}$ has the same unit as $m_{\mathbb{X}}$ while $\sigma_{\mathbb{X}}^2$ is measured in $m_{\mathbb{X}}$

unit squared. Once it is dimensionless, the coefficient of variation is a standardized measure of dispersion.

For our purposes, it is also convenient to define the entropy of $p_{\mathbb{X}}$

$$\mathrm{S}\left(p_{\mathbb{X}}\right) = -\mathbb{E}\left\{\ln\left(p_{\mathbb{X}}(\mathbb{X})\right)\right\}, \tag{16}$$

which (see Eq. 10) is equivalent to

$$\mathrm{S}\left(p_{\mathbb{X}}\right) = -\int_{\mathbb{R}} p_{\mathbb{X}}(x)\, \ln\left(p_{\mathbb{X}}(x)\right) dx. \tag{17}$$

Entropy provides a measure for the level of uncertainty of $p_{\mathbb{X}}$ [42].

3.6 Second-Order Random Variables

The mapping $\mathbb{X}$ is a *second-order random variable* if the expectation of its square (second-order moment) is finite, i.e.,

$$\mathbb{E}\left\{\mathbb{X}^2\right\} < +\infty. \tag{18}$$

The inequality expressed in (18) implies that $\mathbb{E}\{\mathbb{X}\} < +\infty$ ($m_{\mathbb{X}}$ is also finite). Consequently, with the aid of Eq. (14), we see that a second-order random variable $\mathbb{X}$ has finite variance, i.e., $\sigma_{\mathbb{X}}^2 < +\infty$.

This class of random variables is very relevant for stochastic modeling, once, for physical considerations, typical random parameters in physical systems have finite variance.

3.7 Joint Probability Distribution

Given the random variables $\mathbb{X}$ and $\mathbb{Y}$, the *joint probability distribution* of $\mathbb{X}$ and $\mathbb{Y}$, denoted by $P_{\mathbb{X}\mathbb{Y}}$, is defined as

$$P_{\mathbb{X}\mathbb{Y}}(x, y) = \mathbb{P}\left\{\{\mathbb{X} \leq x\} \cap \{\mathbb{Y} \leq y\}\right\}. \tag{19}$$

The function $P_{\mathbb{X}\mathbb{Y}}$ has the following properties:

- $0 \leq P_{\mathbb{X}\mathbb{Y}}(x, y) \leq 1$ (it is a probability);
- $P_{\mathbb{X}}(x) = \lim_{y\to+\infty} P_{\mathbb{X}\mathbb{Y}}(x, y)$, and $P_{\mathbb{Y}}(y) = \lim_{x\to+\infty} P_{\mathbb{X}\mathbb{Y}}(x, y)$ (marginal distributions are limits);

such that

$$P_{\mathbb{X}\mathbb{Y}}(x,y) = \int_{\xi=-\infty}^{x} \int_{\eta=-\infty}^{y} dP_{\mathbb{X}\mathbb{Y}}(\xi,\eta), \tag{20}$$

and

$$\int \int_{\mathbb{R}^2} dP_{\mathbb{X}\mathbb{Y}}(x,y) = 1. \tag{21}$$

$P_{\mathbb{X}\mathbb{Y}}$ is also known as *joint cumulative distribution function*.

3.8 Joint Probability Density Function

If the partial derivative $\partial^2 P_{\mathbb{X}\mathbb{Y}}/\partial x\,\partial y$ exists, for any x and y, then it is called *joint probability density function* of $\mathbb{X}$ and $\mathbb{Y}$, being denoted by

$$p_{\mathbb{X}\mathbb{Y}}(x,y) = \frac{\partial^2 P_{\mathbb{X}\mathbb{Y}}}{\partial x\,\partial y}(x,y). \tag{22}$$

Hence, we can write $dP_{\mathbb{X}\mathbb{Y}}(x,y) = p_{\mathbb{X}\mathbb{Y}}(x,y)\,dy\,dx$, so that

$$P_{\mathbb{X}\mathbb{Y}}(x,y) = \int_{\xi=-\infty}^{x} \int_{\eta=-\infty}^{y} p_{\mathbb{X}\mathbb{Y}}(\xi,\eta)\,d\eta\,d\xi. \tag{23}$$

The joint PDF is a function $p_{\mathbb{X}\mathbb{Y}} : \mathbb{R} \to [0,+\infty)$ which satisfies

$$\int \int_{\mathbb{R}^2} p_{\mathbb{X}\mathbb{Y}}(x,y)\,dy\,dx = 1. \tag{24}$$

3.9 Conditional Probability

Consider the pair of random events $\{\mathbb{X} \leq x\}$ and $\{\mathbb{Y} \leq y\}$, where the probability of occurrence of the second one is non-zero, i.e., $\mathbb{P}\{\{\mathbb{Y} \leq y\}\} > 0$. The *conditional probability* of event $\{\mathbb{X} \leq x\}$, given the occurrence of event $\{\mathbb{Y} \leq y\}$, is defined as

$$\mathbb{P}\left\{\{\mathbb{X} \leq x\} \,\middle|\, \{\mathbb{Y} \leq y\}\right\} = \frac{\mathbb{P}\{\{\mathbb{X} \leq x\} \cap \{\mathbb{Y} \leq y\}\}}{\mathbb{P}\{\{\mathbb{Y} \leq y\}\}}. \tag{25}$$

3.10 Independence of Random Variables

The event $\{\mathbb{X} \leq x\}$ is said to be *independent* of event $\{\mathbb{Y} \leq y\}$ if the occurrence of the former does not affect the occurrence of the later, i.e.,

$$\mathbb{P}\left\{\{\mathbb{X} \leq x\} \,\middle|\, \{\mathbb{Y} \leq y\}\right\} = \mathbb{P}\left\{\{\mathbb{X} \leq x\}\right\}. \tag{26}$$

Consequently, if the random variables $\mathbb{X}$ and $\mathbb{Y}$ are independent, from Eq. (25) we see that

$$\mathbb{P}\left\{\{\mathbb{X} \leq x\} \cap \{\mathbb{Y} \leq y\}\right\} = \mathbb{P}\left\{\mathbb{X} \leq x\right\} \mathbb{P}\left\{\mathbb{Y} \leq y\right\}. \tag{27}$$

This also implies that

$$P_{\mathbb{X}\mathbb{Y}}(x, y) = P_{\mathbb{X}}(x)\, P_{\mathbb{Y}}(y), \tag{28}$$

and

$$p_{\mathbb{X}\mathbb{Y}}(x, y) = p_{\mathbb{X}}(x)\, p_{\mathbb{Y}}(y). \tag{29}$$

3.11 Random Process

A *random process* $\mathbb{U}$, indexed by $t \in \mathcal{T}$, is a mapping

$$\mathbb{U} : (t, \omega) \in \mathcal{T} \times \Omega \to \mathbb{U}(t, \omega) \in \mathbb{R}, \tag{30}$$

such that, for fixed t, the output is a random variable $\mathbb{U}(t, \cdot)$, while for fixed ω, $\mathbb{U}(\cdot, \omega)$ is a function of t. In other words, it is a collection of random variables indexed by a parameter. Roughly speaking, a random process, also called *stochastic process*, can be thought of as a time-dependent random variable.

4 Parametric Probabilistic Modeling of Uncertainties

This section discusses the use of the parametric probabilistic approach to describe uncertainties in physical systems. Our goal is to provide the reader with some key ideas behind this approach and call attention to the fundamental issues that must be taken into account. The exhibition is based on [13, 15] and use a simplistic example to discuss the theory.

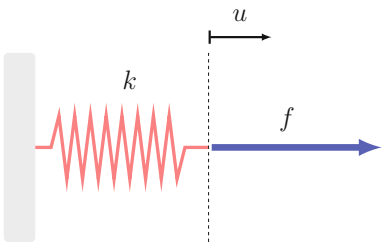


Fig. 10 Mechanical system composed by a fixed spring and a constant force

4.1 A Simplistic Stochastic Mechanical System

Consider the mechanical system which consists of a spring fixed on the left side of a wall and being pulled by a constant force on the right side (Fig. 10). The spring stiffness is k, the force is represented by f, and the spring displacement is denoted by u. A mechanical-mathematical model to describe this system behaviour is given by

$$k\,u = f, \tag{31}$$

from where we get the system response

$$u = k^{-1} f. \tag{32}$$

4.2 Stochastic Model for Uncertainties Description

We are interested in studying the case where the above mechanical system is subject to uncertainties on the stiffness parameter k. To describe the random behaviour of the mechanical system, we employ the parametric probabilistic approach.

Let us use the probability space $(\Omega, \mathbb{\Sigma}, \mathbb{P})$, where the stiffness k is modeled as the random variable $\mathbb{K} : \Omega \to \mathbb{R}$. Therefore, due to result of the relationship imposed by Eq. (32), the displacement u is also uncertain, being modeled as a random variable $\mathbb{U} : \Omega \to \mathbb{R}$, which respects the equilibrium condition given by the following stochastic equation

$$\mathbb{K}\,\mathbb{U} = f. \tag{33}$$

It is reasonable to assume that the deterministic model is minimally representative, and corresponds to the mean of $\mathbb{K}$, i.e., $m_{\mathbb{K}} = k$. Additionally, for physical reasons, $\mathbb{K}$ must have a finite variance. Thus, $\mathbb{K}$ is assumed to be a second-order random variable, i.e., $\mathbb{E}\left\{\mathbb{K}^2\right\} < +\infty$.

4.3 The Importance of Knowing the PDF

Now that we have the random parameter described in a probabilistic context, and a stochastic model for the system, we can ask ourselves some questions about the system response. For instance, to characterize the system response central tendency, it is of interest to know the mean of $\mathbb{U}$, denoted by $m_{\mathbb{U}}$.

Since $m_{\mathbb{K}}$ is a known information about $\mathbb{K}$ (but $p_{\mathbb{K}}$ is unknown), we can ask ourselves: *Is it possible to compute $m_{\mathbb{U}}$ with this information only?* The answer for this question is negative. The reason is that $\mathbb{U} = \mathbb{K}^{-1} f$, so that

$$\begin{aligned} m_{\mathbb{U}} &= \mathbb{E}\left\{\mathbb{K}^{-1} f\right\} \\ &= \int_{\mathbb{R}} k^{-1} f \, p_{\mathbb{K}}(k) \, dk, \end{aligned}$$

and the last integral can only be calculated if $p_{\mathbb{K}}$ is known. Once the map $g(k) = k^{-1} f$ is nonlinear, $\mathbb{E}\{g(\mathbb{K})\} \neq g(\mathbb{E}\{\mathbb{K}\})$.

Conclusion: *In order to obtain any statistical information about model response, it is absolutely necessary to know the probability distribution of model parameters.*

4.4 Why Can't We Arbitrate Distributions?

As the knowledge of the probability distribution of $\mathbb{K}$ is necessary, let's assume that it is Gaussian distributed. In this way,

$$p_{\mathbb{K}}(k) = \frac{1}{\sqrt{2\pi \, \sigma_{\mathbb{K}}^2}} \exp\left\{ -\frac{(k - m_{\mathbb{K}})^2}{2 \, \sigma_{\mathbb{K}}^2} \right\}, \tag{34}$$

whose support is the entire real line, i.e., $\mathtt{Supp}\, p_{\mathbb{K}} = (-\infty, +\infty)$.

The attentive reader may question, at this point, that from the physical point of view, make no sense use a Gaussian distribution to model a stiffness parameter, since $\mathbb{K}$ is always positive. This is true and makes the arbitrary choice of a Gaussian distribution inappropriate. However, this is not the only reason against this choice.

For physical considerations, it is necessary that the model response $\mathbb{U}$ be a second-order (finite variance) random variable, i.e., . $\mathbb{E}\left\{\mathbb{U}^2\right\} < +\infty$. *Is this possible when we arbitrate the probability distribution as Gaussian?* No way! Just do a simple calculation

$$\begin{aligned}\mathbb{E}\left\{\mathbb{U}^2\right\} &= \mathbb{E}\left\{\mathbb{K}^{-2} f^2\right\} \\ &= \int_{\mathbb{R}} k^{-2} f^2 \, p_{\mathbb{K}}(k)\, dk \\ &= \int_{k=-\infty}^{+\infty} k^{-2} f^2 \left(\frac{1}{\sqrt{2\pi\,\sigma_{\mathbb{K}}^2}} \exp\left\{ -\frac{(k-m_{\mathbb{K}})^2}{2\,\sigma_{\mathbb{K}}^2} \right\} \right) dk \\ &= +\infty.\end{aligned} \tag{35}$$

In fact, we also have $\mathbb{E}\{\mathbb{U}\} = m_{\mathbb{U}} = +\infty$.

The Gaussian distribution is a bad choice since $\mathbb{K}$ must be a positive-valued random variable (almost sure). Thus, we know the following information about $\mathbb{K}$:

- $\mathtt{Supp}\, p_{\mathbb{K}} \subseteq (0, +\infty) \Longleftrightarrow \mathbb{K} > 0 \;\; a.s.$
- $m_{\mathbb{K}} = k > 0$ is known
- $\mathbb{E}\left\{\mathbb{K}^2\right\} < +\infty$

All these requirements are verified by the exponential distribution, in which the PDF is given by the function

$$p_{\mathbb{K}}(k) = \mathbb{1}_{(0,+\infty)}(k) \frac{1}{m_{\mathbb{K}}} \exp\left\{ -\frac{k}{m_{\mathbb{K}}} \right\}, \tag{36}$$

where $\mathbb{1}_{(0,+\infty)}$ the indicator function of the interval $(0, +\infty)$.

However, we still have

$$\begin{aligned}\mathbb{E}\left\{\mathbb{U}^2\right\} &= \mathbb{E}\left\{\mathbb{K}^{-2} f^2\right\} \\ &= \int_{\mathbb{R}} k^{-2} f^2 \, p_{\mathbb{K}}(k)\, dk \\ &= \int_{k=0}^{+\infty} k^{-2} f^2 \left(\frac{1}{m_{\mathbb{K}}} \exp\left\{ -\frac{k}{m_{\mathbb{K}}} \right\} \right) dk \\ &= +\infty,\end{aligned} \tag{37}$$

once the function $k \mapsto k^{-2}$ diverges in $k = 0$. Thus, in order to $\mathbb{E}\left\{\mathbb{U}^2\right\} < +\infty$, we must have $\mathbb{E}\left\{\mathbb{K}^{-2}\right\} < +\infty$.

Conclusion: *Arbitrate probability distributions for parameters can generate a stochastic model that is inconsistent from the physical/mathematical point of view.*

4.5 An Acceptable Distribution

In short, an adequate distribution must satisfy the conditions below

- $\texttt{Supp}\, p_{\mathbb{K}} \subseteq (0, +\infty) \Longrightarrow \mathbb{K} > 0 \;\; a.s.$
- $m_{\mathbb{K}} = k > 0$ is known
- $\mathbb{E}\left\{\mathbb{K}^2\right\} < +\infty$
- $\mathbb{E}\left\{\mathbb{K}^{-2}\right\} < +\infty$.

The gamma distribution satisfies all the conditions above so that it is an acceptable choice. Its PDF is written as

$$p_{\mathbb{K}}(k) = \mathbb{1}_{(0,+\infty)}(k) \frac{1}{m_{\mathbb{K}}} \frac{\delta_{\mathbb{K}}^{-2\delta_{\mathbb{K}}^{-2}}}{\Gamma\left(\delta_{\mathbb{K}}^{-2}\right)} \left(k/m_{\mathbb{K}}\right)^{\delta_{\mathbb{K}}^{-2}-1} \exp\left\{-\frac{k/m_{\mathbb{K}}}{\delta_{\mathbb{K}}^2}\right\}, \tag{38}$$

where $0 \leq \delta_{\mathbb{K}} = \sigma_{\mathbb{K}}/m_{\mathbb{K}} < 1/\sqrt{2}$ is a dispersion parameter, and Γ denotes the gamma function

$$\Gamma(\alpha) = \int_{t=0}^{+\infty} t^{\alpha-1}\, e^{-t}\, dt. \tag{39}$$

Conclusion: *Probability distributions for model parameters must be objectively constructed (never arbitrated), and take into account all available information about the parameters.*

4.6 How to Safely Specify a Distribution?

In the previous example, we have chosen a suitable probability distribution by verifying if the candidate distributions satisfy the constraints imposed by physical and mathematical properties of the model parameter/response. However, this procedure is not practical and does not provide a unique distribution as a possible choice. For instance, in the spring example, uniform, lognormal and an infinitude of other distributions are also acceptable (compatible with the restrictions).

Thus, it is natural to ask ourselves if it is possible to construct a consistent stochastic model in a systematic way. The answer for this question is affirmative, and the objective procedure to be used depends on the scenario.

Scenario 1: large amount of experimental data is available

The usual procedure in this case employs nonparametric statistical estimation to construct the random parameter distribution from the available data [13, 15, 43].

Suppose we want to estimate the probability distribution of a random variable $\mathbb{X}$, and for that we have N independent samples of $\mathbb{X}$, respectively denoted by X^1, X^2, …, X^N.

Assuming, without loss of generality, that $X^1 < X^2 < \cdots < X^N$, we consider an estimator for $P_{\mathbb{X}}(x)$ given by

$$\widehat{P}_N(x) = \frac{1}{N} \sum_{n=1}^{N} \mathcal{H}\left(x - X^n\right), \tag{40}$$

where $\mathcal{H}$ is defined as

$$\mathcal{H}\left(x - X^n\right) = \begin{cases} 1 & \text{if } x \geq X^n \\ 0 & \text{if } x < X^n. \end{cases} \tag{41}$$

This estimator, which is mean-square consistent

$$\mathbb{E}\left\{\widehat{P}_N(x)\right\} = P_{\mathbb{X}}(x), \tag{42}$$

and unbiased

$$\lim_{N \to +\infty} \mathbb{E}\left\{\left(\widehat{P}_N(x) - P_{\mathbb{X}}(x)\right)^2\right\} = 0, \tag{43}$$

is known as the *empirical distribution function* or the *empirical CDF* [13, 15, 43, 44].

If the random variable admits a PDF, it is more common to estimate its probability distribution using a *histogram*, that is an estimator for $p_{\mathbb{X}}(x)$. To construct such a histogram, the first step is to divide the random variable support into a denumerable number of bins $\mathcal{B}_m$, where

$$\mathcal{B}_m = [(m-1)\,h, m\,h], \qquad m \in \mathbb{Z}, \tag{44}$$

being h the *bin width*. Then we count the number of samples in each of the bins $\mathcal{B}_m$, denoting this number by ν_m. After that, we normalize the counter (dividing by Nh) to obtain the *normalized relative frequency* $\nu_m/(Nh)$. Finally, for each bin $\mathcal{B}_m$, we plot a vertical bar with height $\nu_m/(Nh)$ [43, 44].

In analytical terms (see [43, 44]) we can write this as PDF estimator as

$$\widehat{p}_N(x) = \frac{1}{N\,h} \sum_{m=-\infty}^{+\infty} \nu_m \, \mathbb{1}_{\mathcal{B}_m}(x), \tag{45}$$

where $\mathbb{1}_{\mathcal{B}_m}(x)$ is the *indicator function* of $\mathcal{B}_m$, defined as

$$\mathbb{1}_{\mathcal{B}_m}(x) = \begin{cases} 1 & \text{if } x \in \mathcal{B}_m \\ 0 & \text{if } x \notin \mathcal{B}_m. \end{cases} \tag{46}$$

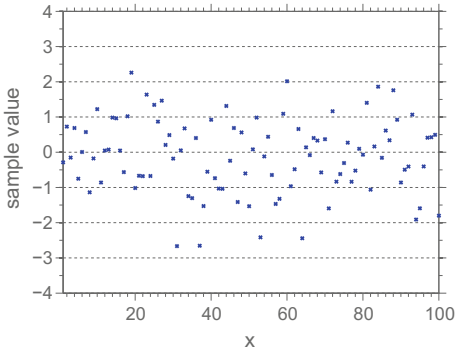


Fig. 11 These samples are realizations of a standard Gaussian random variable

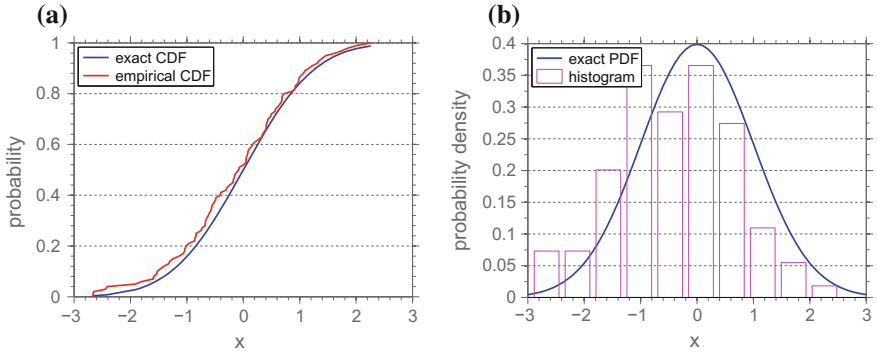


Fig. 12 **a** Estimators for the probability distribution of $\mathbb{X}$: the empirical CDF, and **b** a histogram

Both estimators above are easily constructed, but they require a large number of samples in order to obtain a reasonable approximation [43, 44].

In practice, these estimators are used when we do not know the random variable distribution. However, to illustrate the use of these tools, let us consider a dataset with $N = 100$ samples obtained from the (standard) Gaussian random variable $\mathbb{X}$, with zero mean and unity standard deviation. Such samples are illustrated in Fig. 11. Considering these samples, we can construct the two estimators shown in Fig. 12, with the empirical CDF on the left and a histogram on the right.

Scenario 2: little or even none experimental data is available

When very little or no experimental data is available, to the best of the author's knowledge, the most conservative approach uses the *Maximum Entropy Principle* (MEP) [15, 45, 46, 48], with parametric statistical estimation, to construct the random parameter distribution. If no experimental data is available, this approach takes into account only theoretical information which can be inferred from the model physics and its mathematical structure to specify the desired distribution.

The MEP can be stated as follows: *Among all the (infinite) probability distributions, consistent with the known information about a random parameter, the most unbiased is the one which corresponds to the maximum of entropy PDF.*

Using it to specify the distribution of a random variable $\mathbb{X}$ presupposes finding the unique PDF which maximizes the entropy (objective function)

$$\mathrm{S}\left(p_{\mathbb{X}}\right) = -\int_{\mathbb{R}} p_{\mathbb{X}}(x)\,\ln\left(p_{\mathbb{X}}(x)\right)\,dx, \tag{47}$$

respecting $N+1$ constraints (known information) given by

$$\int_{\mathbb{R}} g_k\left(\mathbb{X}\right) p_{\mathbb{X}}(x)\,dx = \mu_k, \qquad k = 0, \ldots, N, \tag{48}$$

where g_k are known real functions, with $g_0(x) = 1$, and μ_k are known real values, being $\mu_0 = 1$. The restriction associated with $k = 0$ corresponds to the normalization condition of $p_{\mathbb{X}}$, while the other constraints, typically, but not exclusively, represent statistical moments of $\mathbb{X}$.

To solve this problem, the method of Lagrange multipliers is employed, and introduces other $(N+1)$ unknown real parameters λ_k (Lagrange multipliers). We can show that if this optimization problem has a solution, it actually corresponds to a maximum and is unique, being written as

$$p_{\mathbb{X}}(x) = \mathbb{1}_{\mathcal{K}}(x)\,\exp\left(-\lambda_0\right)\,\exp\left(-\sum_{k=1}^{N}\lambda_k\,g_k(x)\right), \tag{49}$$

where $\mathcal{K} = \mathtt{Supp}\,p_{\mathbb{X}}$ here denotes the support of $p_{\mathbb{X}}$, and $\mathbb{1}_{\mathcal{K}}(x)$ is the indicator function of $\mathcal{K}$.

The Lagrange multipliers, which depend on μ_k and $\mathcal{K}$, are identified with the aid of the restriction defined in Eq. (48) using techniques of parametric statistics.

4.7 Using the Maximum Entropy Principle

In this section we exemplify the use of the MEP to consistently specify the probability distribution of a random variable $\mathbb{X}$.

Suppose that $\mathtt{Supp}\,p_{\mathbb{X}} = [a, b]$ is the only information we know about $\mathbb{X}$. In this case, a consistent (unbiased) probability distribution for $\mathbb{X}$ is obtained solving the following optimization problem:

Maximize

$$\begin{aligned} \mathrm{S}\left(p_{\mathbb{X}}\right) &= -\int_{\mathbb{R}} p_{\mathbb{X}}(x)\, \ln\left(p_{\mathbb{X}}(x)\right) dx \\ &= -\int_{x=a}^{b} p_{\mathbb{X}}(x)\, \ln\left(p_{\mathbb{X}}(x)\right) dx, \end{aligned}$$

subjected to the constraint

$$\begin{aligned} 1 &= \int_{\mathbb{R}} p_{\mathbb{X}}(x)\, dx \\ &= \int_{x=a}^{b} p_{\mathbb{X}}(x)\, dx. \end{aligned}$$

To solve this optimization problem, first we define the Lagrangian

$$\mathcal{L}\left(p_{\mathbb{X}}, \lambda_0\right) = -\int_{x=a}^{b} p_{\mathbb{X}}(x)\, \ln\left(p_{\mathbb{X}}(x)\right) dx \; - (\lambda_0 - 1)\left(\int_{x=a}^{b} p_{\mathbb{X}}(x)\, dx - 1\right), \quad (50)$$

where $\lambda_0 - 1$ is the associated Lagrange multiplier. It is worth mentioning that λ_0 depends on the known information about $\mathbb{X}$, i.e. $\lambda_0 = \lambda_0(a, b)$.

Then we impose the necessary conditions for an extreme

$$\frac{\partial \mathcal{L}}{\partial p_{\mathbb{X}}}\left(p_{\mathbb{X}}, \lambda_0\right) = 0, \;\; \text{and} \;\; \frac{\partial \mathcal{L}}{\partial \lambda_0}\left(p_{\mathbb{X}}, \lambda_0\right) = 0, \quad (51)$$

whence we conclude that

$$p_{\mathbb{X}}(x) = \mathbb{1}_{[a,b]}(x)\, e^{-\lambda_0}, \;\; \text{and} \;\; \int_{\mathbb{R}} p_{\mathbb{X}}(x)\, dx = 1. \quad (52)$$

The first equation in Eq. (52) provides the PDF of $\mathbb{X}$ in terms of the Lagrange multiplier λ_0, while the second equation corresponds to the known information about this random variable (the normalization condition).

In order to represent $p_{\mathbb{X}}$ in terms of the known information (a and b), we need to find the dependence of λ_0 with respect to these parameters. To this end, let's go to replace the expression of $p_{\mathbb{X}}$ into the second equation of Eq. (52), so that

$$\int_{\mathbb{R}} \mathbb{1}_{[a,b]}(x)\, e^{-\lambda_0}\, dx = 1 \implies e^{-\lambda_0}(b - a) = 1, \implies e^{-\lambda_0} = \frac{1}{b - a}, \quad (53)$$

from where we get

$$p_{\mathbb{X}}(x) = \mathbb{1}_{[a,b]}(x)\, \frac{1}{b - a}, \quad (54)$$

Table 1 Maximum entropy distributions for given known information

Support	Known information	Maximum entropy PDF
$[a, b]$	–	$p_{\mathbb{X}}(x) = \mathbb{1}_{[a,b]}(x)\, \frac{1}{b-a}$ **(uniform in** $[a, b]$**)**
$[a, b]$	$\mathbb{E}\{\mathbb{X}\} = m_{\mathbb{X}} \in [a, b]$	$p_{\mathbb{X}}(x) = \mathbb{1}_{[a,b]}(x) \exp\left(\lambda_0 - x\,\lambda_1\right)$ $\lambda_0 = \lambda_0(a, b, m_{\mathbb{X}})$ $\lambda_1 = \lambda_1(a, b, m_{\mathbb{X}})$
$[a, b]$	$\mathbb{E}\{\mathbb{X}\} = m_{\mathbb{X}} \in [a, b]$ $\mathbb{E}\left\{\mathbb{X}^2\right\} = m_{\mathbb{X}}^2 + \sigma_{\mathbb{X}}^2$	$p_{\mathbb{X}}(x) = \mathbb{1}_{[a,b]}(x) \exp\left(\lambda_0 - x\,\lambda_1 - x^2\,\lambda_2\right)$ $\lambda_0 = \lambda_0(a, b, m_{\mathbb{X}}, \sigma_{\mathbb{X}})$ $\lambda_1 = \lambda_1(a, b, m_{\mathbb{X}}, \sigma_{\mathbb{X}})$ $\lambda_2 = \lambda_2(a, b, m_{\mathbb{X}}, \sigma_{\mathbb{X}})$
$[0, 1]$	$\mathbb{E}\{\ln(\mathbb{X})\} = p,\ \|p\| < +\infty$ $\mathbb{E}\{\ln(1 - \mathbb{X})\} = q,\ \|q\| < +\infty$	$p_{\mathbb{X}}(x) = \mathbb{1}_{[0,1]}(x)\, \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}$ $a = \left(m_{\mathbb{X}}/\delta_{\mathbb{X}}^2\right)\left(1/m_{\mathbb{X}} - \delta_{\mathbb{X}}^2 - 1\right)$ $b = a\left(1/m_{\mathbb{X}} - 1\right)$ **(beta with shape parameters** a **and** b**)**
$(0, +\infty)$	$\mathbb{E}\{\mathbb{X}\} = m_{\mathbb{X}} > 0$	$p_{\mathbb{X}}(x) = \mathbb{1}_{(0,+\infty)}(x)\, \frac{1}{m_{\mathbb{X}}} \exp\left(-\frac{x}{m_{\mathbb{X}}}\right)$ **(exponential with mean** $m_{\mathbb{X}}$**)**
$(0, +\infty)$	$\mathbb{E}\{\mathbb{X}\} = m_{\mathbb{X}} > 0$ $\mathbb{E}\{\ln(\mathbb{X})\} = q,\ \|q\| < +\infty$	$p_{\mathbb{X}}(x) = \mathbb{1}_{(0,+\infty)}(x) \frac{1}{m_{\mathbb{X}}} \frac{\delta_{\mathbb{X}}^{-2\delta_{\mathbb{X}}^{-2}}}{\Gamma(\delta_{\mathbb{X}}^{-2})} \left(x/m_{\mathbb{X}}\right)^{\delta_{\mathbb{X}}^{-2}-1} \exp\left\{-\frac{x/m_{\mathbb{X}}}{\delta_{\mathbb{X}}^2}\right\}$ **(gamma with mean** $m_{\mathbb{X}}$ **and variation coefficient** $\delta_{\mathbb{X}}$**)**
$(0, +\infty)$	$\mathbb{E}\{\ln(\mathbb{X})\} = \mu \in \mathbb{R}$ $\mathbb{E}\left\{(\ln(\mathbb{X}) - \mu)^2\right\} = \sigma^2,\ \sigma > 0$	$p_{\mathbb{X}}(x) = \frac{1}{x\sqrt{2\pi\,\sigma^2}} \exp\left\{-\frac{(\ln(x)-\mu)^2}{2\,\sigma^2}\right\}$ $\mu = \ln\left(m_{\mathbb{X}}/\sqrt{1+\delta_{\mathbb{X}}^2}\right)$ $\sigma = \sqrt{\ln\left(1+\delta_{\mathbb{X}}^2\right)}$ **(lognormal with location** μ **and scale** σ**)**
$(-\infty, +\infty)$	$\mathbb{E}\{\mathbb{X}\} = m_{\mathbb{X}} \in \mathbb{R}$ $\mathbb{E}\left\{\mathbb{X}^2\right\} = m_{\mathbb{X}}^2 + \sigma_{\mathbb{X}}^2$	$p_{\mathbb{X}}(x) = \frac{1}{\sqrt{2\pi\,\sigma_{\mathbb{X}}^2}} \exp\left\{-\frac{(x-m_{\mathbb{X}})^2}{2\,\sigma_{\mathbb{X}}^2}\right\}$ **(normal with mean** $m_{\mathbb{X}}$ **and variance** $\sigma_{\mathbb{X}}^2$**)**

which corresponds to the PDF of a uniform distributed random variable over the interval $[a, b]$.

Other cases of interest, where the optimization problem solution is a known distribution, are shown in Table 1. In the fourth line of this table the maximum entropy PDF corresponds to a gamma distribution. Once any gamma random variable has finite variance, and $\mathbb{E}\{\ln(\mathbb{X})\} = q,\ |q| < +\infty$, which implies $\mathbb{E}\left\{\mathbb{K}^{-2}\right\} < +\infty$, the known information in this case is equivalent to those listed in Sect. 4.5, required to be satisfied by the distribution of $\mathbb{K}$. For this reason, we presented the gamma distribution as the acceptable choice in Sect. 4.5. It corresponds to the most unbiased choice for that set of information.

For other possible applications of the maximum entropy principle and to go deeper into the underlying mathematics, we recommend the reader to see the references [15, 47–54].

5 Calculation of Uncertainty Propagation

Once one or more of the model parameters are described as random objects, the system response itself becomes random. To understand how the variabilities are transformed by the model, and influence in the response distribution, is a key issue in UQ, known as *uncertainty propagation problem*. This problem can only be attacked after the construction of a consistent stochastic model.

Very succinctly, we understand the uncertainty propagation problem as to determine the probability distribution of model response once we know the distribution of model input/parameters. A schematic representation of this problem is can be seen in Fig. 13.

The methods for calculation of uncertainty of propagation are classified into two types: *non-intrusive* and *intrusive*.

Non-intrusive methods: These methods of stochastic calculation obtain the random problem response by running an associated deterministic problem multiple times (they are also known as sampling methods). In order to use a non-intrusive method, it is not necessary to implement the stochastic model in a new computer code. If a deterministic code to simulate the deterministic model is available, the stochastic simulation can then be performed by running the deterministic program several times, changing only the parameters that are randomly generated [55].

Intrusive methods: In this class of stochastic solvers, the random problem response is obtained by running a customized computer code only once. This code is not based on the associated deterministic model, but on a stochastic version of the computational model [2].

5.1 Monte Carlo Method: A Non-intrusive Approach

The most frequently used technique to compute the propagation of uncertainties of random parameters through a model is the Monte Carlo (MC) method, originally proposed by [56], or one of its variants [57].

An overview of the MC algorithm can be seen in the Fig. 14. First, the MC method generates N realizations (samples) of the random parameters according to their joint

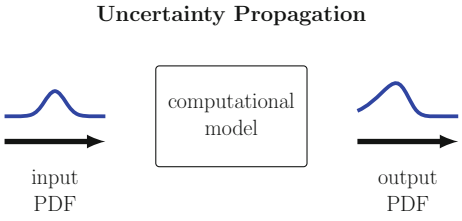


Fig. 13 Schematic representation of uncertainty propagation problem

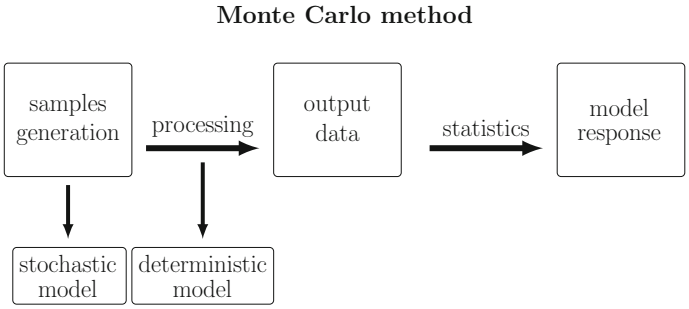


Fig. 14 An overview of monte carlo algorithm

distributions (stochastic model). Each of these realizations defines a deterministic problem which is then solved (processing) using a deterministic technique, generating a certain amount of data. Then, these data are combined through statistics, to access the response of the random system [55, 58]. By the nature of the algorithm, we note that MC is a non-intrusive method.

It can be shown that if N is large enough, the MC method describes very well the statistical behaviour of the random system. However, the rate of convergence of this non-intrusive method is very slow—proportional to the inverse of number of samples square root, i.e., $\sim 1/\sqrt{N}$. Therefore, if the processing time of a single sample is very large, this slow rate of convergence makes MC a very time-consuming method—unfeasible to perform simulation of complex models. Meanwhile, the MC algorithm can easily be parallelized, once each realization can be processed separately and then the results aggregated to compute the statistics [55].

Because of its simplicity and accuracy, MC is the best method to compute the propagation of uncertainties, whenever its use is feasible. Thus, it is recommended that anyone interested in UQ master this technique. Many good references about MC method are available in the literature. For further details, we recommend [58–64].

5.2 *Stochastic Galerkin Method: An Intrusive Approach*

When the use of MC method is unfeasible, the state of art strategy is based on the so-called stochastic Galerkin method. This spectral approach was originally proposed by [65, 66], and became very popular in the last 15 years, especially after work of [67]. It uses a *Polynomial Chaos Expansion* (PCE) to represent the stochastic model response combined with a Galerkin projection to transform the original stochastic equations into a system of deterministic equations. The resulting unknowns are the coefficients of the linear combination underlying to the PCE.

Once PCE theory is quite rich and extensive, we do not have space in this manuscript to cover it in enough detail, but to the reader interested in digging deeper on this subject is encouraged to see the references [2, 3, 8, 68–70].

6 Concluding Remarks

In this manuscript, we have argued about the importance of modeling and quantification of uncertainties in engineering projects, advocating in favor of the probabilistic approach as a tool to take into account the uncertainties. It is our thought that specifying an envelope of reliability for curves obtained from numerical simulations is an irreversible tendency. We also introduced the basic probabilistic vocabulary to prepare the reader for deeper literature on this subject, and discussed the key points of the stochastic modeling of physical systems, using a simplistic mechanical system as a more in-depth example.

Acknowledgements The author's research is supported by the Brazilian agencies CNPq (National Council for Scientific and Technological Development), CAPES (Coordination for the Improvement of Higher Education Personne) and FAPERJ (Research Support Foundation of the State of Rio de Janeiro).

References

1. L. Biegler, G. Biros, O. Ghattas, M. Heinkenschloss, D. Keye, B. Mallick, Y. Marzouk, L. Tenorio, B.B. Waanders, K. Willcox, *Large-Scale Inverse Problems and Quantification of Uncertainty* (Wiley, 2010)
2. O.P. Le Maître, O.M. Knio, *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics* (Springer, 2010)
3. D. Xiu, *Numerical Methods for Stochastic Computations: A Spectral Method Approach* (Princeton University Press, 2010)
4. C. Soize, *Stochastic Models of Uncertainties in Computational Mechanics* (American Society of Civil Engineers, 2012)
5. M. Grigoriu, *Stochastic Systems: Uncertainty Quantification and Propagation* (Springer, 2012)
6. R.C. Smith, *Uncertainty Quantification: Theory, Implementation, and Applications* (SIAM, 2013)
7. H. Bijl, D. Lucor, S. Mishra, C. Schwab, *Uncertainty Quantification in Computational Fluid Dynamics* (Springer, 2013)
8. M.P. Pettersson, G. Iaccarino, J. Nordström, *Polynomial Chaos Methods for Hyperbolic Partial Differential Equations: Numerical Techniques for Fluid Dynamics Problems in the Presence of Uncertainties* (Springer, 2015)
9. R. Ohayon, C. Soize, *Advanced Computational Vibroacoustics: Reduced-Order Models and Uncertainty Quantification* (Cambridge University Press, 2015)
10. T.J. Sullivan, *Introduction to Uncertainty Quantification* (Springer, 2015)
11. S. Sarkar, J.A.S. Witteveen, *Uncertainty Quantification in Computational Science* (World Scientific Publishing Company, 2016)
12. R. Ghanem, D. Higdon, H. Owhadi, *Handbook of Uncertainty Quantification* (Springer, 2017)
13. C. Soize, *Uncertainties and Stochastic Modeling* (Short Course at PUC-Rio, Aug 2008)

14. C. Soize, *Stochastic Models in Computational Mechanics* (Short Course at PUC-Rio, Aug 2010)
15. C. Soize, *Probabilité et Modélisation des Incertitudes: Eléments de base et concepts fondamentaux* (Course Notes, Université Paris-Est Marne-la-Vallée, Paris, 2013)
16. G. Iaccarino, A. Doostan, M.S. Eldred, O. Ghattas, Introduction to uncertainty quantification techniques, in *Minitutorial at SIAM CSE Conference*, 2009
17. G. Iaccarino, *Introduction to Uncertainty Quantification* (Lecture at KAUST, 2012)
18. A. Doostan, P. Constantine, *Numerical Methods for Uncertainty Propagation* (Short Course at USNCCM13, 2015)
19. C. Soize, A comprehensive overview of a non-parametric probabilistic approach of model uncertainties for predictive models in structural dynamics. J. Sound Vib. **288**, 623–652 (2005)
20. Guide for the verification and validation of computational fluid dynamics simulations. Technical Report AIAA G-077-1998 (American Institute of Aeronautics and Astronautics, Reston, 1998)
21. W.L. Oberkampf, T.G. Trucano, Verification and validation in computational fluid dynamics. Technical Report SAND 2002-0529 (Sandia National Laboratories, Livermore, 2002)
22. W. Oberkampf, T. Trucano, C. Hirsch, Verification, validation, and predictive capability in computational engineering and physics. Appl. Mech. Rev. **57**, 345–384 (2004)
23. ASME Guide for Verification and Validation in Computational Solid Mechanics. Technical Report ASME Standard V&V 10-2006 (American Society of Mechanical Engineers, New York, 2006)
24. W.L. Oberkampf, C.J. Roy, *Verification and Validation in Scientific Computing* (Cambridge University Press, 2010)
25. U.M. Ascher, C. Greif, *A First Course in Numerical Methods* (SIAM, 2011)
26. P.J. Roache, Code verification by the method of manufactured solutions. J. Fluids Eng. **124**, 4–10 (2001)
27. C.J. Roy, Review of code and solution verification procedures for computational simulation. J. Comput. Phys. **205**, 131–156 (2005)
28. L.A. Petri, P. Sartori, J.K. Rogenski, L.F. de Souza, Verification and validation of a direct numerical simulation code. Comput. Methods Appl. Mech. Eng. **291**, 266–279 (2015)
29. G.I. Schuëller, A state-of-the-art report on computational stochastic mechanics. Probabilistic Eng. Mech. **12**, 197–321 (1997)
30. G.I. Schuëller, Computational stochastic mechanics recent advances. Comput. Struct. **79**, 2225–2234 (2001)
31. C. Soize, Stochastic modeling of uncertainties in computational structural dynamics—recent theoretical advances. J. Sound Vib. **332**, 2379–2395 (2013)
32. D. Moens, D. Vandepitte, A survey of non-probabilistic uncertainty treatment in finite element analysis. Comput. Methods Appl. Mech. Eng. **194**, 1527–1555 (2005)
33. D. Moens, M. Hanss, Non-probabilistic finite element analysis for parametric uncertainty treatment in applied mechanics: recent advances. Finite Elem. Anal. Des. **47**, 4–16 (2011)
34. M. Beer, S. Ferson, V. Kreinovich, Imprecise probabilities in engineering analyses. Mech. Syst. Signal Process. **37**, 4–29 (2013)
35. C. Soize, A nonparametric model of random uncertainties for reduced matrix models in structural dynamics. Probabilistic Eng. Mech. **15**, 277–294 (2000)
36. C. Soize, Generalized probabilistic approach of uncertainties in computational dynamics using random matrices and polynomial chaos decompositions. Int. J. Numer. Methods Eng. **81**, 939–970 (2010)
37. A. Batou, C. Soize, M. Corus, Experimental identification of an uncertain computational dynamical model representing a family of structures. Comput. Struct. **89**, 1440–1448 (2011)
38. G. Grimmett, D. Welsh, *Probability: An Introduction*, 2nd edn. (Oxford University Press, 2014)
39. J. Jacod, P. Protter, *Probability Essentials*, 2nd edn. (Springer, 2004)
40. A. Klenke, *Probability Theory: A Comprehensive Course*, 2nd edn. (Springer, 2014)
41. A. Papoulis, S.U. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th edn. (McGraw-Hill, 2002)

42. C.E. Shannon, A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)
43. L. Wasserman, *All of Nonparametric Statistics* (Springer, 2007)
44. L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference* (Springer, 2004)
45. E.T. Jaynes, Information theory and statistical mechanics. Phys. Rev. Ser. **II**(106), 620–630 (1957)
46. E.T. Jaynes, Information theory and statistical mechanics II. Phys. Rev. Ser. **II**(108), 171–190 (1957)
47. J.N. Kapur, H.K. Kesavan, *Entropy Optimization Principles with Applications* (Academic Press, 1992)
48. J.N. Kapur, *Maximum Entropy Models in Science and Engineering* (New Age, 2009)
49. F.E. Udwadia, Response of uncertain dynamic systems. I. Appl. Math. Comput. **22**, 115–150 (1987)
50. F.E. Udwadia, Response of uncertain dynamic systems. II. Appl. Math. Comput. **22**, 151–187 (1987)
51. F.E. Udwadia, Some results on maximum entropy distributions for parameters known to lie in finite intervals. SIAM Rev. **31**, 103–109 (1989)
52. K. Sobezyk, J. Trçbicki, Maximum entropy principle in stochastic dynamics. Probabilistic Eng. Mech. **5**, 102–110 (1990)
53. K. Sobezyk, J. Trębicki, Maximum entropy principle and nonlinear stochastic oscillators. Phys. A: Stat. Mech. Appl. **193**, 448–468 (1993)
54. J. Trębicki, K. Sobezyk, Maximum entropy principle and non-stationary distributions of stochastic systems. Probabilistic Eng. Mech. **11**, 169–178 (1996)
55. A. Cunha Jr., R. Nasser, R. Sampaio, H. Lopes, K. Breitman, Uncertainty quantification through Monte Carlo method in a cloud computing setting. Comput. Phys. Commun. **185**, 1355–1363 (2014)
56. N. Metropolis, S. Ulam, The Monte Carlo method. J. Am. Stat. Assoc. **44**, 335–341 (1949)
57. C. Lemieux, *Monte Carlo and Quasi-Monte Carlo Sampling* (Springer, 2009)
58. D.P. Kroese, T. Taimre, Z.I. Botev, *Handbook of Monte Carlo Methods* (Wiley, 2011)
59. J.S. Liu, *Monte Carlo Strategies in Scientific Computing* (Springer, 2001)
60. G. Fishman, *Monte Carlo: Concepts, Algorithms, and Applications*, corrected edn. (Springer, 2003)
61. R.Y. Rubinstein, D.P. Kroese, *Simulation and the Monte Carlo Method*, 2nd edn. (Wiley, 2007)
62. S. Asmussen, P.W. Glynn, *Stochastic Simulation: Algorithms and Analysis* (Springer, 2007)
63. R.W. Shonkwiler, F. Mendivil, *Explorations in Monte Carlo Methods* (Springer, 2009)
64. C.P. Robert, G. Casella, *Monte Carlo Statistical Methods* (Springer, 2010)
65. R. Ghanem, P.D. Spanos, Polynomial chaos in stochastic finite elements. J. Appl. Mech. **57**, 197–202 (1990)
66. R. Ghanem, P.D. Spanos, *Stochastic Finite Elements: A Spectral Approach*, 2nd edn. (Dover Publications, 2003)
67. D. Xiu, G.E. Karniadakis, The Wiener-Askey Polynomial Chaos for stochastic differential equations. SIAM J. Sci. Comput. **24**, 619–644 (2002)
68. P. Vos, Time-dependent polynomial chaos. Master Thesis, Delft University of Technology, Delft, 2006
69. P. Constantine, in *A Primer on Stochastic Galerkin Methods*. Lecture Notes, 2007
70. A. O'Hagan, in *Polynomial Chaos: A Tutorial and Critique from a Statistician's Perspective*, (submitted to publication, 2013)

Towards a More Robust Understanding of the Uncertainty of Wind Farm Reliability

Carsten H. Westergaard, Shawn B. Martin, Jonathan R. White, Charles M. Carter and Benjamin Karlson

Abstract Understanding wind farm reliability from various data sources is highly complex because the boundary conditions for the data are often undocumented and impact the outcome of aggregation significantly. Sandia National Laboratories has been investigating the reliability of wind farms through the Continuous Reliability Enhancement Wind (CREW) project since 2007 through the use of Supervisory Control and Data Acquisition (SCADA) data from multiple wind farms in the fleet of the USA. However, data streaming from sample wind farms does not lead to better understanding as it is merely a generic status of those samples. Economic type benchmark studies are used in the industry, but these do not yield much technical understanding and give only a managerial cost perspective. Further, it is evident that there are many situations in which average benchmark data cannot be presented in a meaningful way due to discrete events, especially when the data is only based on smaller samples relative to the probability of the events and the sample size. The discrete events and insufficient descriptive tagging contribute significantly to the uncertainty of a fleet average and may even impair the way we communicate reliability. These aspects will be discussed. It is speculated that some aspects of reliability can be understood better through SCADA data-mining techniques and considering the real operating environment, as, it will be shown that there is no particular reason that two identical wind turbines in the same wind farm should have identical reliability performance. The operation and the actual environmental impact on the turbine level are major parameters in determining the remaining useful life. Methods to normalize historical data for future predictions need to be developed, both for discrete events and for general operational conditions.

Keywords Wind farms · Reliability · SCADA data · Wakes · Vibration

C.H. Westergaard (✉)
Department of Mechanical Engineering, Texas Tech University, Lubbock, TX, USA
e-mail: carsten.westergaard@ttu.edu

C.H. Westergaard · S.B. Martin · J.R. White · C.M. Carter · B. Karlson
Sandia National Laboratories, Albuquerque, NM, USA

S. Ekwaro-Osire et al. (eds.), *Probabilistic Prognostics and Health Management of Energy Systems*, DOI 10.1007/978-3-319-55852-3_9

1 Introduction

The wind industry has improved its operational practices tremendously over the past decades leading to tremendous global and local successes. According to the Global Wind Energy Council, a total of 370 GW was installed worldwide by the end of 2014, of which approximately 66 GW was installed in the USA. The fifth largest territory in the world is Texas, which according to recent numbers [1] gets 11.7% of its energy from a wind fleet of 16 GW worth of wind turbines. The wind energy power price in the interior of the USA is low and competitive only $23/MWh [2] making new installed wind equally attractive to new installed gas generation. The two sources, wind and gas, each have approximately half the market of new capacity installed in the past years.

A big contribution to the success of wind energy is the new technology introduced in many of the components of the wind turbines [3] which has driven down cost and increased efficiency. The huge rotor size is the most noticeable development; the size has doubled in the past 15 years, and so more than 80% of the rotors entering the market are over 100 m in diameter with an average nameplate rating of 2 MW [2]. This improvement results in capacity factors approaching 50%.

A second contributor to improved cost is both the increased reliability and the operational practices. In the early 80s, reliability was low and availability as low as 20% [4]. Costs were unreasonably high but today the O&M costs of new projects are of the order of $8/MWh [2]. The availability is beyond 98% [5]. This has been achieved by intensive monitoring from large operation centers, where owners of wind farms have reduced the downtime significantly by a fast response and secured availability for new spare equipment. The improved O&M cost also reflects the improvements in technology and design procedures.

In spite of these improvements, owners often comment that O&M expenditures either have increased or are too high. This can, in part, be understood from the large spread of O&M cost one can observe (see, for example [2]). Individual wind farm projects do not behave similarly, meaning that unplanned events are an important part of O&M cost, and, therefore, the reliability of the wind farm.

Three elements contribute to unplanned reliability events. First, new turbine technology is constantly entering the market, manufacturing flaws (acceptable and unacceptable) are present and new and/or unexpected failure modes occur. These types of failures typically show up in the initial life of the wind farm and contribute to the beginning of the classical bathtub curve. The failure modes are often covered under warranty and thus are of a proprietary nature, and therefore the cost and corrective action is not documented for future use. Frequently, the flaws are overestimated in future planning or studies for the same reason. The second contributor is poor operational practice and a lack of consistent documentation, including the tagging of reliability events. However, significant resources have now been put into improvements. The third contributor is truly unforeseen events or systematic bias and the frequency and nature of the events is typically not accounted for.

Removing reliability events of proprietary nature from statistics will remove any technical bias in our understanding. From a managerial perspective, such events should be allocated separately. Better tagging and descriptions of the events could lead to more opportunity for description of rare but costly unforeseen events to be investigated. Those are, for example, driven by external environmental conditions and operational practices. Finally, a data-driven description of each turbine's expected reliability performance could support a more accurate description of future expectations.

In this article, it will be discussed how important data quality and tagging is before aggregating. Further, it will be discussed how the boundary conditions (especially the external environmental impact) also need to be included with the data collection in order for the data comparison to be useful. Finally, it will be shown that SCADA data provides insight to some of these external impacts and how this knowledge can then support improvement of monitoring of wind farms.

2 The Challenge of Collecting Data

The largest single challenge in collecting data on reliability from wind farms is the proprietary nature of the data. The second largest challenge is that no one single participant has an overarching understanding of the entire fleet. Turbine suppliers know their own products and may be collecting data from those in the first few years of the wind farm life while the assets are still under warranty. However, the turbine suppliers do not typically have much knowledge about the remaining time of the turbine life. The owners often know little about the first years of operations (because the wind farm is under service agreement with the turbine supplier), but they do, on the other hand, have 20 years of operational experience with multiple turbine platforms, knowledge which is often partially shared with the independent service providers (see Fig. 1).

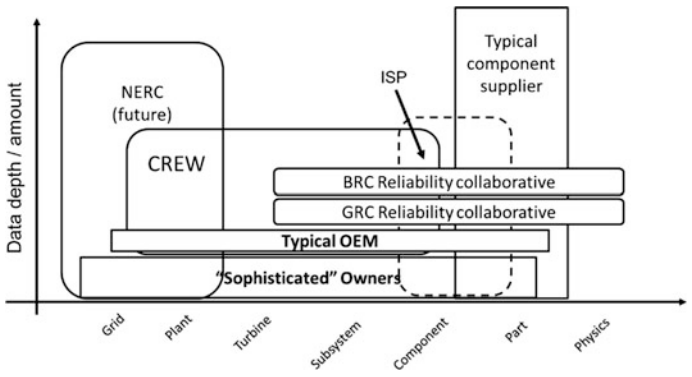


Fig. 1 Knowledge in the reliability space

However, the turbine supplier often has access to data that the owner does not, such as detailed design documentation and high speed turbine controls data (not to be confused with the SCADA data). The component suppliers know everything about their products and the physics of the product, but only a little about the physical conditions to which their product component is subjected. Efforts like the Gearbox Reliability Collaborative (run by National Renewable Energy Laboratory (NREL)), help close the gap for a narrow set of challenges such as initially described in [6]. The Blade Reliability collaborative (BRC) (run by Sandia National Laboratories) is a similar initiative. The efforts are focused on bringing the physics of materials, parts and physical conditions together with testing and monitoring (including condition monitoring or predictive health monitoring).

The Continuous Reliability Enhancement for Wind (CREW) efforts [5, 7–9], also run by Sandia National Laboratories, are aimed at generating macroscopic technical data to better understand the system as a whole, with a focus on developing methods which can be used to describe the external boundary conditions of the wind farm operation down to the turbine level. This is particularly challenging because of the large individual variability from turbine to turbine, and because the individual event data set may be extremely sparse. Further, the descriptive characteristics from event to event can be quite inconsistent. A pathway to a more robust approach is described in [9].

Other overarching data bases do exist (such as grid compliance by North American Reliability Council (NERC)) or are in preparation for other purposes,) such as the economic studies typically performed by private fee-based organizations. Again, such data is generally not sufficiently technically detailed to be useful understanding technical trends.

As discussed above, quantifying the reliability in general terms is difficult. Based on many different sources (such as [2, 5, 10–13] along with experience and conversations with different organizations), Table 1 reflects an indicative weighted approach indicating state-of-the-art for a fictive 2 MW geared turbine. While it is

Table 1 Indicative cost and occurrence of unplanned reliability for a fictive 2 MW turbine in the USA territory. The total lifetime cost is $516,000 whereof $330,000 is replacement cost. The cost accumulates to approximately $5/MWh produced energy. The unplanned cost is about half of the total cost. The objective of this table is not to give accurate numbers, but an order of magnitude

Item	Relative cost	Annual failure rate of repairable items	Fraction of fleet which will experience major replacement in lifetime
Blades	29%	16%	14%
Gear and bearings	36%	6%	42%
Generator	22%	3%	29%
Other	9%	39%	
Force outage or resets	4%		

common wisdom in the community that gear and bearings have caused challenges in the past, blades have been ignored at large. As seen, the cost of blade failures is similar to that of gear and bearings which is one reason it is expected to see more focus on blades in the coming years.

3 Unplanned Discrete Events

As discussed in the introduction, reliability events of a proprietary nature are, generally speaking, inaccurate indicators of future behavior either because they are associated with initial failures due to, for example, manufacturing flaws or because they are associated with some sort of upgrade by the manufacturer. It can be difficult to separate such events purely on a time basis. A recent presentation [11] showed blade failure sorted by blade types across an 8 GW fleet. One particular blade type showed a large number of events 3 years into the data (approximately four times larger than the average over 6 years) but with no failure at the end of the 6 year period. Presumably this was associated with events of proprietary nature. Further, one has to be aware that blade inspections are a manual and labor intensive process which may only be executed every 3 years. Therefore, the probability of detecting flaws can include significant delays from the origin of the damage.

Comparing blade failures (reported in [11] and [12]) from two different wind farm owners, it is apparent that the similar physical symptoms of blade failures are not tagged in a similar manner. This has, in part, to do with a lack of standards but also the practical difficulties actually tracking work orders and field reports. As an example, one of the reports calls out blade damages from lightning (which often results in damages near the tip of the blades and frequently near the trailing edge in the tip region). The second report, however, does not call out lightning and only reported trailing edge damages in general which could have many different root causes. Although this seems like a very simple case, the operational complexity of collecting such data from across many geographical locations and a diverse workforce, cannot be ignored. The uncertainty is real and it makes it difficult (if not impossible) to aggregate statistical data without unnecessary uncertainty.

Even if the above blade failure data could be accurately aggregated precisely for lightning damages, one has to be very careful in the interpretation of the data as common averages do not describe the issues at hand. Firstly, the number of thunderstorm days has large regional differences. For the USA, this ranges from almost no days in California and up to 75 days in the Mid-west. It would not be reasonable to assume a fleet average directly across these regions with such big differences in exposure to risk. Secondly, landscape exposure and turbine height plays an important role. The IEC standard [14] reports a height-square sensitivity, but in [10], a turbine manufacturer examines their historical fleet performance from very small turbines to large modern turbines disclosing a probability of strike sensitive to the height-in-the-fourth-power (with or without failures occurring). Now this level of sensitivity would mean that newer turbines should be extremely exposed since they

are much taller than just a decade ago. Although lightning damages in blades are significant (about 15% of all blade failures according to [11]), the technical improvement of lightning diverter systems has mitigated this large sensitivity.

The above examples highlight the difficulties in quantifying the reliability of wind turbine components because both the boundary conditions for the observations and the failure mechanisms are unknown or inaccurately reported. Furthermore, resorting to reducing the samples to very specific components may be challenging because the number of samples are usually low. If, for example, the annual blade failure rate is 16% (see Table 1), and this is distributed evenly on 20 major failure modes in a fleet of 1000 turbines, then only 8 event samples are retrieved per year. Out of these 8 samples, it will still be necessary to quantify the similarities in boundary conditions and failure modes, so aggregation becomes reasonable. If all data from the USA fleet were collected, we would have about 250 events per failure mode per year on record if all market players were contributing their data to the same standards.

It is clear from this discussion that meaningful aggregation of reliability data begins with a systematic inclusion of the boundary conditions of the failures:

- Extreme wind with static loading
- Extreme wind with dynamic events (such as vibration)
- Lightning characteristics
- Environmentally induced erosion, corrosion or similar deterioration
- Ice and extreme cold
- Operation or maintenance variance

4 SCADA Data Mining and Modeling Potential

Modern wind farms are instrumented with a large number of sensors (as many as 250 sensors per turbine) which can be accessed in different ways. In general, SCADA data is available at least as 10-min averages and can provide an overview of the historical reliability events such as shown in [5, 13]. This approach may reveal reliability issues like sensor faults, but even if these are frequent, the associated costs may not be of big consequence. In the previous section, it was discussed how discrete events need better clarification on the boundary conditions, but what about the average operational parameters which induce the wear and tear? Are they similar between even two neighboring wind turbines in a wind farm? The answer is no, but we can use SCADA data to understand turbine to turbine variations and possibly develop models from such mappings.

Recently, Martin et al. [7, 8] investigated 1.5 years of SCADA data from 67 wind turbines in the mid-west of the USA in order to quantify the impact of turbines shadowing each other with their wakes by mapping the normalized performance in narrow directional sectors. The study found that turbines waking each other indeed impose power deficits and increased power variability when a turbine is directly

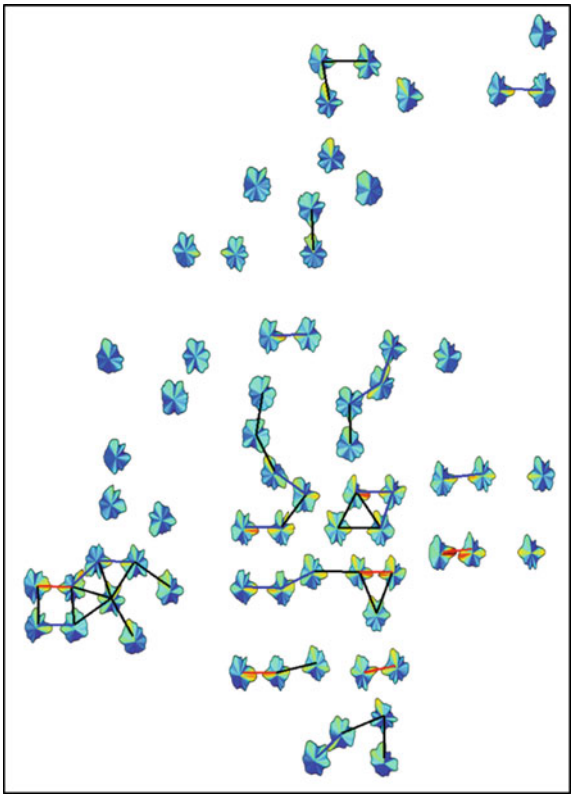


Fig. 2 Directional power variability in an entire wind farm mapped over 1.5 year [7, 8]. The colored lines show turbines which are closely spaced

waked by another turbine. This scenario is generally covered by the design foundation in wind turbine design, where an increased turbulence level will be used for computing turbine loads. However, the study also found that certain combinations of upstream turbine locations actually caused the downstream turbine to produce much higher power than its peers. Counter-intuitively, these high power situations were associated with low power variability, and these effects could affect turbine reliability both in a positive and a negative way. Figure 2 shows the power variability across the entire wind farm, and it is clear that none of the turbines had the same experience in the 1.5 years of investigation, so it should not be expected that the drive train in each of these machines would exhibit the same lifetime wear. Furthermore, this illustrates that one cannot consider a simple fleet average to compare drivetrain reliability performance.

In Fig. 3, three turbines have been selected from the upper middle part of the wind farm, and the power variability shows the clear characteristics just discussed. In the right hand part of the figure, the average tower vibration over 1.5 years is plotted, and a similar pattern is seen across the wind farm. The vibrations show similar trends, (compared to the power variance), in particular in the wake situation. In addition to individual wake deficit profiles, a generic wake deficit effect can be

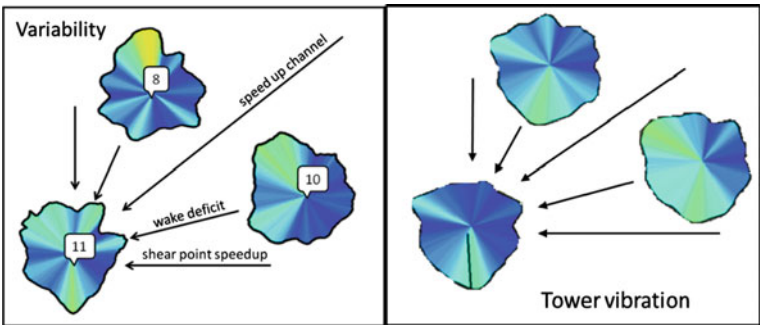


Fig. 3 *Left* Power variability from 3 turbines in the *upper middle* in Fig. 1. *Right* Corresponding average tower vibration levels recorded over 1.5 years, similar to that of the power variability (*Arrows* indicate wind farm flow effects identified in [7, 8])

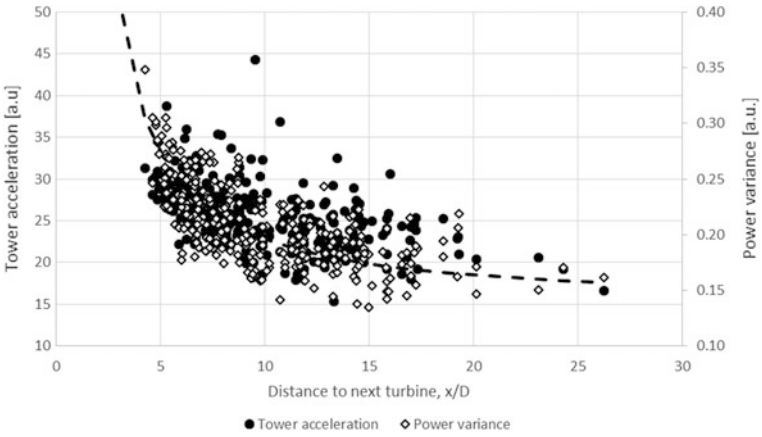


Fig. 4 Power variability in the direction of any neighboring turbine (waked turbine) from Fig. 3 compared to tower vibration in the same direction. Data is 1.5 years of duration

observed across the entire wind farm. To see this effect, 854 turbine pairs were selected within 25 rotor diameters and an undistributed direct path was chosen between them (to observe potential wake effects). From these pairs, the maximum power variance and the maximum tower vibration level of the downwind turbine versus distance between the turbine pairs is shown in Fig. 4.

In [8], a simple directional analysis model is demonstrated for power and power variability to effectively map the inter-turbine variability based only on the geometrical layout of the wind farm. A similar model could potentially be built for a tower, the blades and other main components in order to reduce the uncertainty imposed on the analysis of reliability data.

As a last step in this investigation showing large inter-turbine variability, the number of faults reported in the SCADA system were plotted as a function of

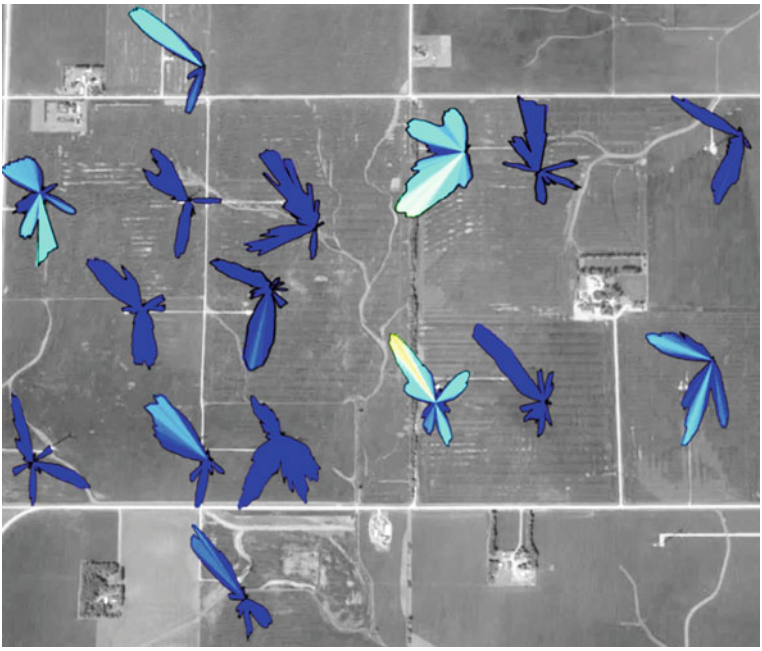


Fig. 5 All fault counts by direction over 1.5 years in subset of the wind farm shown in Fig. 1. It is noticeable that the faults do generally not align with the main wind directions (NE and SSE)

direction. In Fig. 5, a subset of the wind farm is shown. As a first observation, it is clear that none of the turbines exhibit similar fault behavior. It is also surprising that the majority of faults are not aligned with the two main wind directions (NE and SSE), but seem to be rather randomly oriented. A deeper analysis could potentially help understand these patterns, but very little validation opportunity exists for this particular data set so this has not been pursued further.

The directional analysis confirms that, even for a simple flat site in the Mid-west, higher fidelity analysis provides great insight to the reliability performance of each turbine and that bulk averaging may not be a suitable approach. This type of analysis would be useful in complex landscapes where the turbine performance is heavily influenced by the landscape features. Methods to normalize historical data for future predictions of reliability are definitely possible.

5 Conclusion

Discrete events of a proprietary nature need to be isolated from technical benchmarking as they do not support the prediction of the future. Further, it is clear that environmentally-induced reliability issues (originating, for example, from wind,

ice, moisture, lightning, erosion, and corrosion) are relatively undocumented, in part due to the lack of attention and the inspection methods and the common tagging methods. The discrete events are relatively rare so large amounts of data is required, which suggest that a national effort is required if meaningful technical information is to be retrieved for future modeling. Methods for normalization with respect to physical processes (size, technology, location, environment etc.) need to be included in such efforts. The CREW project was initiated to facilitate this national effort to collect, normalize, analyze, and benchmark this type of data essential for understanding wind turbine fleet reliability trends and issues. However, the success of the CREW project will be determined by the willingness of owners to participate and share data with Sandia National Laboratories under the protection of a non-disclosure agreement that ensures the safeguarding of all proprietary data.

A novel directional analysis has been developed for power and power variance showing that individual turbines performance is linked to their location. It is shown that a similar analysis of sensors relating to loading on main components could be successful in modeling the common wear and tear on the individual turbines rather than using common average approaches.

Finally, a directional analysis of faults occurring in the wind farm may prove extremely useful and reveal these individualities from turbine to turbine. Deviation from the expected patterns could yield more accurate detection and accommodation.

Acknowledgements This work is supported and made possible by the Department of Energy (DOE) Wind and Water Power Program. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. Wind farm SCADA data was provided by a strategic industrial partner.

References

1. R. Fares, Texas poised to integrate more wind, solar energy, blogs.scientificamerican.com (2016)
2. R. Wiser, M. Bolinger, 2014 Wind technologies market report, Department of Energy (2015)
3. P. Jamison, *Innovation in Wind Turbine Design* (Wiley, West Sussex, UK, 2011)
4. J.F. Manwell, J.G. McGowan, A.L. Rogers, *Wind Energy Explained: Theory, Design and Application* (Wiley, 2009)
5. V.A. Peters, A.B. Ogilvie, C.R. Bond, *Continuous Reliability Enhancement for Wind (CREW) Database: Wind Plant Reliability Benchmark*
6. W. Musial, S. Butterfield, B. McNiff, *Improving wind Turbine Gearbox Reliability* (National Renewable Energy Laboratory, 2007)
7. S.B. Martin, C.H. Westergaard, J.R. White, New wake effects identified using SCADA data analysis and visualization, in *Proceedings of AWEA Wind Power Conference*, Florida (2015)
8. S.B. Martin, C.H. Westergaard, J.R. White, B. Karlson, Visualizing wind farm wakes using SCADA data. Sandia report (2016)

9. C. Carter, B. Karlson, S. Martin, C.H. Westergaard, Continuous reliability enhancement for wind (CREW), Program Update. Sandia report (2016)
10. Cannata, Lightning protection system (LPS), *Presentation at Windpower Monthly Seminar: Blade Inspection Damage and Repair Forum*, Hamburg (2014)
11. D. Coffey, Blade Reliability case study, in *Sandia Wind Turbine Blade Workshop*, Albuquerque, New Mexico (2014)
12. M. Nissim, Blade maintenance for reliability, an owner/operator perspective, in *Sandia Wind Plant Reliability Workshop*, Albuquerque, New Mexico (2013)
13. M. Wilkinson, Measuring wind turbine reliability—results of the reliawind project, in *Proceedings of EWEA*, Brussels (2011)
14. International Electrotechnical Commission, *Wind Turbine Generator Systems—Part 24: Lightning Protection*, IEC/TR 61400-24:2002(E)

Data Analysis in Python: Anonymized Features and Imbalanced Data Target

Emanuel Rocha Woiski

Abstract Remaining useful life (RUL) of an equipment or system is a prognostic value that depends on data gathered from multiple and diverse sources. Moreover, assumed for the sake of the present study as a binary classification problem, the probability of failure of any system is usually very much smaller than that of the same system to be in normal operating conditions. The imbalanced outcome (largely much more 'normal' than 'failure' states) at any time results from the combined values of a large set of features, some related to one another, some redundant, and most quite noisy. Previewing the development and requirements of a robust framework, it is advocated that by using Python libraries, those difficulties can be dealt with. In the present Chapter, DOROTHEA, a dataset from UCI library with a hundred thousand of sparse anonymized (i.e. unrecognizable labels) binary features and imbalanced binary classes are analyzed. For that, an ipython (jupyter) notebook, pandas are used to import the data set, then some exploratory analysis and feature engineering are performed and several estimators (classifiers) obtained from scikit-learn library are applied. It is demonstrated that global accuracy does not work for this case, since the minority class is easily overlooked by the algorithms. Therefore, receiver operating characteristics (ROC), Precision-Recall curves and respective area under curve (AUCs) evaluated from each estimator or ensemble, as well as some simple statistics, using three hybrid methods, that are, a mix of filter, embedded and wrapper methods, feature selection strategies, were compared.

Keywords Data analysis · Machine learning · Scikit-learn · Python · Imbalanced classes · ROC · Precision-recall

E.R. Woiski (✉)
Department of Mechanical Engineering, São Paulo State University (UNESP), Ilha Solteira, SP, Brazil
e-mail: woiski@dem.feis.unesp.br

S. Ekwaro-Osire et al. (eds.), *Probabilistic Prognostics and Health Management of Energy Systems*, DOI 10.1007/978-3-319-55852-3_10

1 Introduction

Remaining useful life (RUL) of a system is a prediction into the future. The proper estimation of RUL depends on the quality of the data of the current state of the system and on the capability of obtaining the relevant features from this data. On the other hand, actual raw data, even carefully collected from the field, is almost never ready for immediate analysis and interpretation. This is not only because of missing and/or erroneous values, but also because of the unsuitable features exposed [1]. In order to be useful, data needs to be processed, altered and properly conformed to the answers we seek and then, decisions have to be made about missing values. Some guidelines for this phase, that should never be overlooked, are described briefly here.

Regarding the models themselves, Machine Learning (ML) has been used to train the models from the *data features*, that is, common characteristics for each instance (e.g. result or outcome), and then used to infer conclusions from (mostly) unseen new data. Taking into account the overwhelming data production in every field, more and more analysts and researchers are using semi-automated inference from the data, that is, ML. As for the selection of models, there are plenty of ML techniques available, that are adopted to the data and the types of questions to be answered, as will be shown in the following sections.

Python is a well-known interpreted dynamic open-source multi-platform programming language and because of the easiness to learn and the number of libraries, is increasingly becoming popular among scientists and engineers [2]. In fact, with the right libraries, one can do almost anything without ever leaving the language domain, a great advantage for non-programmer scientists and engineers. However, because of the need for C/C++/Fortran compiled extensions into modules, Python libraries installation used to represent a big hurdle, but that is not a problem any longer. For example, Anaconda [3] or Enthought [4] furnish freely all what is needed to have a full scientific Python distribution, more than 200 packages, installed in few minutes, regardless of platform.

Therefore, *numpy* [5], *scipy* [6], *matplotlib* [7], *Ipython (Jupyter) notebook* [8], *pandas* [9], *statsmodels* [10] and *scikit-learn* [11] are becoming household names in Science and Engineering in general and among data scientists and engineers in particular. In the present book chapter, most of those libraries are deployed in the prediction problem for imbalanced binary classes, given a hundred thousand anonymized sparse binary features, the DOROTHEA–UCI dataset [12]. The rest of this chapter is organized as follows: in the Sect. 2 some guidelines for data analysis are summarized. In Sect. 3, some introductory material on ML is presented. In Sect. 4, the scikit-learn Python library is introduced, stressing its API (Application Programming Interface) consistency, which facilitates everything even for the non-specialized user. The problem stated by the authors of DOROTHEA [12] dataset and its specifications are established in Sect. 5. In Sect. 6, there is a brief description of the procedure to load and transform DOROTHEA [12] dataset. The basic definitions of feature processing are in Sect. 7 and the search procedure for

duplicated features in Sect. 8. In Sect. 9, the largest, scoring parameters adequate for imbalanced data are defined, as well as the selected estimators are briefly described and the results of their application for DOROTHEA [12] are presented and discussed. Finally, some conclusions on the feasibility of the use of the Python libraries to DOROTHEA [12] problem, a highly imbalanced classification problem with several thousand of anonymized features close the chapter in Sect. 10.

2 Guidelines for Data Analysis

Applying the algorithms on data is probably one of the final task that an analyst will perform. Before considering that, a number of tasks need to be fulfilled. For the sake of clarity, a recommended checklist is discussed in the sub-sections that follow [1, 13].

2.1 Answering the Question

- Specify the type of analytical question to be formulated (e.g. exploration, association, causality) before analyzing the data;
- Define the metric for the success of answering the question correctly;
- Understand the context for the question and the scientific or business application;
- Record the experimental data; and
- Consider if the question could be answered with the available data.

2.2 Checking the Data

- Plot univariate and multivariate summaries of the data; and
- Check for discrepancies in the data.

2.3 Tidying the Data

- Each variable (feature) should be one column and each observation (instance) one row;
- Record the procedure for moving from raw to tidy data; and
- Record all parameters, units, and functions applied to the data.

2.4 Exploratory Analysis

- Identify missing values;
- Make univariate plots (histograms, density plots, boxplots);
- Consider correlations between variables (scatterplots);
- Check the units of all data points to make sure they are in the right range;
- Identify any errors or miscoding of variables; and
- Consider plotting on a log scale.

2.5 Inference

- Identify what large population you are trying to describe;
- Clearly identify the quantities of interest in your model;
- Consider potential confounders;
- Identify and model potential sources of correlation such as measurements over time or space; and
- Calculate a measure of uncertainty for each estimate on the scientific scale.

2.6 Prediction

- Identify in advance your error measure;
- Split your data into training and validation (testing);
- Use cross validation, resampling, or bootstrapping only on the training data;
- Create features using only the training data;
- Estimate parameters only on the training data;
- Fix all features, parameters, and models before applying to the validation data; and
- Apply only one final model to the validation (testing) data and report the error rate.

3 Fundamentals of Machine Learning

A frequently quoted definition of Machine Learning (ML) from [14] says: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with the experience E". ML systems semi-automatically learn programs from data [15] using a number of models in the form of proper algorithms.

The learning models are categorized in two groups, namely, *supervised* and *unsupervised* models. Supervised learning models make predictions using previously labeled data, whereas unsupervised learning models extract some structure from unlabeled data. Depending on the type of outcome (result), ML can be categorized as *classification*, in the case of unordered categorical classes, or *regression*, in the case of continuous values for the outcomes. There is a third category, *reinforcement learning*, a special case of classification, but with emphasis on the rewards, in which the model has to respond to changes in the environment [16].

There are so many supervised and unsupervised models to choose from, besides ensembles, i.e., several much weaker individual models working cooperatively to build a stronger one. For a given application, the best model depends upon a number of factors, such as the category, the questions to be answered or the inferences to be made, the selected metrics for success, and the kind of data to be dealt with. Moreover, almost every model can be tuned, by altering the values of its many *hyperparameters*, much like turning radio knobs.

Whereas there are numerous algorithms available, ML can be described as constituting of three components [17]:

Representation: The learning algorithms must be represented in some formal language that the computer can handle and conversely, choosing a representation for a learner is choosing the set of structures that it can possibly learn.
Evaluation: An evaluation function (also called objective or scoring function) is needed to distinguish good algorithms from bad ones in some pre-defined sense.
Optimization: A method is needed to search among the algorithms for the highest-scoring one.

4 The Scikit-Learn Package

Open-source BSD licensed scikit-learn package started as part of the Scikits (SciPy Toolkits), and is now, alongside pandas, the core of data science operations on Python. In scikit-learn package there is everything necessary for data preprocessing, supervised and unsupervised learning, model selection, validation, and error metrics [18].

One of the great advantages of scikit-learn for scientists and engineers is the API consistency across all included algorithms, selecting very sensible defaults for the hyperparameters [19]. Once clearly establishing the problem, that API consistency allows one to try several models with very little change, in such a way that experimentation on ML becomes truly accessible to anyone without regard to mathematics skills or programming skills. *Numpy*, *Pandas* and *sckit-learn* libraries were employed in the problem analysis presented in this book chapter. A description of the problem is object of the next section.

5 The Problem

Consider the DOROTHEA [12] dataset applied to UCI ML Repositories [20]. DOROTHEA [12] is just one of the five experiments designed for the NIPS 2003 variable selection benchmark. The dataset with which DOROTHEA [12] was produced is one of the knowledge discovery in data mining (KDD) Cup 2001. DuPont Pharmaceuticals graciously furnished the original data set for the KDD Cup 2001 competition.

The task of KDD dataset was to predict which compounds bind to *thrombin*, a substance of blood clot. This is a two-class classification, that means each compound (instance) results classified in "active" (positive) or "non active" (negative), depending on a very sparse binary vector of features (variables), defining a set of characteristics for each compound. Among many compounds, few truly bind to thrombin in such a way that positive class outcomes are very rare. That is a case of imbalanced classes (outcomes). The data was split into training, validation, and test sets while maintaining the same proportion of samples (examples, instances) of the positive and negative class in each set. The final classes and samples are distributed according to Table 1. Notice that there are less than 10% positive ("active") instances from a total of 1,950 compounds, in such a way that more than 90% are negative ("inactive").

In order to build DOROTHEA [12], only the top ranking 100,000 original features were kept. For the second half lowest ranked features, the order of the instances was individually randomly exchanged (in order to create what the DOROTHEA [12] authors called "random probes"). The order of the instances and the order of the features were globally randomly permuted to mix the original training and the test instances and remove any feature order. All features are binary and anonymized, that is, no feature identification is available, and there are no missing values. The feature set is very sparse, since less than 1% of the entries are nonzero (1,776,363 nonzero in 1.95×10^8 entries). The produced data set was saved as a sparse-binary 1,950 instances × 100,000 features matrix, row-wise starting from 1. In DOROTHEA [12] feature files, e*ach entry* in a row is the *nonzero* column position, also starting from 1. Table 2 shows some statistics on the distribution of nonzero feature values in that matrix. Considering all rows and respective labels, instance classification is inferred from at most 11,475 nonzero in 100,000 entries.

Table 1 Distribution of positive and negative classes among the samples

Label	Positive	Negative	Total
Training	78	722	800
Validation	34	316	350
Test	78	722	800
Total	190	1,760	1,950

Table 2 Statistics of non-zero feature values between classes

Label	Min	Max	Median
Positive	687	11,475	846
Negative	653	3,185	783
All	653	11,475	787

The DOROTHEA [12] dataset is distributed in training, validation and test data files, for the feature arrays, and training and validation label files, for the binary class, originally (−1, +1) distribution.

6 Loading the Data

First of all, the files containing training, validation and test data as well as respective label file (for training and validation), except test label files, not in DOROTHEA [12] file set, were loaded using *pandas* [9]. *Pandas* is a Python library able to deal with all sorts of malformed data source, even with missing values, either in text or binary formats. Using *numpy masks* and taking '*int8*' *(byte)* as *dtype,* the feature array for each set was constructed. Column positions were decreased by 1, because of the zero counting. Training, validation and test data were row-wise concatenated, so that the whole dense 1,950 × 100,000 feature byte array could be regenerated in memory. All −1 labels (negative class) were converted to 0. Each row position values vector was simply treated as an index array in such a fashion that a *pandas* data frame was produced.

7 Performing Feature Engineering

The dimension space is defined by the size of the feature vectors. As the space dimension grows, the instances, limited in number, become further and further apart, under any concept of distance, truly an almost empty space, a phenomenon called "curse of dimensionality" [21, 22]. In the presence of hundreds or thousands of features, a large number of them are not informative because they are either irrelevant or redundant to help predicting a class. When the number of features is high but the number of instances is small, ML gets a difficult task, since the search space will be sparsely populated and the model will not be able to correctly separate the relevant data from the noise. Therefore, the feature number had to be considerably reduced. There are two choices: *feature extraction* and *feature selection*. By using feature selection, several features are assembled together in order to obtain entirely new features. The major difficulty in the interpretation of the relationship between features and outcomes (results, targets) is the main shortcoming of that approach, since the original features are hidden from view. On the other hand,

feature selection has been an important activity in data pre-processing and has been widely studied in the past few years. Feature selection does not change the features in any way, just pruning them (reducing their number down) using some proper algorithm. There are four choices here: *filters*, *wrappers, embedded* and *hybrid* approaches. *Filters* use the statistics within the proper data in order to remove the irrelevant features, without relying in any estimator. *Wrappers* are built into the models, taking the role of sorting all the features according to their relevance along their own estimation procedures. *Embedded* methods use specially developed models for the selection of the features. There are many forms to build a *hybrid* approach. Whatever the method, features can be either individually ranked or a subset is selected, always according to some criteria [21, 23–25]. In the present book chapter, several *hybrid* approaches, using filters as well as wrapper methods have been designed and are described further. Nevertheless, *duplicated* features have to be removed because they are useless for estimation since they waste time and processor cycles. In a hundred thousand features, any procedure bringing the number down without harming their information content is worth.

8 Searching for Duplicated Features

In the search for duplicated features in the whole data, a very fast and efficient algorithm (see the code in the Appendix), using Python *ordered dictionaries, sparse array and string conversions.* The purpose of this algorithm was to reduce the range of comparisons of row and column values in order to avoid useless computational expenses to compare redundant data. The transpose of the feature array was converted row wise in a sparse array and each row was turned into a string and afterwards, stored into a dictionary key. Every time the same key was met by the searcher, a duplicated feature vector was detected thus its index was recorded to be removed afterwards. That way, the size of the remaining feature set was reduced from 100,000 to 83,218.

Suppose there were *duplicated* instances but with flipped classes. That was noise that had to be treated somehow, since would confuse any training procedure. Therefore, after removing the duplicated features, the same algorithm was used to search for duplicate instances. There was none, but at least on can remain assured that there were no duplicated instances in the DOROTHEA [12] data set.

9 The Score Parameters and the Estimators

In supervised classification problems, the model learns from the features about the probability of any instance belongs to each class, given a *threshold value*. This threshold is the probability value chosen to make a decision about the transition of the categorical classification of any sample target from one class to the other.

Table 3 Samples of confusion matrices for a selected threshold value for DOROTHEA [12]

	Training				Validation		
	PREDICTED				PREDICTED		
TRUE VALUES	**TN**	**FP**	**TOTAL**		**TN**	**FP**	**TOTAL**
	722	0	722	**TN + FP**	308	8	316
	0	78	78	**FN + TP**	10	24	34
	FN	**TP**	**808**		**FN**	**TP**	**350**

Imbalanced data still presents a challenge for estimators [26, 27]. Information on binary classification for imbalanced classes is mainly conveyed by the *confusion matrix*. This matrix, as shown in Table 3, is built out from the comparisson between the target values predicted by some algorithm for a selected threshold value. Comparing the estimates from the model with the real targets (the ground truth), the confusion matrix is built. Table 3 is one of the several representations of the confusion matrix in the literature [11]. Along the main diagonal is the TN, classified as *true negative* instances, corresponding to the *majority* class, and the TP, classified as *true positive* instances, corresponding to the *minority (rare)* class. Along the opposite diagonal, from the top there is the FP, *false positive*, instances misclassified as positive but that are actually negative, and the FN, *false negative*, the negative ones that are in fact positive [28–30]. Check that TP + FN is P the total of 34 positive instances in the validation data set, TN + FP is N, the total of 316 negative instances, while TP + FP is the sensitivity to discriminate the true positive classes.

With TP, TN, FP, FN a number of scoring parameters to evaluate and compare the quality of the estimates can be defined. The most common scoring parameters, for a chosen threshold value, are defined in the Eqs. 1–6 [29, 31–33]. Note that the authors in [32] wrongly exchange TN and FN definitions, although the parameter formulations are correct.

- *True Positive Rate or Recall (Sensitivity):*

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

- *False Positive Rate:*

$$FPR = \frac{FP}{FP + TN}, \tag{2}$$

- *Precision:*

$$Precision = \frac{TP}{TP + FP}, \tag{3}$$

- *F-Measure:*

$$F1 = 2\frac{(Precision \times Recall)}{Precision + Recall}, \tag{4}$$

- *Matthew's correlation (phi) coefficient:*

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{5}$$

- *Accuracy:*

$$ACC = \frac{TP + TN}{P + N}. \tag{6}$$

Note that ACC, which is commonly used for supervised classification problems, is not a reliable scoring parameter in the case of highly imbalanced classes, since the much larger TN overwhelms the evaluation [26]. In the present book chapter, receiver operating characteristics (ROC) and Precision-Recall curves were selected as scoring parameters to evaluate the quality of the fit for each selected estimator. The ROC curve corresponds to the representation of the TPR × FPR pairs for the *whole range* of threshold values [0, 1]. One way to summarize the quality of the fit is the area under the curve (AUC) for the ROC curve. The same reasoning is made for the Precision-Recall Curve and its respective AUC, also evaluated using the full range of threshold values [31]. Although AUC ROC has been put in question as performance measure, as in [34], it still can be used, together with ROC curve, to compare estimators [35].

As for the estimators, in order to control over fitting and bias, ensemble classifiers were selected, with the sole exception of the multi-layer perceptron (MLP) classifier. In *scikit-learn*, that class implements a MLP algorithm that trains by adjusting the values of a set of weights along layers of neurons using back-propagation, when submitted to the training data and corresponding labels. In ensemble algorithms, *averaging methods* represent a class of algorithms in which several instances of a black-box estimator are applied on random subsets of the original training set and then their individual predictions are aggregated to produce a final prediction. For the present book chapter, *RandomForest, ExtraTrees* and *Bagging* classifiers were selected. *RandomForest* and *ExtraTrees* are ensembles of binary decision trees, each one built out of specially defined randomized selection of features and instances. *Bagging* classifier is a meta-estimator, that implements a base estimator several times over randomized inputs [11, 36, 37].

On the other hand, in *boosting methods*, several base weak estimators are deployed sequentially in order to reduce the bias of the combined estimator, usually resulting in a much stronger and powerful ensemble. For the present book chapter, *AdaBoost*, *GradientTreeBoosting* and *Xgboost* classifiers were selected. *AdaBoost* implements a sequential application of weak learners, changing weights at each

iteration towards improving the wrongly predicted results. *GradientTreeBoosting* deploys a sequence of weak learners, usually shallow trees, reforcing each other outcomes [11, 18].

Some of the classifiers, among them the *RandomForest* and the *ExtraTrees,* produce a list of feature importance, evaluated according to the role of each feature in the development of the trees. So, in order to reduce further the number of features, three strategies were employed. In the first strategy, the robust *RandomForest* classifier was applied on the training data, 1, 2, 3, 4 and 5 times sequentially, each iteration removing all *zero importance* features and a last one developed to stop when reaching a change less than 5% of the current number of features. That strategy reduced the count to 6,204, 3,395, 2,778, 1,759 and 1,220 features, respectively. In the second strategy, the meta-model *SelectfromModel* was employed to the *ExtraTrees* and then sequentially to the *RandomForest* classifier to remove features with importance equal or less than the mean value. That way, the count was at once reduced to 1,292 features. In the third strategy, a filter was designed to remove irrelevant features relying on the chi^2 and p-values univariate significance test between features and targets [11, 38]. That reduced the number from 83,218 to 10,517. With two pass over *RandomForest* removing all features with importance less than the mean value, this was further reduced to 491 features, a worrying cardinality because of the possibility of increased bias. However, in the tuning, hundreds more estimators (like trees grown) could be employed to compensate. Notice that 491 features correspond to less than 0.5% of the original 100,000 features. Are those the most important or relevant features of the set? This is further discussed alongside the results.

It has to be stressed that most estimator *hyperparameters* were tuned by hand making sure that both classes were always represented in any sampling process. Also, the validation data set was never used to fit the estimators.

With each reduced feature set, ROC and Precision-Recall curves were drawn for validation data, after applying the training data on each model, namely: *mlp, ada, bag, ext, ran, gra and xgb*. All resulting AUCs can be seen in Tables 4 (ROC) and 5 (Precision-Recall), and the curves in Figs. 1 (ROC) and 2 (Precision-Recall) only for the second strategy. Compare with Figs. 3 (ROC) and 4 (Precision-Recall) for the third strategy.

Using the first strategy, the decreasing number of features from 6,204 to 1,220 seemed to produce no great effect on both AUCs. The second strategy brought the number down to 1,292 and that did not make any difference either. So the third strategy was designed to use an univariate filter (chi^2) followed by a wrapper (*RandomForest*) to see what would happen with the AUC scoring under a very reduced feature set. It is interesting to conclude that the third strategy and its 491 features can cope quite well to the estimation procedure. All the results can be analyzed in Tables 4 and 5.

Table 4 AUCs for the ROC curves

Estimator	Number of features						
	1,220[a]	1,759[a]	2,778[a]	3,395[a]	6,204[a]	1,292[b]	491[c]
mlp	0.94	0.90	0.89	0.90	0.91	0.93	0.93
ada	0.95	0.91	0.94	0.91	0.93	0.91	0.93
bag	0.90	0.89	0.92	0.91	0.89	0.92	0.91
ext	0.92	0.91	0.93	0.93	0.94	0.92	0.91
ran	0.92	0.91	0.94	0.91	0.94	0.92	0.93
gra	0.89	0.87	0.91	0.90	0.89	0.89	0.92
xgb	0.88	0.88	0.88	0.88	0.88	0.90	0.88

[a]According to the first strategy
[b]According to the second strategy
[c]According to the third strategy

Table 5 AUCs for the Precision-Recall curves

Estimator	Number of features						
	1,220[a]	1,759[a]	2,778[a]	3,395[a]	6,204[a]	1,292[b]	491[c]
mlp	0.73	0.70	0.70	0.70	0.69	0.75	0.75
ada	0.75	0.74	0.73	0.74	0.75	0.69	0.76
bag	0.68	0.70	0.71	0.71	0.72	0.70	0.71
ext	0.75	0.73	0.76	0.75	0.77	0.72	0.76
ran	0.73	0.74	0.75	0.75	0.76	0.75	0.77
gra	0.69	0.65	0.69	0.67	0.66	0.67	0.73
xgb	0.69	0.69	0.69	0.69	0.69	0.71	0.71

[a]According to the first strategy
[b]According to the second strategy
[c]According to the third strategy

Analyzing the ROC curves and AUC for the positive class with the validation data for each estimator in Figs. 1 and 3 and one can notice the compromise between TPR and FPR, that is, one can cope with some FP, as long as the TP is high enough. The dotted line represents the random choice, for which the AUC is 0.5. Again, comparing Figs. 1 and 3, there is not much difference between the second and third strategies. Pay attention to the steepness at the left side and also to the nonlinear staircase effect. The ideal ROC curve is the one following the *upper left* corner, resulting unitary AUC, in such a way that TPR = 1.0 would be associated to FPR = 0.0. Therefore, the steepness is very important since for TPR (Recall) up to about 0.70 or TP = 24 instances (over 34), FPR can be kept small, of order of about 0.05, or FP = 16 instances (over 316). On the other hand, the staircase means that TPR does not change for large FPR as well as threshold ranges. However, going a little more to the right and the ROC curve is almost useless, as FPR becomes very large for imbalanced data.

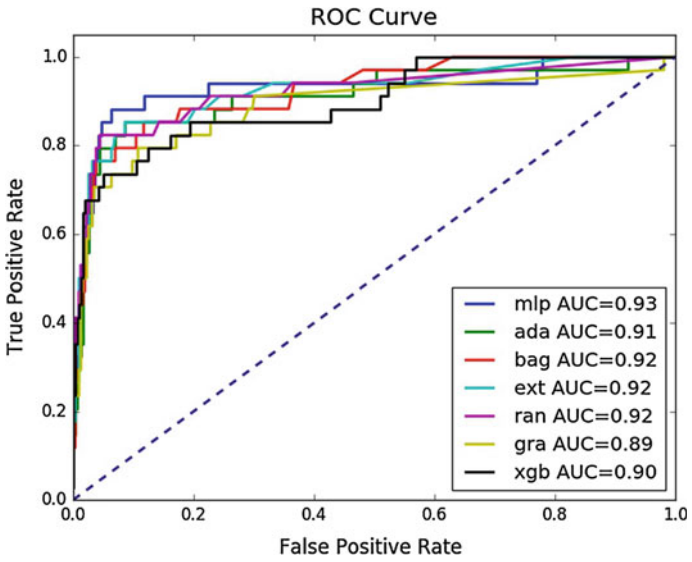


Fig. 1 ROC curves for the second strategy

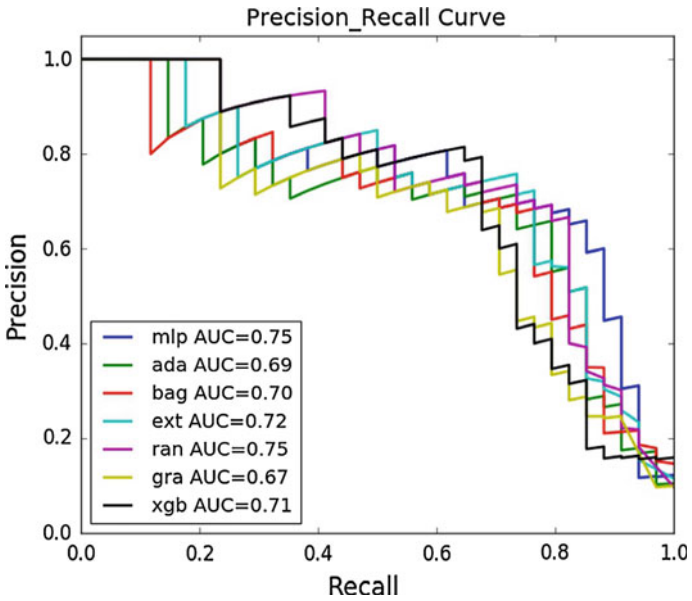


Fig. 2 Precision-Recall curves for the second strategy

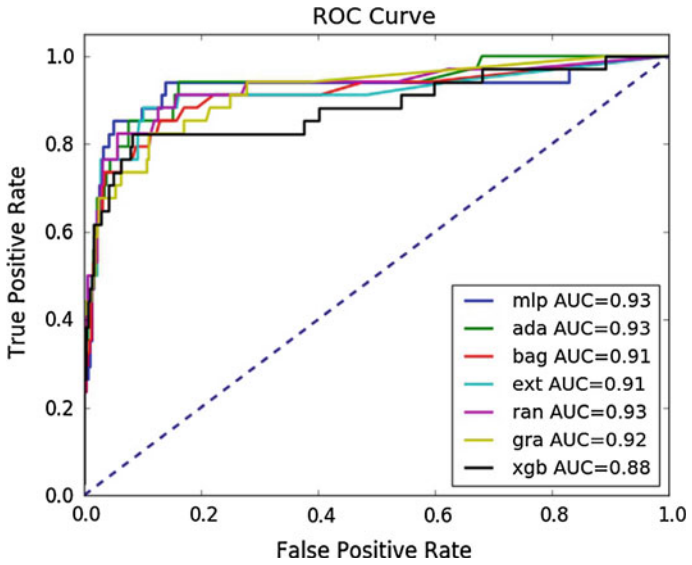


Fig. 3 ROC curves for the third strategy

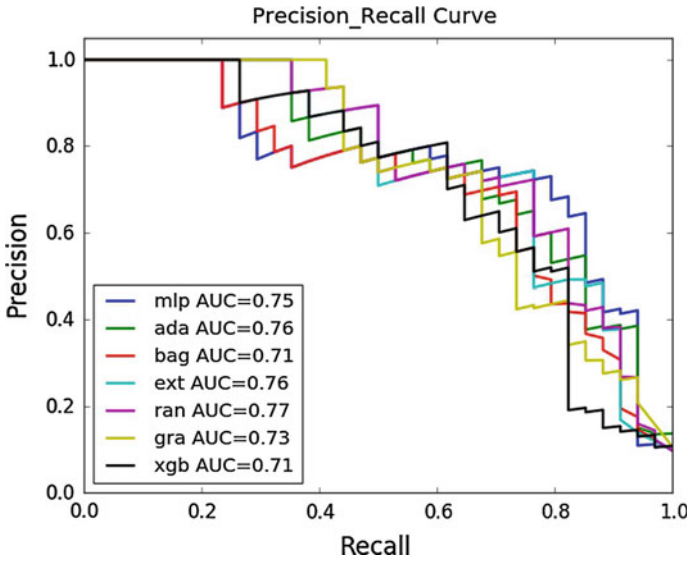


Fig. 4 Precision-Recall curves for the third strategy

Table 6 Mean, standard deviation and median across all estimators *for ROC AUC values*

Number of features	Mean	Standard deviation	Median
1,220[a]	0.91	0.026	0.92
1,759[a]	0.90	0.017	0.90
2,778[a]	0.92	0.024	0.92
3,395[a]	0.91	0.015	0.91
6,204[a]	0.91	0.025	0.91
1,292[b]	0.91	0.014	0.92
491[c]	0.92	0.018	0.92

[a]According to the first strategy
[b]According to the second strategy
[c]According to the third strategy

Now, consider the Precision-Recall curves and AUC for each estimator, according to strategy two in Fig. 2 and strategy three in Fig. 4. Corroborating all previous conclusions, when comparing Figs. 2 and 4, there is not much difference between those strategies. Here the ideal curve would correspond to both *unitary* precision and recall (TPR), that is, the *upper right* corner, also resulting *unitary* AUC. Notice the jagged effect of the curves, resulting from the nonlinear staircase effect in ROC curves. For wide ranges of Recall and threshold values, Precision changes very little. However, for each estimator, along the discontinuities, Precision might drop significantly, meaning that for a value of Recall along those boundaries there is a wide Precision range.

In order to understand the statistical significance in the variation of AUC figures, Tables 6 to 9 were designed to show the values of *mean, standard deviation and median* evaluated from Tables 4 and 5, across all estimators (Tables 6 and 7) and across all number of features (Tables 8 and 9), obtained according to the three strategies already described. It is observed that every AUC value fit quite well within the two standard deviation range, with almost no apparent skewness. The global ROC AUC mean and standard deviation evaluated across estimators and number of features (and strategies) were respectively 0.91 and 0.020, while the corresponding global Precision-Recall AUC mean and standard deviation were 0.72 and 0.031 respectively.

In this sense, *ExtraTrees and RandomForest* estimators came strong, consistent across all strategies, right in accord with the literature [39], but MLP came surprisingly well in all accounts. The worse results came from *XGBoost,* not because of the estimator in itself, but perhaps it could have used some more tuning. Regarding the feature selection strategy and number of features, it can be noticed from Table 6 and 7 that the third strategy with 491 features did not show any signs of appreciable bias, with the final results being even better across all estimators than any other strategy with many more features. Also, due to that small number of features, hundreds more individual estimators could be applied. Moreover, it is long known that there is no single optimum approach to feature selection [21]. Some

Table 7 Mean, standard deviation and median across all estimators *for Precision-Recall AUC values*

Number of features	Mean	Standard deviation	Median
1,220[a]	0.72	0.030	0.73
1,759[a]	0.71	0.033	0.70
2,778[a]	0.72	0.029	0.71
3,395[a]	0.72	0.032	0.71
6,204[a]	0.72	0.042	0.72
1,292[b]	0.71	0.030	0.71
491[c]	0.74	0.025	0.75

[a]According to the first strategy
[b]According to the second strategy
[c]According to the third strategy

Table 8 Mean, standard deviation and median across all number of features (obtained according to the strategies) *for ROC AUC values*

Estimator	Mean	Standard deviation	Median
mlp	0.91	0.019	0.91
ada	0.93	0.016	0.93
bag	0.91	0.012	0.91
ext	0.92	0.011	0.92
ran	0.92	0.013	0.92
gra	0.90	0.014	0.89
xgb	0.88	0.008	0.88

Table 9 Mean, standard deviation and median across across all number of features (obtained according to the strategies) *for Precision-Recall AUC values*

Estimator	Mean	Standard deviation	Median
mlp	0.71	0.026	0.70
ada	0.74	0.023	0.74
bag	0.70	0.013	0.71
ext	0.75	0.018	0.75
ran	0.75	0.013	0.75
gra	0.68	0.026	0.67
xgb	0.70	0.010	0.69

further reduction (to about a hundred features) was tried, but the results degraded very quickly.

The best threshold value for a given estimator to make predictions over test (*unseen*) data could be the one for the maximum Matthew's correlation coefficient (MCC) for each strategy. In order to analyze this relationship, Table 10 was produced only for the third strategy. Table 10 contains for each estimator the recorded values of the largest threshold value for the maximum MCC score and the respective values of TP, TN, FP, FN, Precision, Recall (TPR) and FPR considering the third strategy and the validation dataset. From the point of view of each max MCC, notice how the threshold values ended up skewed regarding the 50% transition probability. Moreover, reading from Fig. 4 it seems that max MCC represents the best compromise between Precision and Recall and, in that case, the winners are

Table 10 Limit threshold value and respective TP, TN, FP, FN, Precision, Recall (TPR) and FPR for each estimator, for maximum MCC, according to the *third* strategy, using validation data

Estimator	Threshold	Max MCC	TP	TN	FP	FN	Precision	Recall (TPR)	FPR
mlp	0.29	0.73	26	307	9	8	0.74	0.76	0.028
ada	0.30	0.64	19	311	5	15	0.79	0.56	0.016
bag	0.40	0.67	24	306	10	10	0.71	0.71	0.032
ext	0.19	0.71	26	306	10	8	0.72	0.76	0.032
ran	0.17	0.71	26	306	10	8	0.72	0.76	0.032
gra	0.10	0.66	22	308	8	12	0.73	0.65	0.025
xgb	0.49	0.68	21	311	5	13	0.81	0.62	0.016

mlp, *ext and ran*. Assuming Recall the most important score, reading from Figs. 3 and 4, a very high 0.90 Recall would take the FPR up to 0.18, and would bring the Precision tumbling down to about 0.40 (*mlp*).

Finally, it is not easy to compare the results from the original NIPS 2003, since DOROTHEA [12] was just one of the sets compounding that challenging event. However, regarding AUC ROC alone, the global mean value obtained in the present work would easily rank among the 20 first winners, as can be seen in [40]. Notice that in that report, the definition of ROC curve has been wrongly defined as TPR versus FNR (*false negative rate*), instead of FPR (*false positive rate*).

10 Conclusions

The growth of data is overwhelming. There is more data than ever in human history available in the internet from every source, even through open repositories, such as the UCI Machine Language Repositories. It is becoming an impossible task for humans alone to sort that out. Thus ML, born from artificial intelligence (AI) and statistics, is getting so much traction among researchers and specialists in all knowledge fields. The case of highly dimensional data with imbalanced classes is particularly relevant to RUL studies, since feature number can grow up very easily. Moreover, instance production can be very expensive or experimentally challenging and "failure" for systems or equipment is clearly a rare instance.

Some fundamentals of data analysis and ML have been presented and the use of Python and Python libraries has been brought to attention, as an approachable way to load, pre-process and perform ML on raw data obtained from the Web. In order to illustrate the use of those libraries, DOROTHEA [12] dataset was selected. A highly imbalanced binary classification problem, with one hundred thousand highly sparse anonymized binary features and limited number of instances, was divided into train, validation and test data. In order to reduce the number of features after removing the duplications, three different hybrid strategies were designed and employed, and several classifiers, mostly ensemble models, were applied.

Since the usual accuracy score (ACC) does not work for imbalanced classes, receiver operating characteristics (ROC) and Precision-Recall curves were obtained for all classifiers, from which only those from strategies three and four are shown in the present work. Area under the curves (AUCs) of ROC and Precision-Recall curves for all strategies as well some simple statistics and estimators are tabulated and compared. Strategy four reduced the data to less than 0.5% of the 100,000 original features and these results showed to be as good as or even better than all the others with many more features. In order to obtain better accuracies than those in the present book chapter, further correlations between instances and classes must be uncovered, perhaps using binary encoding and information theory.

Acknowledgements In order to approach DOROTHEA, *Python, numpy, matplotlib, pandas, scipy sparse,* and mostly *scikit-learn* were employed all over to facilitate all the work. Therefore, the author is very grateful to the developers of those wonderful open-source packages. The author must acknowledge DuPont Pharmaceuticals Research Laboratories as well as KDD Cup 2001, for gracefully allowing the use of the data from which DOROTHEA dataset was built. Finally, the author wishes to thank Dr. João Paulo Dias, from the Department of Mechanical Engineering of the Texas Tech University, for contributing with his comments on the manuscript and for his invaluable help with the references organization.

Appendix

Fast and memory-light procedure to search for duplicated features in large sparse arrays. X is the feature matrix (1950, 100000) and X^T its transpose.
Python code follows:

```
import numpy as np
import  collections
import scipy.sparse as sp
from time import time



start = time()
d = collections.OrderedDict()



remove = np.zeros((100000,),dtype='int8')
Xsp = sp.lil_matrix(X.T)
```

```
for j,row in enumerate(Xsp):
t = str(row)
if t in d and j != 0:
remove[j] = 1
else:
d[t] = 1
if j%10000 == 0: print j,



print '%g s' %(time() - start)
```

References

1. R.D. Peng, E. Matsui, *The Art of Data Science: A Guide for Anyone Who Works with Data* (Skybrude Consulting, LLC, 2016)
2. H. Koepke, 10 Reasons Python Rocks for Research (And a Few Reasons It Doesn't) (University of Washington, 2010), http://www.stat.washington.edu/~hoytak/blog/whypython.html
3. Learn More About Anaconda, https://www.continuum.io/documentation
4. Enthought Scientific Computing Solutions, https://www.enthought.com
5. S. van der Walt, S. C. Colbert, G. Varoquaux, The NumPy array: a structure for efficient numerical computation (2011)
6. E. Jones, E. Oliphant, P. Peterson, *SciPy: Open Source Scientific Tools for Python* (2001), http://www.scipy.org/
7. J. D. Hunter, Matplotlib: a 2D graphics environment. Comput. Sci. Eng. 90–95 (2007)
8. F. Pérez, B.E. Granger, IPython: a system for interactive scientific computing. Comput. Sci. Eng. **9**(3), 21–29 (2007)
9. F. Anthony, *Mastering Pandas* (Packt Publishing Ltd., 2015)
10. S. Seabold, J. Perktold, Statsmodels: econometric and statistical modeling with Python, in *Proceedings of the 9th Python in Science Conference* (2010), pp. 57–61
11. Scikit-learn, Documentation of scikit-learn 0.17 (2014), http://scikit-learn.org/stable/documentation.html
12. I. Guyon, Design of experiments of the NIPS 2003 variable selection benchmark, in *NIPS 2003 Workshop on Feature Extraction* (2003)
13. J. Leek, *The Elements of Data Analytic Style* (Kindle Edi, Leanpub, 2015)
14. T.M. Mitchell, *Machine Learning* (McGraw-Hill Science/Engineering/Math, 1997)
15. K. Markhan, Introduction to machine learning with scikit-learn. *Kaggle's blog* (2015), https://github.com/justmarkham/scikit-learn-videos
16. A. Smola, S.V.N. Vishwanathan, *Introduction to Machine Learning* (Cambridge University Press, 2008)
17. P. Domingos, A few useful things to know about machine learning. Commun. ACM **55**(10), 78–87 (2012)
18. A. Boschetti, L. Massaron, *Python Data Science Essentials* (Packt Publishing Ltd., 2015)

19. L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A.C. Muller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. Vanderplas, A. Joly, B. Holt, G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, in *European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases* (2013), pp. 1–15
20. M. Lichman, *UCI Machine Learning Repository,* (University of California, School of Information and Computer Science, Irvine, CA, 2013)
21. V. Bolon-Canedo, N. Sanchez-Maroño, A.A. Betanzos, *Feature Selection for High-Dimensional Data* (Springer, 2015)
22. J. Fan, R. Li, Statistical challenges with high dimensionality: feature selection in knowledge discovery, in *Proceedings of the International Congress of Mathematicians*, Madrid, Spain (2006), pp. 595–622
23. A. Singh, A. Purohit, A survey on methods for solving data imbalance problem for classification. Int. J. Comput. Appl. **127**(15), 37–41 (2015)
24. S.V. Jadhav, V. Pinki, A Survey on feature selection algorithm for high dimensional data. Int. J. Recent Innov. Trends Comput. Commun. **4**(1), 83–86 (2016)
25. F. Chang, J. Guo, W. Xu, K. Yao, A feature selection method to handle imbalanced data in text classification. J. Digit. Inf. Manage. **13**(3), 169–175 (2015)
26. M. Imran, M. Afroze, A.V. Kumar, A.A.M. Qyser, Learning from imbalanced data of diverse strategies with investigation. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **5**(6), 1285–1290 (2015)
27. A. Ali, S.M. Shamsuddin, A.L. Ralescu, Classification with class imbalance problem: a review. Int. J. Adv. Soft Comput. Appl. **7**(3), 176–204 (2015)
28. A. Sonak, R.A. Patankar, A survey on methods to handle imbalance dataset. Int. J. Comput. Sci. Mobile Comput. **4**(11), 338–343 (2015)
29. A.H.M. Kamal, X. Zhu, A. Pandya, S. Hsu, R. Narayanan, Feature selection for datasets with imbalanced class distributions. Int. J. Softw. Eng. Knowl. Eng. **20**(2), 113–137 (2010)
30. R. Balasubramanian, S.J.S.A. Joseph, Intrusion detection on highly imbalance big data using tree based real time intrusion detection system: effects and solutions. Int. J. Adv. Res. Comput. Commun. Eng. **5**(2), 27–32 (2016)
31. C. Chen, A. Liaw, L. Breiman, Using random forest to learn imbalanced data, (2004)
32. Y. Liu, J. Cheng, C. Yan, X. Wu, F. Chen, Research on the Matthews correlation coefficients metrics of personalized recommendation algorithm evaluation. Int. J. Hybrid Inf. Technol. **8** (1), 163–172 (2015)
33. M. Bekkar, H.K. Djemaa, T.A. Alitouche, Evaluation measures for models assessment over imbalanced data sets. J. Inf. Eng. Appl. **3**(10), 27–38 (2013)
34. J.M. Lobo, A. Jiménez-Valverde, R. Real, AUC: a misleading measure of the performance of predictive distribution models. Glob. Ecol. Biogeogr. **17**(2), 145–151 (2008)
35. P. Flach, J. Hernández-Orallo, C. Ferri, A coherent interpretation of AUC as a measure of aggregated classification performance, in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA (2011), pp. 657–664
36. L. Breiman, *Random Forests*, Berkeley, CA (2001)
37. A. Liaw, M. Wiener, Classification and regression by randomForest. R News **2**(3), 18–22 (2002)
38. G. Hackeling, *Mastering Machine Learning with Scikit-Learn* (Packt Publishing Ltd., 2014)
39. D. Muchlinski, D. Siroky, J. He, M. Kocher, Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. Polit. Anal. **24**(1), 87–103 (2016)
40. I. Guyon, S. Gunn, A. Ben Hur, G. Dror, Result analysis of the NIPS 2003 feature selection challenge. Adv. Neural Inf. Process. Syst. **17**, 545–552 (2003)

The Use of Trend Lines Channels and Remaining Useful Life Prediction

Luciano Barbanti, Berenice Camargo Damasceno, Aparecido Carlos Gonçalves and Hadamez Kuzminskas

Abstract One of the most important aspects in a working machine is the remaining useful life (RUL) of its components. Prognostics in this case depends on establishing the cause-effect entries in the process as well as how it behaves from the series of measures done under experimental conditions. This work introduces two techniques in analyzing series data coming originally from the financial market frame. One of them is the Bollinger Bands theory and another is the Markowitz theory on composite series. Both have a wide spectrum of applications in the cause-effect series prediction.

Keywords Bollinger bands · Trend lines · Forecasting · H. Markowitz theory · Saturated data sequence

1 Introduction

Here we are proposing two methods for determining pairs of cause-effect action and its resulting intensity in a series of data measured in a wind turbine by considering frequency velocity, temperature, viscosity of lubricants, and forces that can potentially cause the structural damage on components due to crack propagation [1].

In experiments, the results outputs are generally done by a sequence of measured data. In this work—through the consideration of a historical data series of such

L. Barbanti (✉) · B.C. Damasceno
Department of Mathematics, São Paulo State University (UNESP), Ilha Solteira, SP, Brazil
e-mail: barbanti@mat.feis.unesp.br

A.C. Gonçalves
Department of Mechanical Engineering, São Paulo State University (UNESP), Ilha Solteira, SP, Brazil

H. Kuzminskas
Department of Electrical Engineering, São Paulo State University (UNESP), Ilha Solteira, SP, Brazil

S. Ekwaro-Osire et al. (eds.), *Probabilistic Prognostics and Health Management of Energy Systems*, DOI 10.1007/978-3-319-55852-3_11

outputs—we will propose a specific technique for forecasting analysis known in the literature as the Bollinger Bands (BB) procedure, widely used to perform analysis on the stock market area. Moreover, we are proposing a method to transform a series in a saturated one (the definition will be provided on sub-section *3.2*) to induce extreme situations when using it in the optimal predictive theory credited to H. Markowitz [6]. In the same way the Markowitz theory has its use in the stock market.

2 Bollinger Bands

Developed by Bollinger [2], the Bollinger Bands (BB) are volatility bands placed above and below a moving average. Volatility is based on the standard deviation with respect to a moving average, which changes as volatility changes: the bands automatically widen when volatility increases and then narrow when volatility decreases. The dynamic nature of BB indicates when data are sub or super evaluated with respect to a "normal" value done by the moving average (MA).

The construction of BB is illustrated as follows: given the sequence of output data in an experiment, let us fix a natural number $n > 1$ and then the n-MA (i.e., the moving average of length equal to n in the data sequence). Let σ be the standard deviation associated to the n-MA, and $k > 0$, a real number. The BB is constituted, then, by the three curves: the n-MA, a lower band $(= MA - k\sigma)$ and the upper band $(= MA + k\sigma)$. The Bollinger strip is the region in the plane confined by the upper and lower bands (Fig. 1).

Fig. 1 Bollinger bands with n = 20 and k = 2

The use of the BB provides us with a powerful tool to determine the relationship of cause-effect especially in extreme situations, as proved by Bollinger [2] in the framework of the financial market. The main characteristic of the BB is, by considering statistical confidence intervals, that at least in 94% of the cases, the next real data in the experiment is expected to be in the Bollinger strip with $n = 20$ and $k = 2$. Then, when the data in the series are out of that strip, we have an instance that is out of the normal statistical pattern. When the data is above the upper band, we say that we having a super-valued data situation, and a sub-valued data in the opposite situation.

The parameters given by Bollinger are the most commonly found, but sometimes we could have more efficient changes. As shown in [3], with the use of weighted average in the field of the derivatives options market, the utilization of the BB with $n = 12$ and $k = 2$ is a more efficient strategy than if BB is considered with $n = 20$.

Despite the importance of this method in the literature and the massive amount of research that has been applied in the past 15 years in the financial market, only a very few number of work applications (specifically in the general engineering literature) are available. As an example of one of those rare works, Ngan and Pang [4] have used BB to inspect and indicate defective areas in patterned fabric.

The method of BB, with its predictive character, surely will be releasing elements for the study of the RUL of a system. In fact, when a local or global tendency is identified in a process, the use of the BB method enables us to analyze the oscillations of the measured data and how far they are around the tendency line.

There are other procedures in the literature that enable us to see if a data in a sequence is sub or super valued by using trend lines. It is the case of the known Parabolic System Approach, the Stop-and-Reverse (SAR) techniques, Relative Strength Index (RSI), the Moving Average Convergence-Divergence (MACD), the Fibonacci Analysis, the Elliot wave analysis, and Ichimoku clouds, among others, [5]. A very extensive research field for future works it is based on the identification of the characteristic parameters for each of the procedures mentioned above by using experimental data, and then applying the prognostic information to RUL analysis.

3 The H. Markowitz Theory and Saturated Series

When measuring aspects of a phenomenon in a data series there are in general several series of cause-effect pairs embedded in such original series.

In the terminology of the financial market, the data series is the composition of all the other series in a "portfolio". The dynamics of this "portfolio" is very efficiently described in the literature as the Optimal Portfolio Theory by H. Markowitz [6, 7]. The fundamental aspect in this theory is represented by the possibility of combining weighted composing series in a specified way, in order to vary the risk (represented by the standard deviation) by varying the average in the original series.

Several attempts were made, when the theory was first published, in order to reduce the portfolio risk. In this sense, slightly different risk definitions were done. As an illustrative example, see the definitions of the average absolute deviation and of the semi-variance as in Chap. 4.2 in Elton and Gruber [7].

Next, we propose a modification of the original series based on the above framework and according to a method called "saturation of series" presented by Damasceno and Barbanti [8]. As it will be shown, it results in a modification of the original series. Moreover, Damasceno and Barbanti [8] have shown that, by using a random example, the saturated series improves the result of the Markowitz theory when the original series is used.

3.1 The Markowitz Strategy

The Markowitz strategy [9] is based on the following general rules:

- It is necessary to diversify the stocks in a portfolio (i.e. choosing the lowest correlated stocks to put together or "do not put all your eggs in a single basket");
- It is necessary to balance the stocks in the Portfolio following the principle;
- In the Markowitz plane (risk − the standard deviation mean × turn − mean value) take the return and risk in the portfolio P as the composite form from the return and risk of the series that compose P. For instance, in the minimum possible case diversification (2 stocks A, B composing P) with the proportion x, $P = x\,A + (1 - x)\,B$, the return and risk of P are:

$$\overline{R_P} = x\ \overline{R_A} + (1-x)\ \overline{R_B} \tag{1}$$

and

$$\sigma_P^2 = x^2\sigma_A^2 + (1-x)^2\sigma_B^2 + 2x\,(1-x)\,\sigma_{AB} \tag{2}$$

where σ_{AB} is the AB—covariance. The minimum risk in the composition of the stocks series in the Portfolio is done by, using derivations, the value x_0 that solves the equation (in x):

$$2x\,\sigma_A^2 + 2\,(1-x)\,\sigma_B^2 + 2\,(1-2x)\,\sigma_{AB} = 0 \tag{3}$$

then

$$x_0 = \frac{-\sigma_B^2 - \sigma_{AB}}{\sigma_A^2 - \sigma_B^2 - 2\sigma_{AB}}. \tag{4}$$

Moreover, if we use σ (non-usual sign) to denote the AB—correlation, the curve composition of A, B in the Markowitz plane looks like the Fig. 2.

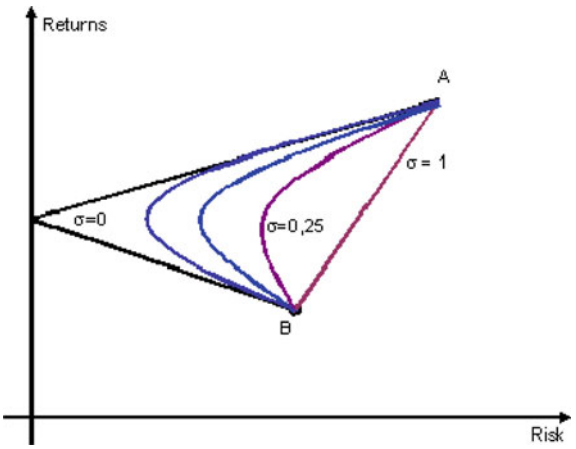


Fig. 2 Risk × Return curves for composition of A and B

3.2 The Saturation Method for Series

Given the original series S, we introduce a modification on it to obtain a new series [8]. In an inductive manner we construct the series S_0, S_1, …, beginning with:

$$S_0 = S. \tag{5}$$

Assuming that

$$S_j: s_1^j, s_2^j, \ldots, s_k^j, \ldots \tag{6}$$

This allows us to construct the series $S_{j+1}, j = 0, 1, 2, 3, \ldots,$ and define for the series S_j linear regression in the plane (x, y)

$$L_j: y = \alpha_j x + \beta_j. \tag{7}$$

The S_{j+1} is the series with the elements:

$$s_k^{j+1} = s_k^j \quad if \quad s_k^j \leq \alpha_j k + \beta_j, \tag{8}$$

$$s_k^{j+1} = \alpha_j k + \beta_j \quad if \quad s_k^j > \alpha_j k + \beta_j. \tag{9}$$

Note that we could apply in the definition of S_{j+1} the signs $\geq$ and $<$ instead of $\leq$ and $>$, respectively, creating in this way a new series T_{j+1} from T_j by fixing $T_0 = S$.

It can be notice also that the series $(S_j)_{0 \leq j}$ and $(T_j)_{0 \leq j}$ have upward and downward bias, respectively.

Damasceno and Barbanti [8], also pointed out the existence a situation in Brazilian stock market (BM&F-BOVESPA) in which information provided by the Markowitz Theory was improved with the use of saturated series, when compared to the use of the original series *S*.

4 The Next Step

The next step will be towards the identification of parameters in a series concerning data obtained from wind turbine measurements, specifically of n and k above, grounded on the physics underlying the experiments themselves. The general purpose is to establish some cause–effect connections by making inferences on the value of future data, and in this way, leading to RUL predictions. This can also be done through the saturation of the original series obtained in the process.

References

1. S. Sankararaman, Significance, interpretation and quantification of uncertainty in prognostics and remaining useful life prediction. Mech. Syst. Signal Process. **52–53**, 228–247 (2015)
2. J. Bollinger, *Bollinger on Bollinger Bands* (Mc-Graw Hill, 2001)
3. W. Liua, X. Huanga, W. Zhenga, Black–Scholes' model and Bollinger bands. Physica A **371**(2), 565–571 (2006)
4. H.Y.T. Ngan, G.K.H. Pang, Novel method for patterned fabric inspection using Bollinger bands. Opt. Eng. **45**(8), 45–56 (2006)
5. S-C.T. Chou, H.-J. Hsu, C-C. Yang, F. Lai, A stock selection DSS combining AI and technical analysis. Ann. Oper. Res. **75**(0), 335–353 (1997)
6. H. Markowitz, Portfolio selection. J. Finance **7**(1), 77–91 (1952)
7. E.J. Elton, J. Gruber, *Theory and Investment Analysis* (Wiley, 2005)
8. B.C. Damasceno, L. Barbanti, Composing a portfolio in stock markets by using a series saturation method. Anais do CNMAC **2**, 277–283 (2009)
9. C. Rowland, J.M. Lawson, *The Permanent Portfolio: Harry Browne's Long-Term Investment Strategy* (Wiley, 2012)

The Derivative as a Probabilistic Synthesis of Past and Future Data and Remaining Useful Life Prediction

Berenice Camargo Damasceno, Luciano Barbanti, Hadamez Kuzminskas and Márcio Antonio Bazani

Abstract The concept of remaining useful life (RUL) is crucial when dealing with mechanical systems. RUL is taken into account in a system through a series of prognostics and the study of stability in a related data series. This paper is focused on a powerful optimal technique in prognosis coming from the Grünwald–Letnikov definition of derivative.

Keywords Grünwald–Letnikov derivative · Probability · Forecasting

1 Introduction

Remaining useful life (RUL) of a mechanical systems or equipment is the survival time that the unit has. The predictions allowed by RUL models provide schedules for decision-making in production management and the maintenance of the unit itself [1, 2].

In this work consider the Grünwalk–Letnikov derivative definition, which consists in a version of the Newtonian derivative. It allows the n-th derivative in a point as being the probabilistic synthesis of present and past data that can be used to forecast the behavior of future data.

B.C. Damasceno (✉) · L. Barbanti
Department of Mathematics, São Paulo State University (UNESP),
Ilha Solteira, SP, Brazil
e-mail: berenice@mat.feis.unesp.br

H. Kuzminskas
Department of Electrical Engineering, São Paulo State University (UNESP),
Ilha Solteira, SP, Brazil

M.A. Bazani
Department of Mechanical Engineering, São Paulo State University (UNESP),
Ilha Solteira, SP, Brazil

S. Ekwaro-Osire et al. (eds.), *Probabilistic Prognostics and Health Management of Energy Systems*, DOI 10.1007/978-3-319-55852-3_12

However, the application of results presented in this chapter to analyse real data coming from experiments may demand for more complex structured models, as the ones using fractional calculus and/or with large interval of data [3–8].

2 The Grünwald–Letnikov Derivative

Let this be the first derivative of a function $f(x)$, given by the Newton form:

$$f^{(1)}(x) = \lim_{h\to 0}\left[\frac{f(x)-f(x-h)}{h}\right]. \tag{1}$$

In a recursive way, we can write:

$$f^{(2)}(x) = \lim_{h\to 0}\left\{\frac{\lim_{h\to 0}\left[\frac{f(x)-f(x-h)}{h}\right]-\lim_{h\to 0}\left[\frac{f(x-h)-f(x-2h)}{h}\right]}{h}\right\} \tag{2}$$

and

$$f^{(2)}(x) = \lim_{h\to 0}\left[\frac{f(x)-2f(x-h)+f(x-2h)}{h^2}\right]. \tag{3}$$

Then, the Grünwald–Letnikov formula for the n-th derivative of $f(x)$ is given by:

$$f^{(n)}(x) = \lim_{h\to 0}\left\{\left[\frac{1}{h^n}\left(f(x)-nf(x-h)+n(n-1)\frac{f(x-2h)}{2}-n(n-1)(n-2)\frac{f(x-3h)}{6}+\cdots\right)\right]\right\} \tag{4}$$

in which n = 1, 2, 3, …, for $n \in N$ indicates the drivative order.

Now, consider $r \in N$ in which $r \geq n$. Then we have that $n < r$ and the factorial relationship is true.

$$\frac{n!}{r!(n-r)!} = 0. \tag{5}$$

Thus, in general terms, Eq. (4) can be written as,

$$f^{(n)}(x) = \lim_{h\to 0}\left\{\frac{1}{h^n}\left[\sum_{r=0}^{n}\left[(-1)^r\left(\frac{n!}{r!(n-r)!}\right)f(x-rh)\right]\right]\right\}. \tag{6}$$

In (Eq. 6) it is possible to identify that, roughly speaking, the n-th derivative in a point is the probabilistic synthesis of the present ($x = 0$) and the displaced points are $x-h, x-2h, \ldots, x-nh$. This is an advantage in the method, because the

expected value of f in a displaced point can be verified in several manners for different n.

Some immediate results can be inferred when observing Eq. (6) in probabilistic terms: for every $n = 1, 2, 3, \ldots$ we have, considering $r = 0$, that

$$\frac{n!}{r!(n-r)!} = 1. \tag{7}$$

This result shows us that for the initial value x, the probability of $f(x), f^{(1)}(x), \ldots, f^{(n)}(x)$ is equal to 1, and then the expression

$$-\sum_{r=1}^{n}\left[(-1)^r\left(\frac{n!}{r!(n-r)!}\right)\right] = 1, \tag{8}$$

for n = 1, 2, 3, …, and $r \leq n$.

Thus,

$$-\sum_{r=1}^{n}\left[(-1)^r\left(\frac{n!}{r!(n-r)!}\right)\right] f(x-rh), \tag{9}$$

can be seen as, the expected value of the variable $Y = f(rh)$ with

$$\mathbb{E}\,(Y = f(rh)) = \frac{n!}{r!(n-r)!} = \left|\left[(-1)^r\left(\frac{n!}{r!(n-r)!}\right)\right]\right| \quad \text{for} \quad n = 1, 2, 3, \ldots, \text{ and } r \leq n. \tag{10}$$

The above process enables us to deduce the expected value for f at future or past points (relative to a fixed point x).

3 An Example

Choose a series of numbers (e.g. as the data in an experiment) and the function $f: \mathbb{R} \to \mathbb{R}$, an optimal continuous approximation of the series itself.

Let us take as an instance $n = 3$ *for* $r = 1, 2$ *and* 3. The the resulting Grünwald–Letnikov formula in Eq. (6) is:

$$f^{(3)}(x) \cong \frac{f(x)}{h^3} - 3\frac{f(x-h)}{h^3} + 3\frac{f(x-2h)}{h^3} - \frac{f(x-3h)}{h^3}, \tag{11}$$

where the signal ≅ means that when considering a fixed h (which can made as small as we like) the symbol for equality, =, is applied.

Since we have the expected values for some $f(x-h)$, $f(x-2h)$, $f(x-3h)$ [$h>0$, and $h<0$], this example shows that equilibrium equation in (11) enabling us to find the probability of the unknowns values of f among the points

$$x - sh;\ s = 1, 2, 3. \tag{12}$$

Notice that as long as we increase n, we are refining the prognostics values in the original series. Furthermore, we see that the same values in (12) can be considered for other values of n.

4 An Extension

As previously mentioned, the Grünwald–Letnikov derivative works well in the field of Dynamics described by fractional derivative. In fact, the Grünwald–Letnikov derivative is the most suitable to permit numerical treatment [5] and can be extended from the integer derivatives to the fractional ones.

In fact, the equality in Eq. (8) can be transformed when extending the notion of factorial on natural n by the function Γ on the real positive numbers α.

Thus, we have in the case of the fractional derivative D^{α} (for $0<\alpha<1$):

$$D^{\alpha} f(x) = \lim_{h \to 0} \left(\frac{1}{h^{\alpha}} \left(\sum_{r=0}^{\infty} \gamma(\alpha, r) f(x - rh) \right) \right), \tag{13}$$

where

$$\gamma(\alpha, r) = (-1)^{r} \frac{\Gamma(\alpha + 1)}{r! \Gamma(\alpha - r + 1)}. \tag{14}$$

This definition can be extended to Eqs. (4), (5), and (7), allowing in this way the extension of the result as in Eq. (10) to systems modeled with fractional derivatives, that is, a real new possibility in modeling systems based upon experimental data.

5 Conclusion

By making the above considerations, we could see that the formulation of derivatives due to Grünwald–Letnikov allows us to synthetize the expected value at points of a data series (represented here by points of an optimum approximation function f), contributing in this way to the enrichment of the predictive techniques in the domain of the RUL theory.

References

1. J.Z. Sikorska, M. Hodkiewicz, L. Ma, Prognostic modeling options for remaining useful life estimation by industry. Mech. Syst. Signal Process. **25**(5), 1803–1836 (2011)
2. S. Sankararaman, Significance, interpretation and quantification of uncertainty in prognostics and remaining useful life prediction. Mech. Syst. Signal Process. **52–53**, 228–247 (2015)
3. B. Tremeac, F. Meunier, Life cycle analysis of 4.5 MW and 250 W wind turbines. Renew. Sustain. Energy Rev. **13**(8), 2104–2110 (2009)
4. D. Baleanu, K. Diethelm, E. Scalas, *Fractional Calculus Models and Numerical Methods* (World Scientific Book, 2011)
5. V.E. Tarasov, Lattice model of fractional gradient and integral elasticity: long-range interaction of Grünwald–Letnikov–Riesz type. Mech. Mater. **70**, 106–114 (2014)
6. R. Scherer, S.L. Kalla, Y. Tang, J. Huang, The Grünwald–Letnikov method for fractional differential equations. Comput. Math. Appl. **62**(**3**), 902–917 (2011)
7. R, Garrappa, A Grunwald–Letnikov scheme for fractional operators of Havriliak–Negami type. Recent Adv. Appl. Math. Modell. Simul. **34**, 70–76 (2014)
8. R.J. Hyndman, G. Athanasopoulos, *Forecasting: Principles and Practice* (OTexts, 2014)

Part III
Condition Monitoring

Monitoring and Fault Identification in Aeronautical Structures Using an Wavelet-Artificial Immune System Algorithm

Fernando P.A. Lima, Fábio R. Chavarette, Simone S.F. Souza and Mara L.M. Lopes

Abstract This chapter presents a Wavelet-Artificial Immune System (WAIS) algorithm to diagnose failures in aeronautical structures. Basically, after obtaining the vibration signals in the structure, the wavelet module is used to transform the signals into the wavelet domain. Afterward, a negative selection artificial immune system performs the diagnosis via identifying and classifying the failures. The main application of this methodology is in the auxiliary structures inspection process in order to identify and characterize the flaws as well as assist in the decision making process that is aiming at avoiding accidents or disasters. In order to evaluate this methodology, we carried out the modeling and simulation of signals from a numerical model of an aluminum beam that represent an aircraft structure such as a wing. The proposed algorithm presented good results, with 100% matching in detecting and classifying of the failures tested. The results demonstrate the robustness and accuracy of the methodology.

Keywords Wavelet-artificial immune systems (WAIS) · Monitoring and fault identification · Aeronautical structures · Artificial intelligence

1 Introduction

In the last few decades, the aeronautical industry have placed significant investments in research and technological development in order to obtain efficient methods to analyze the integrity of structures and to prevent disasters and/or accidents from happening to ensure the safety of people's lives and to avoid economic damages.

Fault diagnosis systems, aka Structural Health Monitoring Systems (SHMS), perform tasks such as: acquisition and data processing, validation and analysis,

F.P.A. Lima · F.R. Chavarette (✉) · S.S.F. Souza · M.L.M. Lopes
Department of Mathematics, São Paulo State University (UNESP),
Ilha Solteira, SP, Brazil
e-mail: fabioch@mat.feis.unesp.br

S. Ekwaro-Osire et al. (eds.), *Probabilistic Prognostics and Health Management of Energy Systems*, DOI 10.1007/978-3-319-55852-3_13

detection, characterization and interpretation of adverse changes in a structure so to assist in making decisions and identifying structural faults [1].

Structural failures occur as a consequence of factors such as component wear, cracks, loosening of screw connections or simply a combination of these. Regardless of the source, in most cases, structural failure causes a variation of spatial parameters of the structure, generating a reduced structural rigidity, mass, and also increased damping so that the dynamic behavior of the structure is changed [2].

To solve this problem, several solutions have been proposed such as traditional SHMS based on ultrasonic inspection, radiography (X-ray) or acoustic emission testing. However, these traditional techniques cannot meet increasing demands of industries, especially when the structures are in motion [3]. Thus, one solution to develop the most modern and efficient SHMS is the utilization of intelligent techniques, and efficient data acquisition systems.

In literature, several studies that utilize smart materials and SHMS that have robustness, accuracy and good performance are available. The following few paragraphs present the most relevant papers in this field.

Krawczuk et al. [4] presented the application of a genetic algorithm in conjunction with a Perceptron Multi-Layer neural network with back-propagation to perform fault detection and location in a numerical model of a beam. Giurgiutiu [5] used the method of electro-mechanical impedance to monitor aerospace structures with piezoelectric sensors attached. Palaia [6] presented a methodology for structural analysis of buildings using a non-destructive method (NDT). Chandrashekhar and Ganguli [7] proposed a fuzzy system to detect structural faults using curvature mode shapes.

Chen et al. [8] used a model that implement wavelet transform to evaluate the integrity of bridge structures through the vibration signals. A system for identification and location of damage in an airplane wing using a probabilistic neural network was proposed in [9]. Wang et al. [10] proposed a multimodal genetic algorithm for diagnosing damage in a steel truss bridge. Song et al. [11] proposed an experimental method to perform structural analysis of buildings. Souza et al. [12] proposed an ARTMAP-Fuzzy neural network applied to diagnosis of faults in buildings. Lima et al. [13], proposed an immune algorithm with negative selection to diagnose failures in aircraft structures.

Lima et al. [14] has presented a SHMS based on ARTMAP-Fuzzy neural network and wavelet transform, to diagnose faults in buildings. Lima et al. [15] presented a hybrid method based on ARTMAP-Fuzzy neural network and wavelet transform to diagnose failures in aluminum beams. Abreu et al. [16] presented a failure analysis tool in aircraft structures using complex wavelet transform.

In this paper, a new approach to fault diagnosis in aeronautical structures using a Wavelet-Artificial Immune System (WAIS) algorithm is presented. This methodology is divided into three main modules: the acquisition and processing of data, fault detection and classification. From the signal acquisition, the wavelet transform is applied to decompose the signals into four levels of resolution. After obtaining

the processed signals, the Negative Selection Algorithm (NSA) is applied to perform the detection of abnormalities in the structure, and thus the characterization of structural faults can be detected.

The Artificial Immune System (AIS) is a promising algorithms in Artificial Intelligence (AI). The concept is based on Biological Immune Systems (BIS) and aims to computationally reproduce its principal characteristics, properties and abilities [7]. As emphasized by [17], AIS is an adequate tool to be applied in failure diagnosis due to the natural characteristics of diagnoses.

The wavelet transform is a mathematical tool for signal analysis that decomposes or breaks the constituent signals into parts, allowing scientists to analyze the data at different levels of frequency with the resolution of each component in its range. In summary, the wavelet transform allows to view the approximation of the discontinuous data in functions (i.e., view the abnormalities in the signals) so that it can become an important tool in the analysis and diagnosis of abnormality in aeronautical structures. The use of a wavelet transform provides a sensitivity to the diagnosis system that allows the system to identify signal abnormalities easily.

Unlike several studies that have been presented in the literature, the main advantage of the method presented in this work is the ability to filter the signals using wavelet module, and thereafter applying the NSA, one of the most efficient techniques for failure diagnosis. This combination generates a powerful failure analysis tool that is demonstrated by the results obtained in this work. Thus, the main contribution of this work is a new efficient and accurate hybrid failure diagnosis approach composed of a signal processing mathematical tool, i.e. the wavelet transform, and an intelligent method, i.e. AIS.

In order to evaluate the proposed methodology, we have used one database containing the signals numerically simulated from a model of an aluminum beam that represents the wing of aircraft. This structure was modeled by finite elements and simulated in MATLAB. The results demonstrate the efficiency, accuracy and robustness of the proposed method.

This text is organized as follows: Sect. 2 presents the negative selection algorithm. Section 3 describes the wavelet transform. The modelling and simulation is presented in Sect. 4. Section 5 presents the proposed methodology and finally, the results and conclusions are presented, respectively, in Sects. 6 and 7.

2 Negative Selection Algorithm

The Negative Selection Algorithm (NSA), which was proposed in [18], detects changes in systems based on the biological process of negative selection of T lymphocytes, that occur in the thymus. This process works on the discrimination of proper versus non-proper cells. The algorithm is executed in two phases, according to the following description [19, 20]:

1. Censor:
   - Define a set of proper chains (S) to be protected;
   - Generate random chains and evaluate the affinity (Match) between each chain and the proper chains. If the affinity is greater than a predefined value, then reject the chain. Otherwise, file the chain into a detector set (R).
2. Monitor:
   - Given a set of chains to be protected (protected chains), evaluate the affinity with each chain and the detector set. If the affinity is superior to a predefined value, then a non-proper element is identified.

The censor-phase of the NSA primarily consists of generating a detector set from the data that were randomly chosen and verifying which data can then recognize a non-proper pattern. The detectors are similar to mature T cells, which can recognize pathogenic agents [21].

The monitoring phase consists of monitoring a system to identify a change in the behavior; thus, this phase classifies the change using the detector set that was created in the censor-phase. The censor-phase occurs offline, and the monitoring-phase occurs in real time [19, 21].

The antigen (Ag) is the signal to be analyzed in the negative selection algorithm and can be represented by

$$Ag = Ag_1, Ag_2, Ag_3, Ag_4, \ldots, Ag_L. \tag{1}$$

The detectors represent the antibodies (Ab) and are expressed as [17, 20]:

$$Ab = Ab_1, Ab_2, Ab_3, Ab_4, \ldots, Ab_L \tag{2}$$

where L is the dimension of the space of the antigen and the antibody.

2.1 Matching Criterion

To evaluate the affinity with the chains and to prove that they are similar, a matching criterion is used, which has the same meaning as the combination. The matching can be perfect or partial [22]. The matching is perfect when the two analyzed chains have the same value in every position, and the matching is partial when the patterns have only one identical position value to confirm the matching (which has been previously defined in [17]). This quantity is known as the affinity rate, and represents a similar grade for matching to occur between two analyzed chains [20]. Reference [22] defines the affinity rate as:

$$TAf = \left(\frac{An}{At}\right) * 100 \tag{3}$$

where

- TAf is the affinity rate,
- An is the quantity of normal rates in the problem (proper rates), and
- At is the total number of chains in the problem (proper and non-proper chains).

Equation (3) allows the precise calculation of the affinity rate for the proposed problem and represents the statistical analysis with the samples of the problem.

To dynamically improve the diagnosis, a deflection is proposed that is attached to the antibody (detector pattern—Ab), i.e., a tolerance with which it is possible to accept the combination with the patterns. This tolerance is defined according to Eq. (4) [17]. This deflection acts individually in each position i of vector (Ab), allowing verification of the matching in each position:

$$\underline{Ab_i} \leq Ag_i \leq \overline{Ab_i} \tag{4}$$

where:

- Ag_i is the nominal value of position i of the antigen (pattern under analysis),
- Ab_i is the nominal value of position i except for the deflection adopted at the antibody (detector pattern), and
- $\overline{Ab_i}$ is the nominal value of position i plus the deflection adopted at the antibody (detector pattern).

In this way, if the value of position i of antigen (Ag) is in the interval expressed in Eq. (4), then the position is considered to match. Thus, it is possible to quantify the affinity using the patterns analyzing position-by-position (point-by-point).

Equation (5) shown below represents the method for quantifying the total affinity with the analyzed patterns [23]:

$$Aft = \sum_{i=1}^{L} Pc_i \tag{5}$$

where:

- Af_T is the percentage of the affinity with the patterns analyzed,
- L is the total quantity of positions, and
- Pc is the matched position.

Thus, if Aft is greater than TAf, then the combination/matching with the patterns occurs, and the patterns are considered to be equal/similar. Otherwise, there is no matching with the patterns.

3 Wavelet Transform

The wavelet functions are mathematical transformations that can decompose functions, which allows these functions to be re-written in a more detailed form, i.e., with a global vision. Thus, it is possible to differentiate the local characteristics of a signal with different sizes (resolutions) and to analyze all of the signals by translations. Because most of the wavelets have compact support, they are useful in analyzing non-stationary signals. In this way, the wavelet analysis is better than the Fourier analysis [24].

There are several wavelet families. This work considers the orthonormal family functions and the Daubechies discrete family [25] due to having faster computational algorithms [24].

3.1 Discrete Wavelet Transform (DWT)

Define a signal $y[t] = (y_0, \ldots, y_{n-1}, y_n)$, which represents a discrete vector; then, it can be represented by a wavelet series, as follows [24]:

$$y[t] = \sum_{k=0}^{N_j} C_{j,k}\phi_{j,k}(t) + \sum_{j=J}^{1}\sum_{l=0}^{N_j} d_{j,k}v_{j,k}(t), \ \nabla t \in [0, N_0] \tag{6}$$

where J represents the resolution level, $N_j = (N/2) - 1$ represents the quantity of points in each new vector obtained by transformation; $\phi_{j,k}(t)$ and $v_{j,k}(t)$ are the wavelet and scale functions that perform the transformation; j is the scale (dilation); and k is the position (translation).

The Discrete Wavelet Transform (DWT), when applied directly to a signal to generate a set of coefficients, is calculated by several entrances into a G filter (low pass) and H filter (high pass), which are known as resolution levels. The filters G and H are calculated constant values vectors that provide an orthogonal base related to the scale and wavelet functions, respectively. This process is known as the Mallat Pyramidal algorithm [24] and is shown in Fig. 1.

In Fig. 1, C_0 corresponds to the original discrete signal $(C_0 = y[t])$, and H and G represent the low-pass and high-pass filters, respectively. The parameters d_1, d_2 and d_3 are the wavelet coefficients or the detail at each resolution level, and C_3 are the scale coefficients or approximations at the last level of the transform. These coefficients are obtained by a convolution of the constants with the filters represented in Eqs. (7) and (8) [24]:

$$C_{j+l,k} = \sum_{l=0}^{D-1} h_l C_{j,2k+l} \tag{7}$$

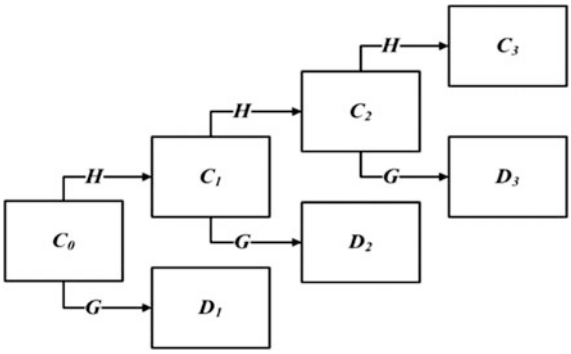


Fig. 1 Flowchart of the algorithm for DWT

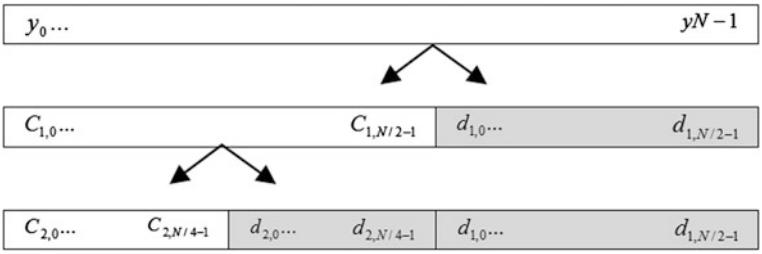


Fig. 2 Adaptation of the pyramidal algorithm for DWT

$$d_{j+l,k} = \sum_{l=0}^{D-1} g_l C_{j,2k+l} \tag{8}$$

where $k = [0, \ldots, (N/2^j) - 1]$, and D is the number of constants in the filter. Thus, the coefficients $C_{j,k}$ represent the average local media, and the wavelet coefficients $d_{j,k}$ represent the complementary information or the details that depart from the average media. Therefore, the transform coefficients, when ordered by scale (j) and position (k), are represented as follows [24]:

$$\psi = \left[[C_{J,k}]_{k=0}^{N_J}, \ \left[(d_{j,k})_{k=0}^{N_J} \right]_{j=J}^{1} \right] \tag{9}$$

in which ψ is a finite representation in terms of the coefficients of the signal decomposition in Eq. (6). Figure 2 shows the decomposition process of a signal at two resolution levels. Observe that at each transformation level, the size of the vectors is reduced by half $(N/2^J)$. Figure 2 represents an adaptation of Fig. 1 that represents the pyramidal algorithm for DWT.

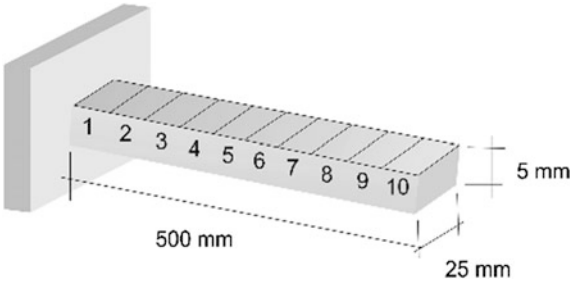


Fig. 3 Aeronautical structure model [26]

Table 1 Number of signals simulated

Wear level	Number of simulations
Normal condition (0%)	500
5%	150
10%	150
15%	150
20%	150
25%	150
30%	150
Total	1400

4 Modeling and Simulations

The proposed methodology is demonstrated considering a finite element model of an aluminum cantilever beam discretized by 10 finite elements, each having two degrees of freedom. The material has a modulus of elasticity, $E = 700$ GPa and a density of $\rho = 2710$ kg/m^3. The dimensions of the beam are 500 mm long, 25 mm wide and 5 mm thick. Figure 3 illustrates the discretized beam [26].

Using the beam model, several simulations were performed with different percentages of wear and locations of faults. The database consists of generated signals captured by an accelerometer attached to the beam. In all simulations, the beam was excited in the 3rd degree of freedom (element 2) and the signal was captured on the 19th degree of freedom (element 10). Thus, 1400 signals were simulated in the structure, 500 without wear (baseline condition) and 900 signs with wear (structural failure) as presented in Table 1. In the present analysis, 150 signals were simulated in each type of failure and 500 signals were simulated in normal conditions.

5 Proposed Methodology

The WAIS algorithm proposed in this work to detect and classify failures was based on the negative selection principle, and the phases are presented as follows:

5.1 Censor-Phase

This phase generates the proper detectors and the disturbance detector set. The detector sets are used by the diagnosis system during the monitoring process and are generated for each kind of signal of the database generated by modeling and simulation.

The proper detectors represent the baseline or normal condition of the structure. To generate this kind of detector, normal signals were randomly selected, and are defined as proper detectors. Once a proper detector is generated, it is then possible to generate the failure detectors. This process is illustrated in Fig. 4.

The next procedure is divided into three modules: the reading of the signals to create the detectors, the wavelet module that decomposes the signals using a discrete wavelet transform with four resolution levels, and the censor module with randomly chosen signals and that verifies the matching in relation to the proper detector set. If the affinity criterion is satisfied, the signals are rejected because they have proper characteristics. Otherwise, the signals are placed in the failure detector set.

The quantity of detectors that are used is determined by the operator. However, it is recommended to use 30% of the available data. The matching criterion is proposed in [27], which uses a deviate of 3%.

5.2 Monitoring-Phase

The monitoring-phase is divided into four modules: the input or the reading of the signals (by the acquisition data system), the wavelet module that decomposes the signals into four resolution levels, the detector module, which performs the discrimination of proper/non-proper, and the classification module to classify the failures. Figure 5 illustrates the monitoring-phase.

The wavelet module is executed after the acquisition of the signal and decomposes the signals by transforming them into the wavelet domain. Afterwards, the detector module compares the signals that are under analysis with the proper detectors to identify the matching with the signals. This module performs the diagnosis of the analyzed signals and classify them into proper and non-proper categories.

When an abnormality is detected, the abnormal signal is separated, and the classification module is executed. The classification module compares the abnormal signal with the failures detector set, and the matching is then verified. Thus, the abnormal signal is classified according to the detector class that the signal matches. This phase uses the partial matching criterion proposed in [22], and hence adopts a standard deviation of 3% in the detectors.

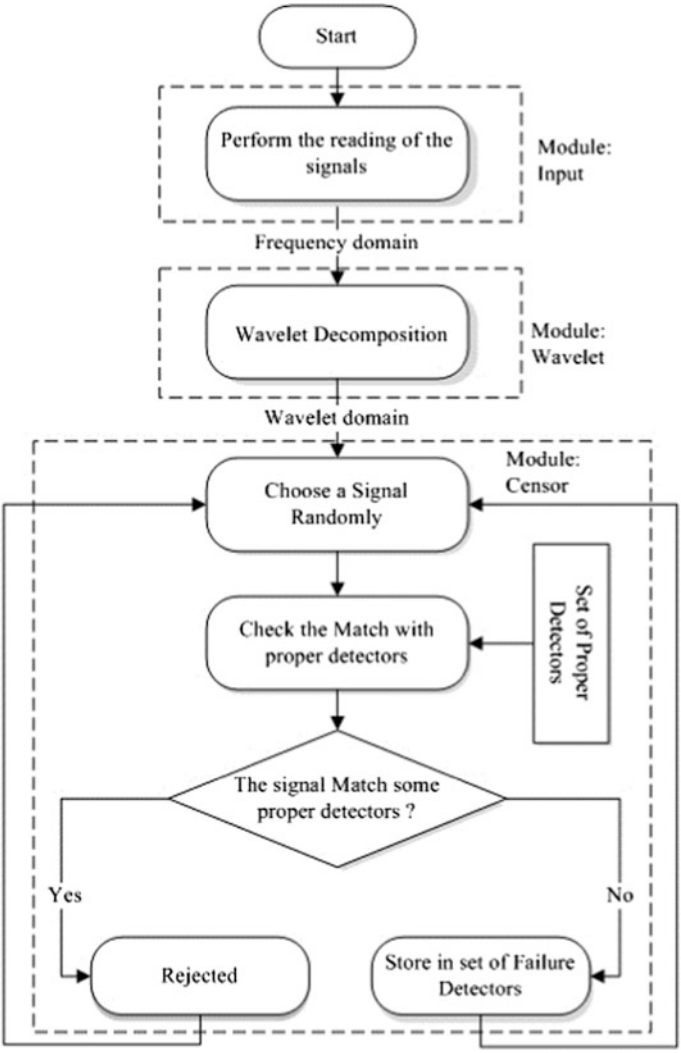


Fig. 4 Flowchart of the censor-phase

5.3 Wavelet Decomposition Module

The wavelet decomposition module is important to extract and emphasize the signal characteristics, which are easily detected in the wavelet world.

In this work, we have used four levels of decomposition for the DWT. This procedure was adopted aiming to allow the signals abnormalities representation more easily. Table 2 presents the frequency ranges for each level of resolution in the DWT.

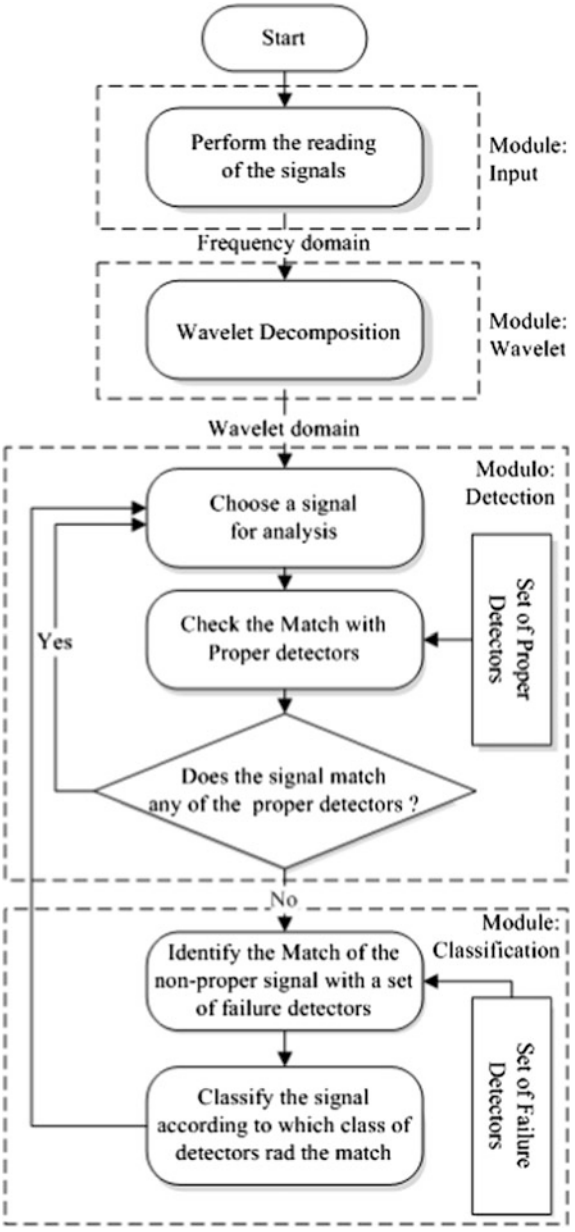


Fig. 5 Flowchart of the monitoring-phase

Figure 6 illustrates a signal with a normal condition and a signal with 15% of damage. These signals were presented at the input of the wavelet decomposition module, and after the signal processing, the results shown in Fig. 7.

These figures show the importance of wavelet decomposition for the diagnosis system. The failures are emphasized when the signal is decomposed into the

Table 2 Frequency ranges for each level of resolution in the DWT

Resolution level	Parameter	Frequency range (KHz)
1	D_1 component	7.68–3.84
2	D_2 component	3.84–1.92
3	D_3 component	1.92–0.96
4	D_4 component	0.96–0.48
4	C_4 component	0.00–0.48

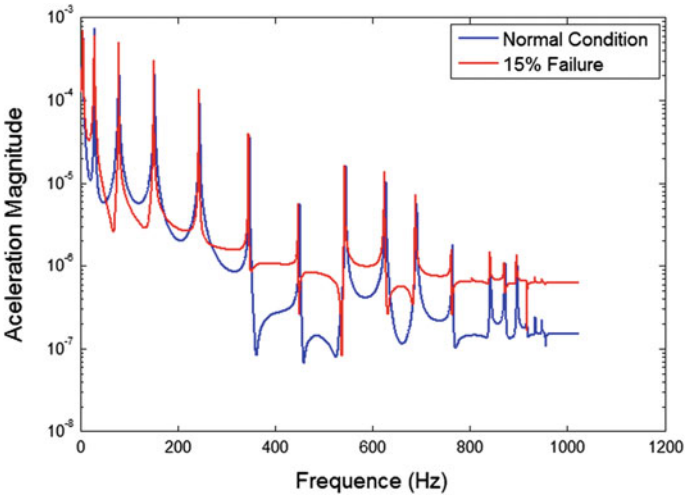


Fig. 6 Frequency domain signal

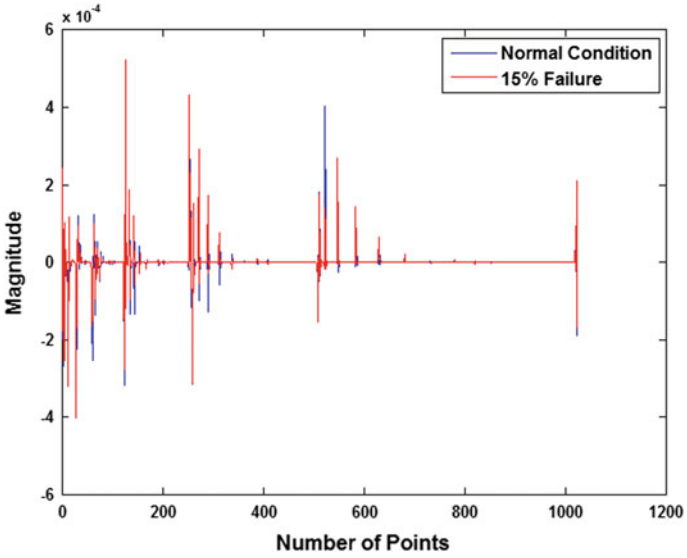


Fig. 7 Wavelet decomposition signal

wavelet world, and thus, the wavelet module contributes to the NSA. This is because its sensitivity when analyzing patterns which allows easy recognition of any abnormality.

6 Applications and Results

This section presents the results that were obtained with the proposed method implemented in a test database. The algorithm was developed in MATLAB® [28]. The proposed algorithm is applied to a database composed of signals in the frequency domain obtained from a numerical model of an aluminum beam, representing the wing of the aircraft.

6.1 Parameter Used in the Method

In the tests proposed in this work, an assessment of the proposed methodology was applied by checking the efficiency, accuracy and the computational time for different configurations of the set of detectors of the WAIS. Accordingly, three sets of detectors (CD_1, CD_2 and CD_3) have been generated using, respectively, 10%, 20% and 30% of the normal signal (baseline). For instance, a CD value 10% means that 50 signals were selected to be proper. Identical percentages were also considered to generate failure detectors. The parameters used for the tests are shown in Table 3.

6.2 Results

In order to evaluate the proposed methodology, tests were performed considering different settings of the WAIS. The results obtained are shown in Table 4, and represents the best configuration of the WAIS. The results presented in Table 4 represent the average values obtained by a cross-reference test that was performed 20 times while performing the WAIS for each set of detectors in order to guarantee the veracity of the results. The cross-reference test is a statistical test to analysis of the results.

It was observed that the WAIS has a good performance (with an accuracy rate equal to 100% for the best configuration as shown in Table 4), and that the quantity

Table 3 Parameters used in the tests

Parameters	Value
TAf	66.66%
Deviation (ε)	3%
CD_1	10% of the data
CD_2	20% of the data
CD_3	30% of the data

Table 4 Results of the tests

Analyzed signals	CD_1		CD_2		CD_3	
	Samples tests	Match correct	Samples tests	Match correct	Samples tests	Match correct
Normal condition (0%)	500	496	500	498	500	500
5%	150	146	150	148	150	150
10%	150	147	150	148	150	150
15%	150	149	150	149	150	150
20%	150	143	150	147	150	150
25%	*150*	*142*	*150*	*147*	*150*	*150*
30%	*150*	*146*	*150*	*148*	*150*	*150*
Accuracy (%)	97.78		98.92		100%	
Time (ms)	96.03		97.32		95.43	

Table 5 Comparative study

References	Data type	Technique used	Accuracy (%)
[10]	Experimental	Multi-objective Genetic Algorithm	93.70
[29]	Experimental	Multilayer Perceptron (Levenberg-Marquardt)	98.52
[7]	Simulated	Fuzzy Logic	98.74
[12]	Simulated	ARTMAP-Fuzzy	100.00
[14, 15]	Simulated	ARTMAP-Fuzzy-Wavelet	100.00
This work	Simulated	WAIS	100.00

of detectors used in censor-phase directly influences the failure diagnosis process. Thus, we recommend to use 30% of database information to generate the set of detectors to bring robustness to the system. That is, the more knowledge is available in the learning phase, the more efficient is the process of diagnosis of the WAIS.

Finally, we highlight that the WAIS was ran with a time of less than 100 ms, which provides the application of this system in real time, as decisions must be taken in time to prevent tragedies and disasters.

6.3 Comparative Study

In this section, we present a comparative study between the present methodology and the methodology proposed by other authors [7, 10, 12, 14, 15, 29]. For this comparison, the total accuracy of the methodologies for the detection and classification of structural failure has been taken into consideration.

Table 5 shows the comparison between the accuracy obtained by the proposed method and the main methods available in the literature.

In Table 5, we note that the proposed method had a very good success rate (matched 100%), when compared to other methods. It is important to emphasize that from the comparison of the results obtained, it is clear that the application of the proposed method in real problem will also bring good efficiency levels.

6.4 Positive and Negative Aspects of the Proposed Methodology

After performing all tests and getting the results of WAIS algorithm proposed in this work, we present an analysis highlighting the main positive and negative aspects of the proposed methodology.

- Positive Aspects:
  - Regarding the accuracy in diagnosing, the WAIS showed to have excellent performance;
  - The proposed WAIS runs with low processing time which allows this method to be applied in real situations as decision making should be taken instantly to avoid disasters;
  - WAIS is robust because using only 30% of the available information it was able to diagnose 100% of actual signals (high level of learning);
  - Compared with different neural networks, employing WAIS means that it is not necessary to execute the learning phase (training) every time monitoring runs.
- Negative Aspects:
  - WAIS has parameters that must be calibrated, especially in the wavelet module.

7 Conclusion

This work presented a new approach to detect and classify failures in aeronautical structures using WAIS algorithm. A finite element numerical model was used to simulate the failure signals, to generate a data set to be analyzed and test the methodology. The proposed algorithm presented good results, with 100% matching in detecting and classifying the failures tested. The detector generation phase was executed off-line with no bias for the algorithm. The monitoring-phase is quickly executed in a total time of less than 100 ms, which allows for it to be used in real time to aid the decision making process. The combination of the wavelet transform with the NSA (Negative Selection Algorithm) provides more precision to the diagnosis due to the high resolution level in decomposing signals, making it easy to

identify abnormalities. Thus, the proposed Wavelet Immune System Algorithm showed to be precise, robust, efficient and suitable in several applications, particularly in real systems as aircraft structures.

The authors believe that this work will contribute to the SHM research area introducing a new hybrid approach to perform the monitoring of aeronautical structures using intelligent techniques and wavelet transforms.

References

1. S.R. Hall, The effective management and use of structural health data, in *Proceedings of the International Workshop on Structural Health Monitoring* (1999), pp. 265–275
2. S. Zheng, X. Wang, L. Liu, Damage detection in composite materials based upon the computational mechanics and neural networks, in *Proceedings of the European Workshop on Structural Health Monitoring* (2004), pp. 609–615
3. V.R. Franco, D.D. Bueno, M.J. Brennan, A.A. Cavalini Jr., C.G. Gonsalez, V. Lopes Jr., Experimental damage location in smart structures using Lamb wave's approaches, in *Proceedings of the Brazilian Conference on Dynamics, Control and Their Application* (2009), pp. 1–4
4. M. Krawczuk, W. Ostachowicz, G. Kawiecki, Detection of delamination in cantilevered beams using soft computing methods, in *Proceedings of the Conference on System Identification and Structural Health Monitoring*, Madrid (2000), pp. 243–252
5. V. Giurgiutiu, Tuned lamb wave excitation and detection with piezoelectric wafer active sensors for structural health monitoring. J. Intell. Mater. Syst. Struct. **16**, 291–305 (2005)
6. L. Palaia, Structural Failure Analysis of timber floors and roofs in ancient buildings at Valencia (Spain), in *Proceedings of the International Conference on Mechanical Behavior and Failures of the Timber Structures* (2007), pp. 1–11
7. M. Chandrashekhar, R. Ganguli, Structural damage detection using modal curvature and fuzzy logic. Struct. Health Monit. **8**, 267–282 (2009)
8. X.J. Chen, Z.-F. Gao, Y.-E. Ma, Q. Guo, Application of wavelet analysis in vibration signal processing of bridge structure, in *Proceedings of the International Conference on Measuring Technology and Mechatronics Automation* (2010), pp. 671–674
9. T. Shen, F. Wan, B. Song, Y. Wu, Damage location and identification of the wing structure with probabilistic neural networks, in *Proceedings of the Prognostics and System Health Management Conference* (2011), pp. 1–6
10. F.L. Wang, T.H. Chan, D.P. Thambiratnam, A.C. Tan, Damage diagnosis for complex steel truss bridges using multi-layer genetic algorithm. J. Civil Struct. Health Monit. **3**(2), 117–217 (2013)
11. B.I. Song, H. Seze, K.A. Giriunas, Collapse performance evaluation of steel building after loss of columns, in *Proceedings of the Structures Congress* (2012), pp. 213–224
12. A.S. Souza, F.R. Chavarette, F. Lima, M. Lopes, S.S.F. Souza, Analysis of structural integrity using an ARTMAP-Fuzzy Artificial Neural Network. Adv. Mater. Res. **838–841**, 3287–3290 (2013)
13. F.P.A. Lima, F.R. Chavarette, A.S. Souza, S.S.F. Souza, M. Lopes, Artificial immune systems with negative selection applied to health monitoring of aeronautical structures. Adv. Mater. Res. **871**, 283–289 (2013)
14. F.P.A. Lima, F.R. Chavarette, S.S.F. Souza, M. Lopes, A.E. Turra, V. Lopes Jr., Analysis of structural integrity of a building using an artificial neural network ARTMAP-Fuzzy-Wavelet. Adv. Mater. Res. **1025–1026**, 1113–1118 (2014)

15. F.P.A. Lima, F.R. Chavarette, S.S.F. Souza, M. Lopes, A.E. Turra, V. Lopes Jr., Monitoring and fault identification in aeronautical structures using an ARTMAP-Fuzzy-Wavelet Artificial Neural Network. Adv. Mater. Res. **1025–1026**, 1107–1112 (2014)
16. C.C.E. Abreu, F.R. Chavarette, F.V. Alvarado, M. Duarte, F. Lima, Dual-Tree complex wavelet transform applied to fault monitoring and identification in aeronautical structures. Int. J. Pure Appl. Math. **97**, 89–97 (2014)
17. F.P.A. Lima, A.D.P. Lotufo, C.R. Minussi, Disturbance detection for optimal database storage in electrical distribution systems using artificial immune systems with negative selection. Electr. Power Syst. Res. **109**, 54–62 (2014)
18. S. Forrest, A. Perelson, L. Allen, R. Cherukuri, Self-nonself discrimination in a computer, in *Proceedings of IEEE Symposium on Research in Security and Privacy* (1994), pp. 202–212
19. L.N. Castro, J. Timmis, *Artificial Immune Systems: A New Computational Intelligence Approach* (Springer, 2002)
20. L.N. Castro, Immune engineering: development and application of computational tools inspired by artificial immune systems. Ph.D. thesis, UNICAMP, 2001 (in Portuguese)
21. D. Dasgupta, *Artificial Immune Systems and Their Applications* (Springer, 1998)
22. D.W. Bradley, A.M. Tyrrell, Immunotronics—novel finite-state-machine architectures with built-in self-test using self-nonself differentiation. IEEE Trans. Evol. Comput. **6**, 227–238 (2002)
23. F.P.A. Lima, C.R. Minussi, R.B. Bessa, J.N. Fidalgo, A modified negative selection algorithm applied in the diagnosis of voltage disturbances in distribution electrical systems, in *Proceedings of 18th International Conference on Intelligent System Application to Power Systems* (2015), pp. 1–6
24. S. Mallat, *A Wavelet Tour of Signal Processing*, 2 edn. (Academic Press, New York, 1999), 637 pp.
25. I. Daubechies, *Ten Lectures on Wavelets* (Society for Industrial and Applied Mathematics, 1992)
26. F.P.A. Lima, F.R. Chavarette, S.S.F. Souza, A.S. Souza, M. Lopes, Artificial immune systems applied to the analysis of structural integrity of a building. Appl. Mech. Mater. **472**, 544–549 (2014)
27. F.P.A. Lima, A.D.P. Lotufo, C.R. Minussi, Wavelet-artificial immune system algorithm applied to voltage disturbance diagnosis in electrical distribution systems. IET Gener. Transm. Distrib. **9**, 1104–1111 (2015)
28. MATLAB 7.8 version, MathWorks Company
29. L. Roseiro, U. Ramos, R. Leal, Neural networks in damage detection of composite laminated plates, in *Proceedings of the 6th International Conference on Neural Networks* (2005), pp. 115–119

An Illustration of Some Methods to Detect Faults in Geared Systems Using a Simple Model of Two Meshed Gears

Fabrício Cesar Lobato de Almeida, Aparecido Carlos Gonçalves, Michael John Brennan, Amarildo T. Paschoalini, A. Arato Junior and Erickson F.M. Silva

Abstract Gears are the components in many mechanical systems that are likely to develop faults due to their dynamic characteristics, such as the cyclic loading applied to the meshing teeth. The main faults in gears are pitting and scuffing, where the tooth profile (involute) is heavily affected, and hence, signal processing techniques have been developed to aid in the detection of gear faults in their early stages. It is already known that the dynamic behaviour of a mechanical system changes when its characteristics are affected (such as in the presence of a fault), and as a result, the vibration of such a system can be used to detect a fault in its early stage. To investigate and develop techniques based on vibration analysis, a physical understanding of the system involving meshing gears is required. In this chapter, a model is introduced that can be used for simulating vibration data of toothed meshing gears. The data generated by the simulations is then used to investigate some classic techniques used in gear fault detection problems.

Keywords Detection · Gears · Fault · Simulation

F.C.L. de Almeida (✉)
Department of Biosystems Engineering, São Paulo State University (UNESP), Tupã, SP, Brazil
e-mail: fabricio@tupa.unesp.br

A.C. Gonçalves · M.J. Brennan · A.T. Paschoalini · A. Arato Junior
Department of Mechanical Engineering, São Paulo State University (UNESP), Ilha Solteira, SP, Brazil

E.F.M. Silva
Department of Exact Science and Engineering, State University of Santa Cruz (UESC), Ilhéus, BA, Brazil

S. Ekwaro-Osire et al. (eds.), *Probabilistic Prognostics and Health Management of Energy Systems*, DOI 10.1007/978-3-319-55852-3_14

1 Introduction

Meshing gears are ubiquitous in industry, and are being applied to many mechanical systems to transmit power. Additionally, gears are used to smooth the angular velocity transmission from one gear to another. This characteristic, however, is only possible due to the involute tooth form which leads to the fundamental law of gearing where the angular velocity ratio between two meshing gears remains constant during the mesh [1]. It is conventional to refer to the smaller gear in the gear set (two gears) as the pinion and the other as the gear. Because of their widespread use and importance, there is a need to monitor the health of gearboxes to detect incipient faults as gears are the parts more likely to present faults in gearboxes due to their dynamic characteristics (cyclic loads applied to the teeth). The main faults are due to contact stress fatigue (known as pitting) and damage generated by wear due to sliding gear contact (which is called scuffing) [2]. These defects change the tooth profile so that the contact between the meshing gears is affected causing non-uniform gear rate, reduced efficiency, increased dynamic effects, and may lead to severe tooth failure [3].

Signal processing techniques have been developed to aid in predicting gear faults in their early stages. As the dynamic behaviour of a mechanical system changes due to the presence of a fault, the vibration of the system can be used to detect faults in their early stages. However, to investigate and develop such techniques, we need to use an analytical model of meshing gears so that controlled conditions can be obtained using this model and physical insight can also be achieved. The aim of this chapter is to show how the complicated problem of meshing gears can be turned into the simple model of a one degree of freedom system which can then be used to investigate techniques to detect faults in gears. Moreover, techniques to detect faults in gears in the time domain (time synchronous averaging), frequency domain (discrete Fourier transform) and time-frequency domain (wavelet transform) will be carried out using simulated data.

2 Modeling the Dynamic Response of Toothed Gears Pair

The model used in this work is the one developed by Harris [4] as this model is well known in the literature and has been used for many years. Furthermore, this model describes the meshing forces for a pair of gears, and it is these forces, together with the variation of the meshing stiffness (time varying stiffness during tooth-mesh) that are considered to be the source of gear vibration [5]. This time varying stiffness can be modelled in many ways as shown in [6], but in this work, however, the varying stiffness is assumed to be a cosine function varying around its mean value.

2.1 The Equivalent One Degree-of-Freedom System for Toothed Geared Systems

In this model it is considered that the gears are connected to a rigid support so that the tooth-mesh is the only parameter responsible for the vibrating behaviour of the mechanical system. Figure 1 shows a schematic of the dynamics of a pair of meshing gears. Figure 1a shows the pair of meshing gears considering a rigid support (pinned). Figure 1b shows the free-body diagram of meshing gears which is used to derive the fundamental equations that the dynamic model is based on. The equations of motion are given by

$$T_1 - Pr_1 \cos\varphi - T_1{}' = I_1\ddot{\theta}_1, \tag{1}$$

$$Pr_2 \cos\varphi + T_2{}' - T_2 = I_2\ddot{\theta}_2, \tag{2}$$

where P is the contact force between meshing gears, r is the pitch circle radius, T is the torque, $T^{'}$ is torque due to the friction, φ is the pressure angle and I is the polar moment of inertia. The subscripts 1 and 2 denote the driving and driven gears respectively.

Assuming that the torques are a function of the gear angular displacement θ (i.e. $T_1(\theta)$ and $T_2(\theta)$), so that these torques can be expressed as the product of a constant force $\bar{P}_o$ along the path of contact and the gear radius, then $T_1 = \bar{P}_o r_1 \cos\theta$ and $T_2 = \bar{P}_o r_2 \cos\theta$. Considering that the relative displacement between the driving and driven gears is expressed as $x = (r_1\theta_1 - r_2\theta_2)\cos\theta$ and combining this with Eq. (1) and Eq. (2) results in

$$\ddot{x} + c'\dot{x} + \sigma F'(x,\theta) = \sigma F, \tag{3}$$

in which $\sigma = \left(r_1^2/I_1 + r_2^2/I_2\right)$ is a constant, $c'\dot{x} = (T_1{}'r_1/I_1 + T_2{}'r_2/I_2)\cos\varphi$ is the damping force, $F'(x,\theta) = P\cos^2\phi$ is a time-varying force, and $F = \bar{P}_0\cos^2\phi$.

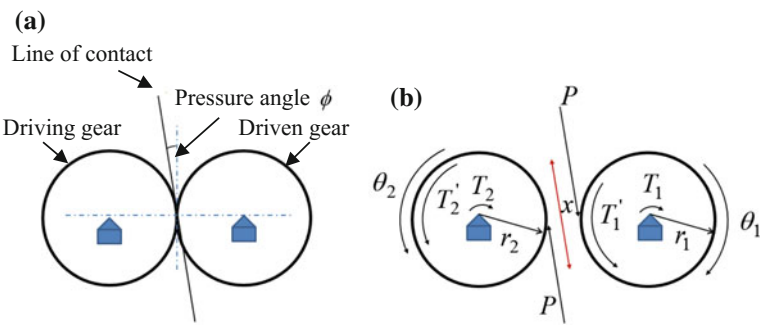


Fig. 1 Schematic of meshing gears. **a** Pair of meshing gears and the nomenclature used in this work. **b** Free-body diagram of meshing gears

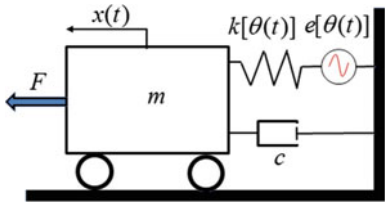


Fig. 2 The schematic of a linear oscillator given by Eq. (6)

According to the work carried out by Harris [4], the contact force between meshing teeth can be expressed as

$$P(x,\theta) = K(1 + a\cos\theta)(x - x_0\cos\theta), \tag{4}$$

where K is a constant stiffness, a is the maximum stiffness, and x_0 is the maximum deviation from the ideal meshing surface (involute). Substituting Eq. (4) into Eq. (3) gives

$$\ddot{x} + c'\dot{x} + \sigma k(\theta)x = \sigma F + \sigma k(\theta)e(\theta), \tag{5}$$

where $e'(\theta) = x_0\cos\theta$ is a spatially periodic displacement excitation function and $K(1 + a\cos\theta)$ is a spatially periodic stiffness function. Multiplying Eq. (5) by the equivalent mass, which is $m = 1/\sigma$, then Eq. (5) becomes

$$m\ddot{x} + c\dot{x} + k(\theta)x = F + k(\theta)e(\theta), \tag{6}$$

in which $m\ddot{x}$ is the inertial force, $c\dot{x}$ is the damping force, $k(\theta)x$ comes from the variation of meshing stiffness, $c = c'/\sigma$, F is a constant load, and $k(\theta)e(\theta)$ comes from the displacement excitation. This equation describes a linear oscillator, which is shown in a schematic way in Fig. 2. Moreover, Eq. (6) shows that the two degrees of freedom system was simplified to a one degree of freedom system.

3 Time Histories Generated by the Dynamic Simulator of Toothed Gear Pair

A simulation is carried out using the linear oscillator given by Eq. (6). Solutions of Eq. (6) are calculated numerically, and although the simulation is based on a theoretical model, some data was collected from the work developed in [7], such as the damping ratio. The parameters used in the simulations are as follows:

- Gear: 15 teeth, pitch radius 30 mm. Pinion: 13 teeth, pitch radius 14 mm.
- Angular speed of the driven gear: 385 RPM. Angular speed of the driving gear (pinion): 833 RPM.

- Equivalent mass 0.5 kg. Damping ratio 0.0625. Load force (F) 1.56 kN. Meshing Stiffness $k(\theta) = 22 \times 10^6(1 + 0.1\cos\theta)$ N/m.

Figure 3 shows the results obtained using the model given by Eq. (6) considering two cases. The labels 'i' and 'ii' are for cases where there is no fault and when a fault is introduced by reducing the time-varying stiffness in the meshing gears, respectively. Figure 1a show the schematic of the 7 pinion teeth used in the analysis. Figure 3b shows the time-varying stiffness. Figure 3c shows the time series (acceleration). As observed, the presence of a fault (simulated by reducing the stiffness of one tooth) affects the simulated time series. Hence, such a model can be used to investigate signal processing techniques for gear faults as it is representative of the system dynamics.

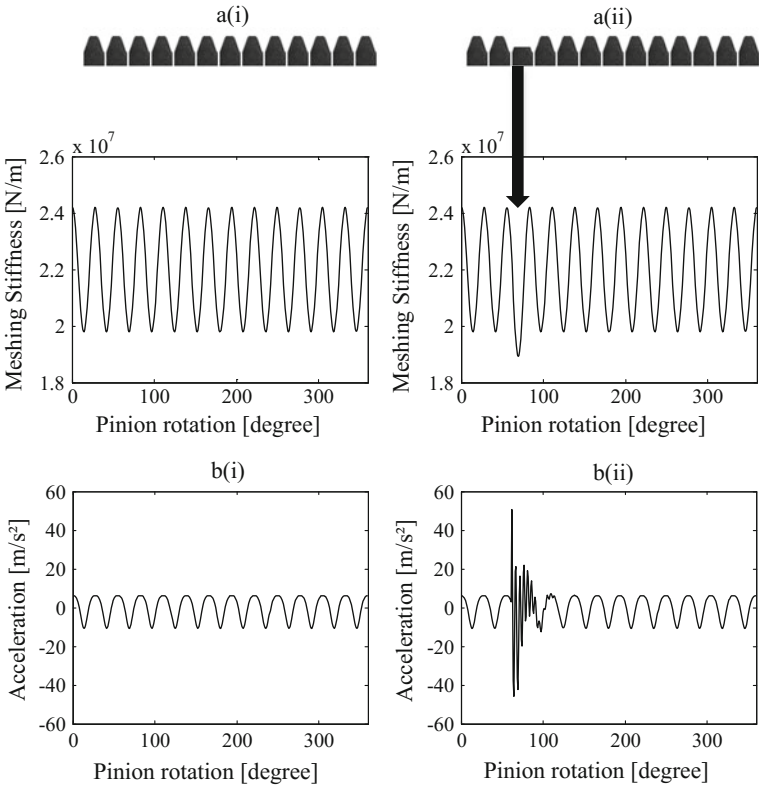


Fig. 3 Meshing stiffness and acceleration of meshing gears using the toothed model. The labels "i" and "ii" stand for a healthy and a damaged gear, respectively. **a** Meshing stiffness. **b** Acceleration

4 The Effect of Uncorrelated Noise on the Time Histories of Simulated Data and Time Synchronous Averaging (TSA) Technique

Uncorrelated noise due to external sources is present in any actual vibration data so techniques should be used to attenuate its effect prior to analysis. One of these techniques is synchronous averaging or time synchronous averaging (TSA) or time domain average [8, 9]. The technique generally used in gear fault detection problems. To obtain a synchronous sample of the vibration signal from meshing gears, the average between these vibration signals collected over many gear revolutions (or cycles) (in which any block (sample length) starts at the same angular position) is required. Figure 4 shows a schematic highlighting how this technique (averaging) is conducted. Every gear cycle (block) q has the same period, T_c, so that the signal x can be averaged Q times generating an averaged signal y.

The result of this technique is a synchronous sample that is then dominated by synchronous components (such as the gear meshing response). Moreover, the uncorrelated noise is also attenuated in the time histories (since such noise is random in each block) so that its average tends to zero. In a practical situation, however, a device (such as a trigger like a tachometer) should be used to detect the angular position (reference) from which the averaging will be carried out. Figure 5a shows an infrared optical sensor used as a trigger. Figure 5b shows the optical sensor mounted in an experimental test rig.

This sensor gives a pulse when the light beam generated by it crosses a reflecting tape (angular position) glued, in this case, to the input shaft. Figure 6a, b shows the time history of the trigger and the time history of an accelerometer mounted on the gear box, respectively, to highlight the use of such a device in the calculation of TSA.

One way of calculating the TSA (shown in Fig. 4) is by conducting the recursive average, and this average is conducted by calculating a weighted residual for each block, which is added to the TSA. The mathematical equation for this average is given by

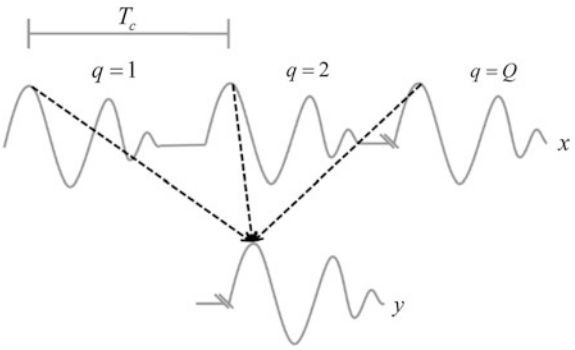


Fig. 4 The schematic of how the time synchronous averaging is conducted

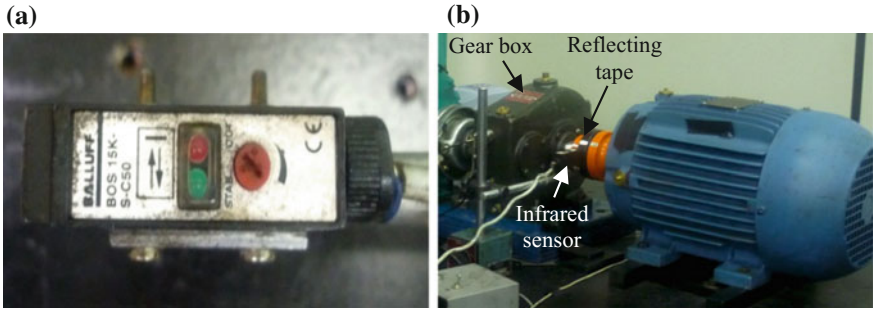


Fig. 5 Experimental test rig **a** the infrared sensor (trigger), and **b** the test rig

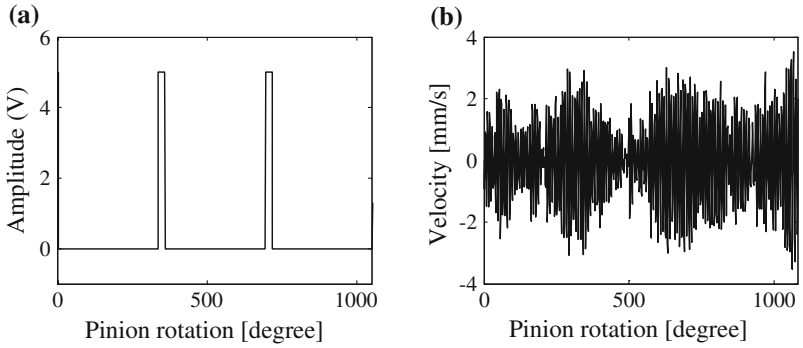


Fig. 6 The time history of: **a** Infrared optical sensor (trigger), **b** Vibration signal (velocity) measured on the gear box shown in Fig. 5

$$y_q = y_{q-1} + \frac{x_q - y_{q-1}}{Q}, \tag{7}$$

where y is the time synchronous averaging, x is the actual block of the signal and n is the block number.

Returning to the simulated data, it has been seen that noise is present in actual vibration signals. Hence, to check how such noise affects the time histories and also how TSA works, uncorrelated noise will be added to the simulated data given in Fig. 3. The signal-to-noise ratio (SNR) of −6.5 dB is adopted in this simulation. This is achieved by adding Gaussian noise to the signal. Figure 7a(i), b(i) show the simulated data without and with a defect, respectively, for one revolution of the pinion with no time averaging. Figure 7a(ii), b(ii) show the simulated data without and with a defect, respectively, after applying TSA. It can be seen that this technique is effective in reducing the uncorrelated noise added to the signal. Furthermore, TSA can also be used as a pre-processing tool before conducting other fault

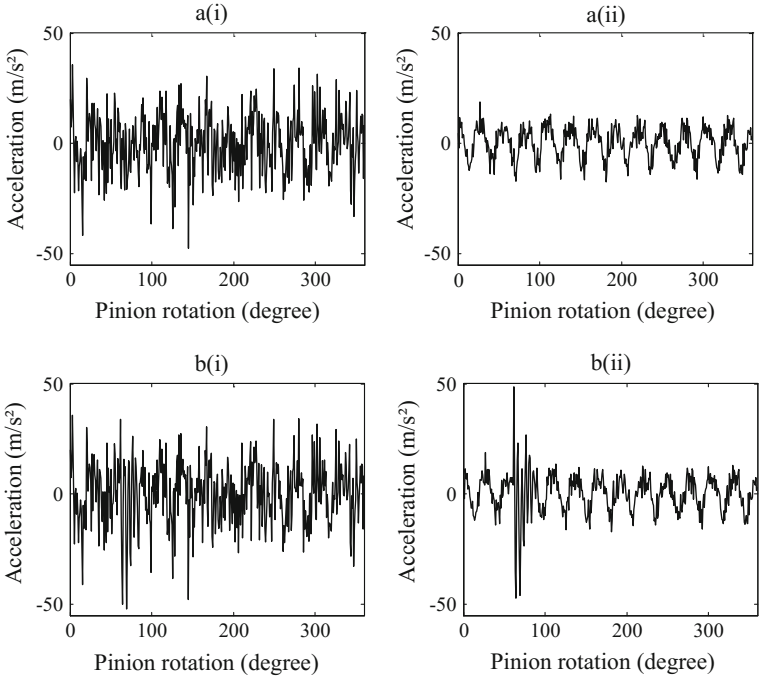


Fig. 7 Time histories of simulated data generated by the toothed meshing gear model. The labels "i" and "ii" stand for cases without and with the application of the time synchronous average technique, respectively, **a** Gear without fault, **b** Gear with fault

detection methods (such as the discrete Fourier transform and the wavelet transform which are presented later in this chapter).

5 Statistical Analysis in the Time Domain

Statistical moments [10, 11] and their normalizations are included in this group as these moments are commonly used to describe the shape of probability density functions (PDF) so that they are able to detect any change in its original form. In gears, for example, if the interactions between meshing gears (tooth surface interaction) are in good conditions, then the PDF has a particular shape which is similar to a bell. However, if a fault and/or a severe wear cause a change in the load condition over the tooth surfaces, then the PDF shape will also change so that such change can be detected via statistical moments [10].

The mean value of a sample is the first moment and the variance is the second one. The RMS (Root-mean-square) value, which is very often used in gear fault detection problems, is the square root of the variance (second moment). So, the

RMS value is a variation (normalization) of the second moment. In addition to those, the third moment (which is called the skewness), the fourth moment (which is called kurtosis), and the crest factor (which is the ratio between the peak in the time histories) and the RMS value are all used in fault detection in gears. These moments and their normalizations are described in this section.

The first moment (or mean value) is a measurement of the central tendency of a set of numbers characterized by a probability distribution function [12]. The mean shows where the scatter of points of a sample is centred so that it can also be interpreted as a measure of location [13]. For a discrete signal, the mean is given by

$$\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n, \tag{8}$$

where x is the discrete time history and N is the total sample elements.

The root-mean-square (RMS) value is a measurement of the amount of energy contained in the vibration time histories (vibration signature). Although this tool is valuable for measuring vibration level, it does not provide any evidence that a fault is occurring in the mechanical system (such as a severe tooth wear in a gearbox). The RMS equation for a discrete signal is given by

$$\text{RMS} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} x_n^2}. \tag{9}$$

This tool can also be used to aid the maintenance management as vibration severity (and given by its calculation) can be used as an indication of the machine health condition. There are charts which show that a machine can operate without any severe damage (e.g., the ISO 2372 which is defined according to the machine horsepower and operating angular speed, and gives a threshold value, and, in this case, an RMS value). In this situation, the vibration measurement is velocity given in mm/s.

The crest factor (also called peak-to-RMS-ratio) is given by the ratio of the highest peak in the vibration signal to its RMS value. This technique is good for transient signals, such as a broken gear tooth, where the RMS value is not too sensitive to such transient change. Hence, the presence of peaks in the vibration signature will increase the crest factor value, and indicate the presence of a fault occurring in the mechanical system. For a normal operation, the crest factor should be between 2 and 6 [14], and is a non-dimensional number. The crest factor can be calculated by

$$CF = \frac{x_{peak}}{\text{RMS}}, \tag{10}$$

where x_{peak} is the highest peak in the vibration signature.

The third moment about the mean (or the so-called skewness) is a measure of the asymmetry of a probability density function or any random variable associate with it. It is given by

$$S = \frac{1}{N} \frac{\sum_{n=1}^{N} (x_n - \bar{x})^3}{(\sigma^2)^3}, \tag{11}$$

where σ^2 is the variance of the time histories.

The fourth moment about the mean (or kurtosis) is a measure of the relative "peakedness" or "flatness" of probability distribution function compared to the normal (Gaussian) distribution [14]. This moment is used as an indication of the degradation of a mechanical system, such as gears, and an increment in the kurtosis indicates an increment in the crest factor. Kurtosis can be evaluated by using the following mathematical equation

$$K = \frac{1}{N} \frac{\sum_{n=1}^{N} (x_n - \bar{x})^4}{(\sigma^2)^4}. \tag{12}$$

To investigate how these measures work, the simulated data shown in Figs. 3 and 7 are used to evaluate 4 out of the 5 measures mentioned previously. Table 1 shows the result for six simulated cases where the RMS, crest factor, skewness and kurtosis were evaluated. It is observed that the RMS value and crest factor are more sensitive to noise. Moreover, when TSA is conducted, the effect of noise is drastically reduced as already shown in Fig. 7. Furthermore, the values calculated for the two measurements with noise attenuated by the use of TSA are close to the ones calculated using the ideal case without noise.

6 Frequency Domain Analysis: The Discrete Fourier Transform (DFT)

The severity of vibration or vibration level, as mentioned in the previous section, is a valuable tool to detect critical vibration conditions such as severe wear in gears. Although this technique provides a good indication when there is a fault developing in a mechanical system (which is given by the increment of the RMS value, for example), such a technique does not allow identification of the source of the fault.

It is known that a mechanical system has many sources of excitation from its mechanical parts (such as gears, bearings, shafts, among others). The time history of such a system is given by the summation of the vibrations of each component so

Table 1 Evaluation of the RMS, crest factor, skewness and kurtosis for simulated data considering cases with and without gear faults, noise and TSA

Simulated cases		RMS (m/s^2)	Crest factor	Skewness	Kurtosis
Signal without noise	Gear without fault	5.95	1.06	−0.54	1.78
	Signal shown in Fig. 3a				
	Gear with fault	8.06	6.27	−0.7	11.1
	Signal shown in Fig. 3b				
Signal with noise	Gear without fault and no TSA	13.4	2.68	−0.2	3.12
	Signal shown in Fig. 7a(i)				
	Gear without fault with TSA	7	2.68	−0.33	2.38
	Signal shown in Fig. 7a(ii)				
	Gear with fault and no TSA	14.3	2.5	−0.4	3.5
	Signal shown in Fig. 7b(i)				
	Gear with fault and with TSA	8.8	5.53	−0.6	8.33
	Signal shown in Fig. 7b(ii)				

that the time histories are complicated to analyse, and the vibration information from each mechanical part is masked by the global response of the system. Hence, a technique that overcomes such problem is needed. This can be achieved in the frequency domain by analysing the frequency components from each mechanical part. Figure 8 shows a schematic highlighting vibration sources from different mechanical parts of a mechanical system and their responses in the time and frequency domains. In this specific case, the vibration sources are periodic so that each component has its own amplitude and phase.

As observed, the signal in the time domain (time histories) does not give clear information about the vibration sources whereas the components from each vibration source can be clearly seen in the frequency domain. One way of calculating the signal in the frequency domain is via the Fourier transform. In rotating mechanical systems (such as gear vibration), this technique is fundamental for vibration analysis. The Fourier transform $X(f)$ of a vibration signal $x(t)$ is given by

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt, \tag{13}$$

where f is frequency in Hz and $j = \sqrt{-1}$. However, analog-to-digital converter (ADC) systems are used in actual vibration problems in signal acquisition systems, which means that the continuous vibration signal $x(t)$ is then digitalized at a sampling frequency f_s to its discrete form, so that

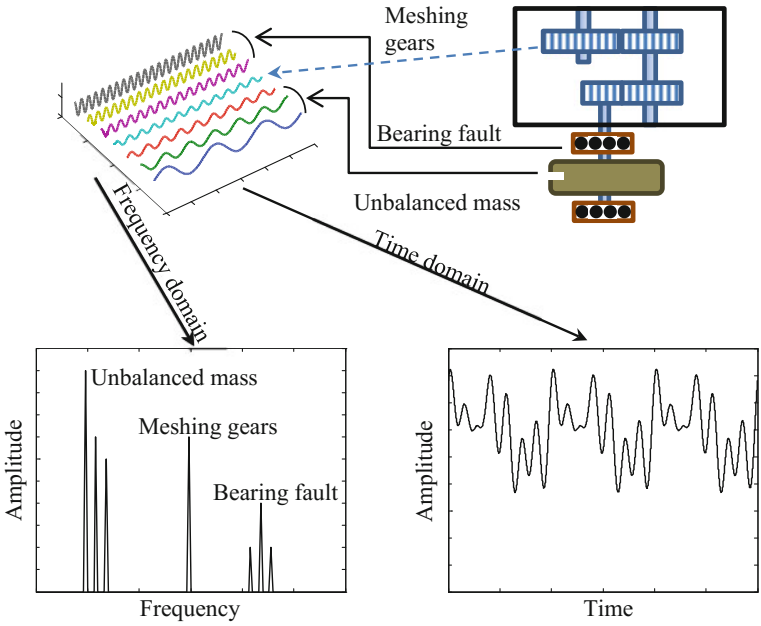


Fig. 8 Schematic of different sources of vibration excitation present in a mechanical system together with its representation in frequency and time domain

$$x(n\Delta) = x(t) \sum_{n=-\infty}^{\infty} \delta(t - n\Delta), \tag{14}$$

where δ is the delta function and $\Delta = 1/f_s$ is the time resolution. Hence, the discrete form of Eq. (13) for a sampled signal $x(n\Delta)$ is then given by

$$X(e^{j2\pi f\Delta}) = \sum_{n=-\infty}^{\infty} x(n\Delta)e^{-j2\pi fn\Delta}. \tag{15}$$

However $f = w/(N\Delta)$, where w is an integer. Moreover, in practical situations the length of data acquisition is finite, hence Eq. (15) can be rewritten as

$$X(w) = \sum_{n=1}^{N} x(n\Delta)e^{-j(2\pi/N)nk}. \tag{16}$$

Equation (16) is the Discrete Fourier Transform (DFT). This equation shows that the sampled data for a finite length in the time domain generates a discrete spectrum equally sampled in the frequency domain (which is an approximation of the Fourier series) [15]. As the data is finite, then the DFT will be distorted by data truncation

(or "windowing effect"), which is known as leakage. One way of reducing leakage problems is by multiplying the truncated data by another window (such as the Hamming window) so that these distortions can be attenuated.

For meshing gears, it is expected that frequencies at the meshing frequency (which is the angular speed of the gear multiplied to the number of gear teeth) and its multiples are present in the spectrum. Figure 9 shows cases where the DFT has been calculated. Figure 9a shows the case where no additive Gaussian noise is

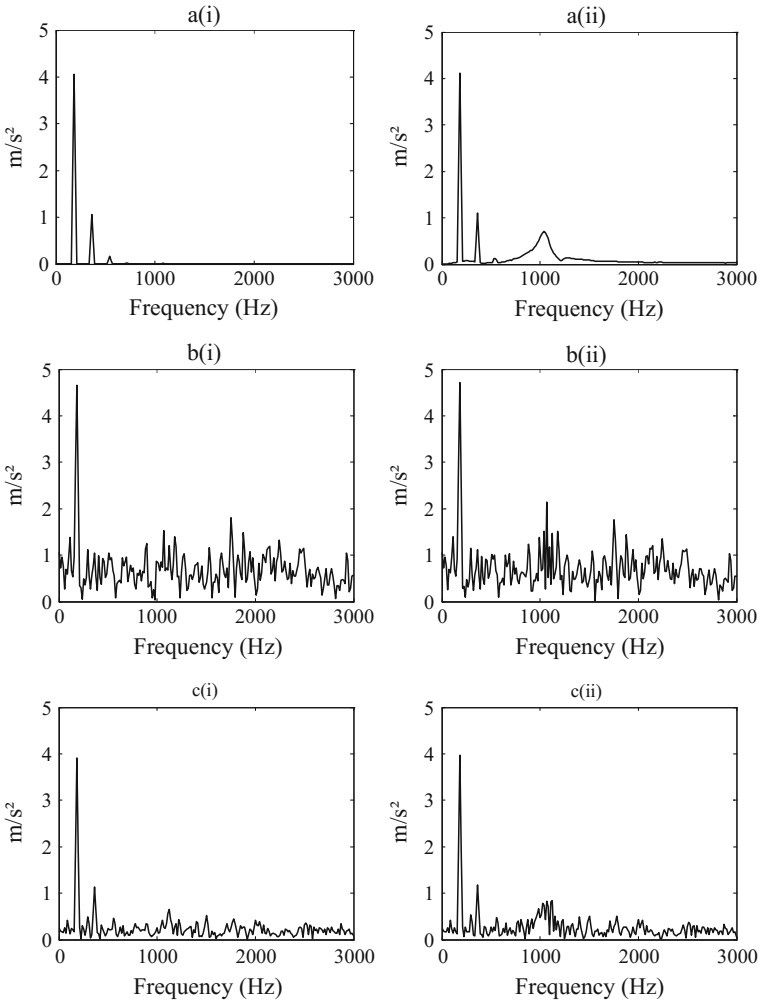


Fig. 9 The discrete Fourier Transform of the simulated signal generated by the toothed gear model. The labels "i" and 'ii' stand for cases of health and damaged gear, respectively, **a** Signals without any additive Gaussian noise present, **b** Signals with additive Gaussian noise and no use of time synchronous average technique, and **c** Signals with additive Gaussian noise smoothed by the use of time synchronous average technique

present in the data which is also shown in Fig. 7. The labels "i" and "ii" stand for healthy and damaged gears, respectively. It can be observed that for the damaged gear, there is a frequency band around 1000 Hz which is not observed for the healthy gear. However, when Gaussian noise is added to the data, such noise masks the main features present in the DFT as shown in Fig. 9b(i), b(ii). Figure 9c(i), c(ii) show the use of the TSA to attenuate external noise, and thus, the TSA technique is effective in enhancing the signal, therefore, highlighting the main features present in the DFT.

7 Time-Frequency Domain Analysis: The Wavelet Transform

Wavelets (or little waves) are functions that contain information in both the time and the frequency domains. They have one advantage when compared to the Fourier transform which deals only with the frequency domain where the temporal information is not available. The wavelet transform uses a scaled and shifted version of a basis function $h_w(t)$, which is also called mother wavelet, together with the signal $x(t)$ to compose the inner product that evaluates a decomposition of the signal into a weighted set of scaled waves [16]. The continuous wavelet transform is defined as [17]

$$CWT(b,a) = \frac{1}{\sqrt{a}} \int x(t) h_w^* \left(\frac{t-b}{a} \right) dt, \tag{17}$$

where b and a are the translation (shifting) and scaling coefficients, respectively. It is observed that Eq. (17) is similar to that the Fourier transform. However, the basis function for the Fourier transform is $e^{-j2\pi ft}$, which means that the signal $x(t)$ can be decomposed in sine and cosine functions. For the wavelet, however, this is a time-scale distribution, where the time-frequency analysis can be performed by establishing a relationship between the scale coefficient a and the frequency f. For signal analysis, the wavelet transform is suitable for non-stationary signals, such as vibration signals from damaged gears. To isolate signal discontinuities, it is desirable to have short basis functions, but at the same time, in order to obtain a detailed frequency analysis, it is desirable to have long basis functions. The wavelet transform provides good time resolution when analyzing high-frequency components, and provides good frequency resolution when analyzing low-frequency components because this transform is limited by the Heisenberg's Uncertainty Principle where the BT product remains constant. Figure 10 shows the time-frequency resolution plane highlighting that low-frequency components (high scale) have high-frequency resolution, and high-frequency components (low scale) have high time resolution.

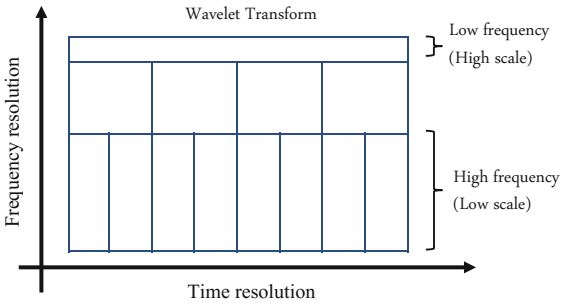


Fig. 10 The frequency-time plane showing how the resolutions in these domains are related

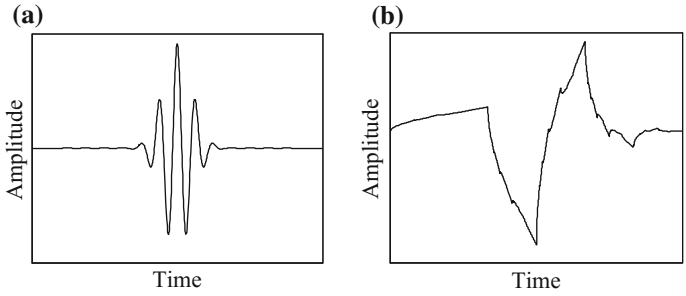


Fig. 11 The wavelets **a** Morlet and **b** Symlet

Although there are many basis function (wavelets) available in the literature, two of them are well-known: the Morlet wavelet and the Symlet wavelet. Figure 11a, b show the Morlet wavelet and the Symlet wavelet of order 2, respectively.

These wavelets are used to investigate how the wavelet transform can be used to detect faults in gears. Figure 12a, b show the wavelet transform performed using Morlet wavelet for simulated data of a healthy gear and a damaged gear, respectively. The frequency and time domain signals are also shown for convenience. Thus, it can be observed that the meshing gear frequency is clearly seen in this wavelet transform. Additionally, for the case with damage present in the system, it is observed that the wavelet transform can show when the damage occurs in time so that it is possible to know which tooth is damaged. Figure 13a, b show the wavelet transform performed using Symlet wavelet for simulated data of a healthy and a damaged gear, respectively. The frequency and time domain signals are also shown for convenience. It can be observed that the meshing gear frequency is also clearly seen in this wavelet transform, and it is even clearer to see where the damage is in this particular wavelet than when using the Morlet wavelet.

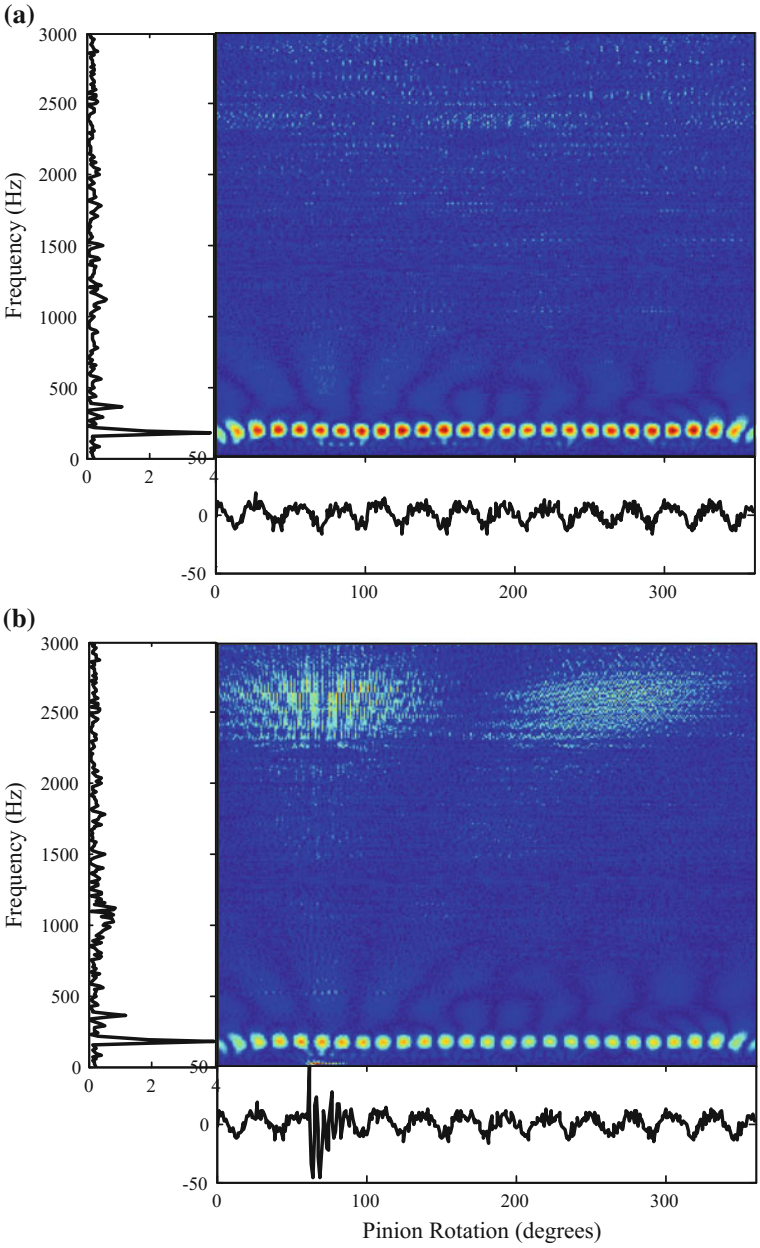


Fig. 12 Wavelet transform evaluated using simulated data and Morlet wavelet. The time and frequency domain histories are shown for convenience, **a** Gear in a good condition, and **b** Gear with a fault at tooth 3

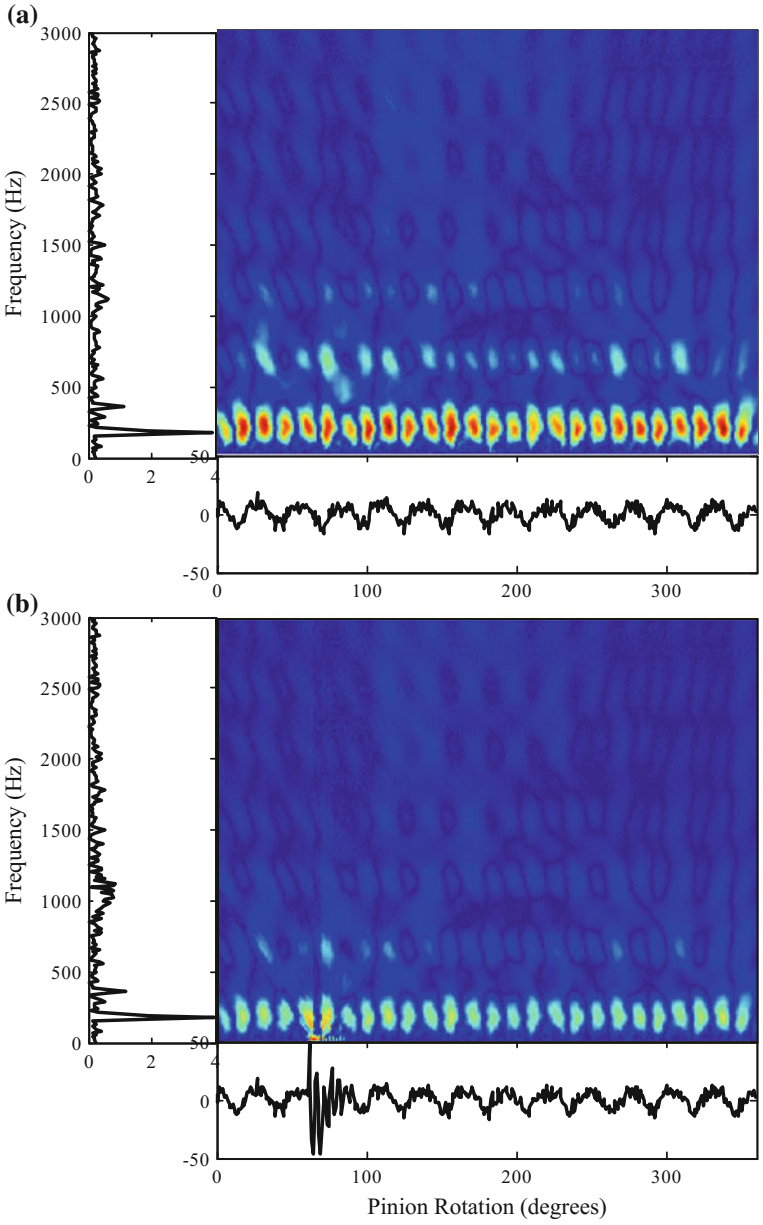


Fig. 13 Wavelet transform evaluated using simulated data and Symlet wavelet, the time and frequency domain histories are shown for convenience, **a** Gear in good condition, and **b** Gear with a fault at tooth 3

8 Conclusions

In this chapter a vibration simulator for toothed gearing systems has been described. This simulator was used to reproduce vibration signals considering gears without a fault or with a fault by reducing the mesh stiffness to simulate a broken tooth. Moreover, these signals were used to illustrate how classical signal processing techniques can be applied to detect fault in gears. These techniques can be carried out in the time domain, frequency domain or in the time-frequency domain. In the time domain the Time Synchronous Average (TSA) and statistical analysis methods were evaluated. The first (TSA) is very helpful in reducing the effects of uncorrelated noise present in the data, so that it can be used as a pre-processor prior to evaluating a signal processing method to detect faults. The statistical analysis gives an indication of the healthy condition of the mechanical system, but does not indicate which fault is in the system. This can be achieved by analysing the data in the frequency domain using the Fourier transform or by using time-frequency domain techniques, such as the Wavelet transform. These techniques together with Probabilistic Prognostics and Health Management tools can be used to enhance the estimation of the remaining lifetime of the mechanical system before failure.

References

1. R.L. Norton, *Kinematics and Dynamics of Machinery* (McGraw Hill Higher Education, 2008)
2. D.W. Dudley, *Handbook of Practical Gear Design* (Technomic Publishing Company, Lancaster, UK, 1994)
3. A. Flodin, S. Andersson, Simulation of mild wear in spur gears. Wear **207**, 16–23 (1997)
4. S.L. Harris, Dynamic loads on the teeth of spur gears. Proc. Inst. Mech. Eng. **172**, 87–112 (1958)
5. R.C.N. Leung, R.J. Pinnington, Vibrational power transmission of an idealized gearbox. J. Sound Vib. **128**, 259–273 (1989)
6. K. Umezawa, T. Sato, J. Ishikawa, Simulation on rotational vibration of spur gears. Bull. Jpn. Soc. Mech. Eng. **27**, 102–109 (1984)
7. M.H. Chen, M.J. Brennan, Active control of gear vibration using specially configured sensors and actuators. Smart Mater. Struct. **9**, 342–350 (2000)
8. P.D. Macfadden, A revised model for the extraction of periodic waveforms by time domain averaging. Mech. Syst. Signal Process. **1**, 83–95 (1987)
9. P.D. Macfadden, Examination of a technique for the early detection of failure in gears by signal processing of the time domain average of meshing vibration. Mech. Syst. Signal Process. **2**, 173–183 (1987)
10. H.R. Martin, Detection of gear damage by statistical vibration analysis. Proc. Inst. Mech. Eng. 395–401 (1992)
11. F.A. Andrade, I.I. Esat, M.N.M. Badi, Gearbox fault detection using statistical methods, time-frequency methods and harmonic wavelet: a comparative study, in *International Congress on Condition Monitoring and Diagnostic Engineering Management* (1999), pp. 77–85
12. J.W. Barnes, *Statistical Analysis for Engineers and Scientists: A Computer-Based Approach* (McGraw-Hill, Singapore, 1994)

13. G.E.P. Box, J.S. Hunter, W.G. Hunter, *Statistics for Experimenters: Design, Innovation and Discovery*, 2nd edn. (Wiley, New Jersey, 2005)
14. M. Lebold, K. McClintic, R. Campbell, C. Byington, K. Maynard, Review of vibration analysis methods for gearbox diagnosis and prognosis, in *Proceedings of the 54th Meeting of Society for Machinery Failure Prevention Technology* (2000), pp. 623–634
15. K. Shin, J.K. Hammond, *Fundamentals of Signal Processing for Sound and Vibration Engineers* (Wiley, Chichester, England, 2008)
16. M.H. Chen, Combining the active control of gear vibration with conditioning monitoring, Ph. D. Thesis, University of Southampton, UK (1999)
17. P.M. Bentley, J.T.E. McDonnel, Wavelet transforms: an introduction. J. Acoust. Soc. Am. **89**, 175–186 (1994)

Condition Monitoring of Structures Under Non-ideal Excitation Using Low Cost Equipment

Paulo J. Paupitz Gonçalves and Marcos Silveira

Abstract Monitoring the integrity of structures and machines is an evergrowing concern in engineering applications. Better knowledge of structural conditions allows optimized maintenance cycles, increasing the availability and return of investment, and preventing failure of various systems from manufacturing equipment to air and land vehicles. A common way of evaluating the integrity of mechanical systems is capturing and analyzing vibration signals during operation. Many of the condition monitoring systems are highly specialized, incurring high initial investment. In this context, the objective of this work is to demonstrate the possibility of using low-cost systems for monitoring the integrity of structures. The use of piezoelectric sensors to capture vibration signals is currently ubiquitous, and acquisition and conditioning of these signals can be performed by low cost and open source logic programmable microcontrollers such as Arduino. Structures coupled to non-ideal motors (such that the phenomenon of resonance capture can occur) are used in this study. Controlled structural modifications are performed by the addition of point masses along the length of the beam, and by the application of magnetomotive forces with the use of an electromagnet at a fixed point on the beam. The experimental data is compared to analytical and numerical results, and to an established commercial system, demonstrating the possibility of satisfactory monitoring of structural integrity with such system.

Keywords Condition monitoring · Non-ideal excitation · Sommerfeld effect · Low cost

1 Introduction

Structural integrity monitoring (more specifically, damage detection at the earliest possible stage) is an increasingly studied topic in many engineering applications in order to improve manufacturing planning, machinery performance, and to prevent

P.J. Paupitz Gonçalves (✉) · M. Silveira
Faculty of Engineering, São Paulo State University – UNESP, São Paulo, Brazil
e-mail: paulo.jpg@feb.unesp.br

S. Ekwaro-Osire et al. (eds.), *Probabilistic Prognostics and Health Management of Energy Systems*, DOI 10.1007/978-3-319-55852-3_15

downtime, failures, financial loss and disasters. Damage to structures can be defined as undesirable changes to a given system. Such changes can be either geometric such as deformations or wear, as well as changes to the parameters defining the material properties such as stiffness or mass.

Many examples of machines attached to flexible structures can be found in civil, mechanical and aerospace engineering. For example, aircraft wings with jet engines or propellers, cranes for loading vessels and industrial hoists. Rotors tend to become unbalanced due to the accumulation of particles, magnetic asymmetry and uneven wear of its components and may become a source of vibratory mechanical excitation. These induced vibrations may coincide with the eigenfrequencies of vibration modes of the structure, often resulting in undesirable effects, such as resonance capture, which causes energy to be transferred to increase structural vibration instead of motor rotation.

Farrar et al. [1, 2] define the scope of condition monitoring as a part an area known as Prognosis and Heath Monitoring of structures. They define two classes of monitoring. The first is the *Usage Monitoring*, which is the process of acquiring operational data from a structure or system. On the other hand, *Health Monitoring* is the process of identifying the presence and quantifying the quantity of damage in a system based on information extracted from measured data.

Specialized integrity monitoring systems exist, and may consist of sensors integrated in the structure, data acquisition hardware and signal analysis software. Due to their high initial costs, the complexity and size of such systems, only critical equipment or components are monitored. In other cases, it is impractical to install the necessary number of sensors due to size or harsh conditions restrictions. In this context, the use of modular and low cost equipment for data acquisition has become an interesting solution. An important compromising relationship exists between cost, reliability and signal processing capabilities of the systems. Currently, it is possible to obtain sensors and microcontrollers with high acquisition rates which are relatively low cost, which makes it interesting to understand their applicability as structural health monitoring systems in various engineering areas. The recent push in the direction of connectivity and big data analysis (such as Internet of Things) relies on the accessibility, availability and reliability of sensors and signal processing to enable wide adoption.

1.1 Smart Structures and Structural Health Monitoring

Structures are said to be intelligent or smart when they are able to detect and resolve problems before a failure occurs, by receiving signals from sensors and then processing them via a central control unit. In general, smart structures should employ sensors that record internal and external information, contain actuators to apply designated forces and have a central control system capable of acquiring signals and making decisions. In these terms, any structure that is capable of capturing different signals in response to any change in the environment or integrity can be considered a smart structure.

The techniques involved in Structural Health Monitoring (SHM) allow optimization of the use of the structure (reducing downtime and preventing disasters), enable maintenance planning based on performance or true working condition and, in general, assist the designer in improving the structure. Non-destructive evaluation (NDE) techniques used in SHM include electromechanical impedance, and consist of comparing the signals obtained from a structure without damage (known as the baseline signal) to signals of a structure to be inspected. However, the sensitivity to environmental changes (such as temperature and noise) need to be taken into account when using SHM.

Rytter and Kirkegaard [3] describes the use of inspection based on the vibration and modal analysis to determine structural damage and to estimate the life of the system. Damage identification is divided into four levels: determination of damage present in the structure, determination of the geometric location of the damage, quantification of the severity of damage, and prediction of the remaining life of the structure. Narkis [4] conducted simulations in a simple supported beam under bending and axial vibrations with damage increased along the beam. Comparisons with numerical simulations calculated by the finite element method for validation indicated that data from two natural frequencies of the system is sufficient for the location of the damage.

Doebling [5] described several damage identification methods using mechanical vibrations, including damage detection based on changes in modal properties (defined by frequency of resonance, damping and mode shape), methods based on dynamic measurement of stiffness, methods based on changes of the matrices of the structural model (such as mass, stiffness and damping) and modal information indicating the vibration modes and natural frequencies. The optimal matrix method was presented by Zimmerman and Kaouk [6] who also formulated the algorithm of the minimum rank perturbation, demonstrating how the perturbation of two matrix properties can be estimated simultaneously. Sohn et al. [7] conducted a literature review of structural health monitoring with damage settings, integrity and the methods used to perform the monitoring of the condition of the structure (which involves sensing, acquisition, signal conditioning, development of statistical models to detect changes in the modal parameters).

Yan et al. [8] developed methods of detecting structural damage based on mechanical vibrations with changes in natural frequencies, vibration modes, structural stiffness, transfer function or frequency response of the system and based on statistical information. The development of modern techniques such as wavelet analysis, neural networks and genetic algorithm methods are also mentioned. Signal treatment, processing and analysis using discrete Fourier series and state space methods are exemplified by Lathi [9]. Statistical techniques are required for analysis of samples. The presence of noise is inherent and may interfere to the extent of damage and the sensitivity of amplitude or natural frequency to damage extent is usually very low. The measure of the degree of flattening of the distribution (known as kurtosis), can be used as a filter to the signal [10, 11].

1.2 Non-ideal Excitation and Health Monitoring

As described by Gonçalves et al. [12, 13], rotating machines suffer from unbalance and alignment problems that can lead to excessive levels of vibration causing various undesirable problems and failures. Critical speed occurs when the shaft angular speed matches the shaft bending natural frequency. Laval was the first to perform an experiment with a steam turbine to observe that quick passage though critical speed would significantly reduce the levels of vibration when compared to steady state excitation [14]. This procedure would require a motor with enough power to be accelerated quickly in the range of resonance frequency. In some cases, motors have limited power to perform such operations, and the angular velocity increases so slowly that the passage through resonance becomes a problem.

Another class of problem related to unbalanced motors with limited power was discussed by Sommerfeld [15]. He proposed an experiment of a motor mounted on a flexible wooden table and observed that the energy supplied to the motor was converted in the form of table vibration instead of being converted to increase angular velocity of the motor. This observation was used to explain a class of motors called non-ideal energy sources. The non-ideal energy source have an influence on the system near the resonance regime. When considering a DC motor, usually the angular velocity increases according to the power supplied by the source. However, due to the Sommerfeld effect, near the resonance and with additional energy, the average angular velocity of the DC motor remains unchanged until it suddenly jumps to a much higher value upon exceeding a critical input power. Simultaneously, the amplitude of oscillations of the excited system jumps to a much lower value. Before the jump, the non-ideal oscillating system cannot pass through the resonance frequency of the system, or requires an intensive interaction between the vibrating system and the energy source to be able to do so [14, 16, 17].

The interaction between non-ideal motors and flexible structures has been studied by many authors. A review of non-ideal energy sources is presented by Balthazar et al. [14] and Cveticanin [18]. Eckert [19] presents a brief review of the problem investigated by Sommerfeld. Blekhman et al. [20] discuss the motion of an unbalanced rotor when passing through a resonance zone solved by the iteration method combined with the method of the direct separation of motions. Dimentberg et al. [21] presents a method to avoid resonance capture by switching on and off a mechanism to change the stiffness of an engine mount, while Castão et al. [22] makes use of magneto-rheological dampers to avoid resonance capture. Tsuchida et al. [23] studied the dynamics of a non-ideal system with two coupled oscillators, with results that showed that jump phenomena and chaos are present for certain values of the parameters in the resonant regime. Zukovic and Cvetićanin [16, 17] detected the Sommerfeld effect and chaotic regimes on the dynamics of a non-ideal system comprised of an oscillator connected with an unbalanced motor with clearance. Moraes et al. [24] analyzed the dynamics of a vibro-impact system with a non-ideal source by means of a DC motor with limited power supply and an unbalanced rotor. Three different situations were presented: in the first situation, the motor has reached a

steady state angular frequency similar to the angular velocity constant of the motor, while in the second and third situations, the motor exhibits resonance capture.

Considering the analysis of such systems, Palacios et al. [25] applied the Bogoliubov Averaging Method to study the vibrations of an elastic foundation with a non-ideal energy source. They considered a model consisting of a planar portal frame with quadratic nonlinearities and internal resonance 1:2 that supported a direct current motor with limited power. Quinn et al. [26] presented an approximated method to identify which sets of initial conditions lead to resonance capture. Kerschen et al. [27] reported an experimental study of transient resonance capture that may occur in a system of two coupled oscillators with essential nonlinearity, that showed that, during transient resonance capture, the two oscillators are in a state of resonance, the frequency of which varies with time. Lee et al. [28] studied the dynamics of a two-degree-of-freedom (DOF) nonlinear system consisting of a grounded linear oscillator coupled to a light mass by means of an essentially nonlinear stiffness. They first considered the undamped system and performed a numerical study based on non-smooth transformations to determine its periodic solutions in a frequency-energy plot. Bishop and Galvanetto [29] considered the behavior of a mechanical oscillator with cubic nonlinearity subjected to a forcing excitation whose frequency remained constant while the amplitude was ramped. They found that the reduced level of forcing at the initial stages of ramping produces a delay in bifurcational events when compared to the constant sinusoidally forced counterpart. Felix et al. [30] studied a nonlinear control method based on the phenomenon of mode saturation which was applied to a portal frame support and unbalanced motor with limited power. An alternative method was analized in [31] which consists in the energy transfer of a structure (cantilever beam) with a non-ideal motor using Linear Electromechanical Vibration Absorber (LEVA) and a Nonlinear Electromechanical Vibration Absorber (NEVA).

In terms of continuous systems with coupled motors, Krasnopolskaya [32] studied an infinite plate immersed in an acoustic medium. The plate was subject of a point excitation by an electric motor of limited power-supply, and it was shown that chaos might occur in the system due to the feedback influence of waves in the infinite hydroelastic subsystem in the regime of motor shaft rotation. In terms of damage detection, Ko et al. [33] presented a method combining sensitivity analysis and MAC/COMAC analysis that showed that some methods can predict structural changes without information of the system parameters by comparing healthy with damage states by means of signal processing.

1.3 Piezoelectric Sensors

A piezoelectric sensor is a passive transducer that converts mechanical excitation energy into electrical energy (and vice versa) due to the phenomenon of the generation of electrical charges on the surface of a material under mechanical strain, a process called piezoelectricity. Piezoelectricity properties are related to the crystalline structure of the Perovskite type. Piezoelectric elements began to be studied in

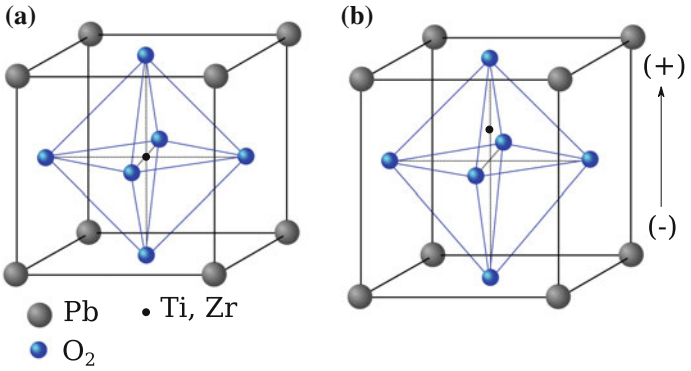


Fig. 1 Cristalline structure of piezoelectric ceramics

1880 [34] for generating an electric charge by means of pressure applied to crystals. The reverse effect was studied shortly after [35], and the electric potential difference application in crystals was already called piezoelectric, leading to mechanical deformations.

From a structure as shown in Fig. 1a, the unpolarized ceramic is centralized (cubic structure), which occurs when the temperature is above the Curie point. The structure shown in Fig. 1b has tetragonal symmetry in which the center of symmetry of the positive electric charge does not coincide with the center of symmetry of negative charges and generates an electric dipole. The Curie point defines the critical temperature that divides the two situations shown [36].

The equivalent electrical circuit of a piezoelectric element is a resonant serial RLC circuit in parallel to a capacitor equivalent to the parallelism of the materials (as can be seen in Fig. 2a). The characteristic impedance curve (a function of frequency) reveals two points resulting in impedance at resonance which are the

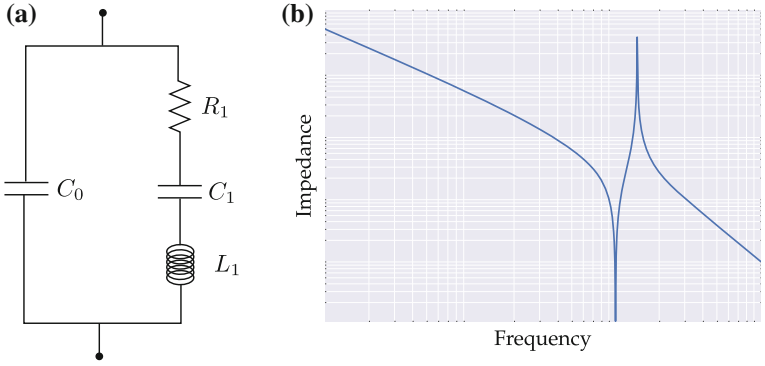


Fig. 2 Equivalent electrical circuit of piezoelectric sensor and characteristic impedance curve

minimum impedance point (Z_{min}) at frequency f_1 and maximum impedance (Z_{max}) at frequency f_2 (as shown in Fig. 2b). This is why piezoelectric elements are widely used in components that promote stability in oscillator circuits (called XTAL).

The impedance as a function of the frequency is given by

$$Z(\omega) = \frac{1 - C_1 L_1 \omega^2}{j\omega \left(C_0 + C_1 - C_0 C_1 L_1 \omega^2\right)} \tag{1}$$

where the resonance frequency is given by

$$f_1 = \frac{1}{2\pi} \sqrt{\frac{1}{C_1 L_1}} \tag{2}$$

and the anti-resonance is written as

$$f_2 = \frac{1}{2\pi} \sqrt{\frac{C_0 + C_1}{C_0 C_1 L_1}}. \tag{3}$$

The objective of this work is to demonstrate the possibility of using low-cost systems for monitoring the integrity of structures. A frame structure coupled to a non-ideal motor is used in this study so that the phenomenon of resonance capture can occur. Controlled structural modifications are performed by the addition of point masses along the length of the beam, and also by the application of magnetomotive forces with the use of an electromagnet at a fixed point on the beam. The experimental data is compared to analytical and numerical results and to an established commercial system in order to evaluate the possibility of satisfactory monitoring of structural integrity with such system. In the sequence, a 2-DOF discrete (lumped) parameter model of a mass vibrating in a plane is presented in Sect. 2 which is used to investigate the phenomenon of resonance capture and demonstrate the Sommerfeld Effect. An experimental set-up was built (which is compatible with the 2-DOF model) and its sensors and data acquisition system are presented in Sect. 3. The experimental results are presented in Sect. 4, including the damage emulations and comparisons with a commercial monitoring system.

2 Mathematical Modeling

The system considered in this section is presented in Fig. 3, which consists of a block with mass M supported by springs and viscous dampers in two orthogonal directions (x and y). The spring constants are defined by k and the dampers by c. The subscripts x and y indicate the displacement direction. Attached to the mass is a rotating motor,

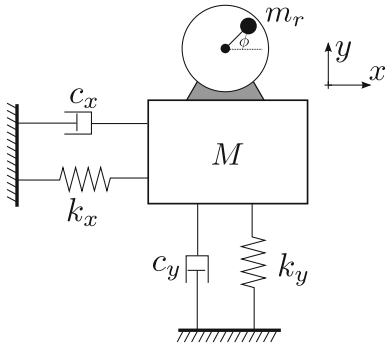


Fig. 3 Discrete parameter system with two degrees of freedom and coupled non-ideal unbalanced motor

with an unbalanced mass m at a distance r from the center of the motor shaft. The motor shaft has moment of inertia defined by J_0.

The equations of motion of an electrical motor attached to a structure are developed based on Lagrange equations and the model is related to an experimental device described in later sections.

2.1 Energy Equations

To apply Hamilton's principle, expressions for the kinetic and the potential energy need to be written in terms of the unknown degrees of freedom. The kinetic energy is defined as

$$T = \frac{1}{2}M\dot{x}^2 + \frac{1}{2}M\dot{y}^2 + \frac{1}{2}J_0\dot{\phi}^2 + \frac{1}{2}m\left(\dot{x}_m^2 + \dot{y}_m^2\right). \tag{4}$$

The term J_0 defines the motor shaft moment of inertia, and the terms $x_m = x + r\cos\phi$ and $y_m = y + r\sin\phi$ define the position of the motor's unbalanced mass m, with r being the distance of this mass to the motor's center of rotation. Thus, Eq. 4 can be written as

$$\begin{aligned} T = {} & \frac{1}{2}(M+m)\dot{x}^2 + \frac{1}{2}(M+m)\dot{y}^2 + \frac{1}{2}\left(J_0 + mr^2\right)\dot{\phi}^2 \\ & + mr\dot{\phi}\left(\dot{y}\cos\phi - \dot{x}\sin\phi\right). \end{aligned} \tag{5}$$

If the gravity potential energy is neglected, then the system's potential energy is simply

$$U = \frac{1}{2}k_x x^2 + \frac{1}{2}k_y y^2. \tag{6}$$

2.2 Equations of Motion

The equations of motion of the system are obtained by writing the Lagrangian, $L = T - U$, and first-order stationary conditions in the form of Hamilton's equation

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \left(\frac{\partial L}{\partial q_i}\right) = F_i \tag{7}$$

in which F_i are the non-conservative forces, which are the viscous damping forces $F_x^{damp} = -c_x\dot{x}$, $F_y^{damp} = -c_y\dot{y}$ and the torque $\mathfrak{M}$ supplied by the motor.

Applying Eqs. 5 and 6 into 7, it is possible to obtain the block's equations of motion for x and y directions

$$(M+m)\,\ddot{x} + k_x x + c_x\dot{x} = mr\left(\dot{\phi}^2\cos\phi + \ddot{\phi}\sin\phi\right) \tag{8}$$

$$(M+m)\,\ddot{y} + k_y y + c_y\dot{y} = mr\left(\dot{\phi}^2\sin\phi - \ddot{\phi}\cos\phi\right) \tag{9}$$

and the equation of motion for the unbalanced mass

$$\left(J_0 + mr^2\right)\ddot{\phi} = mr\left(\ddot{x}\sin\phi - \ddot{y}\cos\phi\right) + \mathfrak{M}\left(\dot{\phi}\right). \tag{10}$$

Equations 8, 9 and 10 can be conveniently written in terms of the parameters

$$\omega_x = \sqrt{\frac{k_x}{M+m}} \qquad \xi_x = \frac{c_x}{2(M+m)\omega_x} \qquad \mu_1 = \frac{mr}{M+m}$$

$$\omega_y = \sqrt{\frac{k_y}{M+m}} \qquad \xi_y = \frac{c_y}{2(M+m)\omega_y} \qquad \mu_2 = \frac{mr}{J_0+mr^2}$$

such that

$$\begin{aligned}
\ddot{x} + \omega_x^2 x + 2\xi_x\omega_x\dot{x} &= \mu_1\left(\dot{\phi}^2\cos\phi + \ddot{\phi}\sin\phi\right)\\
\ddot{y} + \omega_y^2 y + 2\xi_y\omega_y\dot{y} &= \mu_1\left(\dot{\phi}^2\sin\phi - \ddot{\phi}\cos\phi\right)\\
\ddot{\phi} &= \mu_2\left(\ddot{x}\sin\phi - \ddot{y}\cos\phi\right) + \mathfrak{M}\left(\dot{\phi}\right)/\left(J_0+mr^2\right).
\end{aligned} \tag{11}$$

2.3 Model Order Reduction

The order of the equations describing the motion of the system (Eq. 11) is reduced by the use of the state variables $q_1 = x$, $q_2 = y$, $q_3 = \phi$, $q_4 = \dot{x}$, $q_5 = \dot{y}$ and $q_6 = \dot{\phi}$, such that the velocities are re-written as

$$\dot{q}_1 = q_4 \qquad \dot{q}_2 = q_5 \qquad \dot{q}_3 = q_6. \tag{12}$$

The accelerations can then be calculated by solving the linear system of differential equations

$$\begin{bmatrix} 1 & 0 & -\mu_1 \sin\phi \\ 0 & 1 & \mu_1 \cos\phi \\ -\mu_2 \sin\phi & \mu_2 \cos\phi & 1 \end{bmatrix} \begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{\phi} \end{bmatrix} = \begin{bmatrix} -\omega_x^2 x - 2\xi_x \omega_x \dot{x} + \mu_1 \dot{\phi}^2 \cos\phi \\ -\omega_y^2 y - 2\xi_y \omega_y \dot{y} + \mu_1 \dot{\phi}^2 \sin\phi \\ \mathfrak{M}(\dot{\phi}) / \left(J_0 + mr^2\right) \end{bmatrix} \tag{13}$$

in which $\dot{q}_4 = \ddot{x}$, $\dot{q}_5 = \ddot{y}$ and $\dot{q}_6 = \ddot{\phi}$.

2.4 Non-ideal Motor

To define a limited power (or non-ideal) motor, two parameters are used to represent the torque as a function of the angular velocity, given by

$$\mathfrak{M}(\dot{\phi}) = M_0 \left(1 - \frac{\dot{\phi}}{\Omega_0}\right) \tag{14}$$

in which M_0 and Ω_0 are constants of the motor, the first related to static torque and the second related to zero torque.

Equation 14 is represented by the curve shown in Fig. 4 where the torque $\mathfrak{M}$ is a function of the angular velocity $\dot{\phi}$. For values of angular velocity equal to Ω_0, the torque reduces to zero, and when the angular velocity is zero, the torque is maximum and equal to M_0.

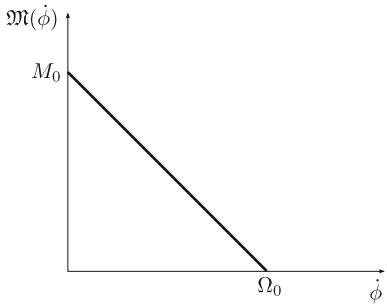


Fig. 4 Motor torque characteristic curve (adopted from Ref. [12])

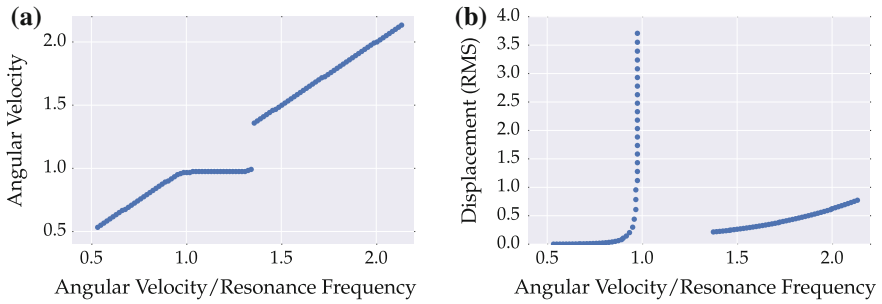


Fig. 5 The illustration of the Sommerfeld effect. **a** Angular velocity as a function of Ω_0 and **b** displacement amplitude as a function of Ω_0

2.5 Sommerfeld Effect

This numerical example considers the case of setting the motor angular velocity to a fixed value. The motor is accelerated from rest to a fixed velocity by changing the parameter Ω_0. The simulations were performed for frequencies around the mass block resonance frequency ω_0 and are presented in Fig. 5. Figure 5a shows that when Ω_0 is slightly bigger than ω_0 the angular velocity does not increase. For instance, when setting $\Omega_0 = 1.1\omega_0$, the motor does not reach the angular velocity $1.1\omega_0$; instead it will oscillate with angular velocity ω_0. The consequence is that this energy is transferred to cart displacement amplitude. The mass magnitude (in RMS (root mean square)) is shown in Fig. 5a as a function of the oscillation frequency as the parameter Ω_0 increases.

3 Experimental Set-Up

The experimental system considered in this work is a frame structure consisting of a long horizontal beam supported by two shorter vertical beams, as shown schematically in Fig. 6. The bending stiffness of the horizontal beam corresponds to k_y of the 2-DOF discrete system shown in Fig. 3, while the equivalent bending stiffness of the two vertical beams corresponds to k_x. Attached to the center of the horizontal beam is a non-ideal electrical DC motor with an unbalanced mass that excites the structure at a frequency determined by its angular velocity. The first two mode shapes were investigated. In the first mode, the horizontal beam had large bending motion, while

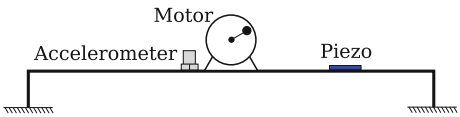


Fig. 6 The experimental set-up illustrating the sensors and the motor

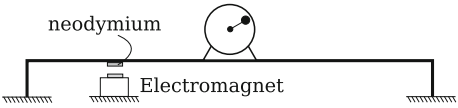


Fig. 7 The experimental set-up showing the use of electromagnetic transducer to introduce structural changes in the system

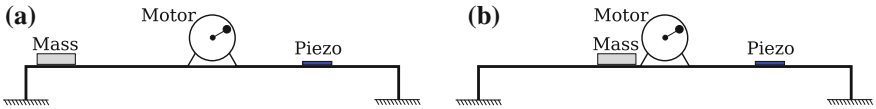


Fig. 8 The experimental set-up showing the use of concentrated masses in two positions to introduce structural changes in the system. **a** Mass at 35 mm and **b** mass at 160 mm from the corner

the vertical beams had very small motion. In the second mode, all beams presented large bending motion.

Structural modifications were applied to the structure in two forms: the first was by the introduction of an electromagnet and a permanent magnet as illustrated in Fig. 7. The second form of structural modification was introduced by addition of concentrated masses in two different positions as shown in Fig. 8.

3.1 Using Arduino UNO as a DAQ

The success of the monitoring system depends heavily on the sensors and the data acquisition. The application of piezoelectric elements as sensors was explained by Park et al. [37] and Wang et al. [38] (the latter specifically for concrete structures). Guechaichia and Trendafilova [39] conducted experiments of fault detection on a beam, detecting and locating damage using only the first natural frequency of the system, and the Arduino UNO microcontroller.

A summing non-inverter amplifier circuit was built using a LM741 operational amplifier with two inputs: one for the signal generated by the piezoelectric element and another for a DC input adjustable by a potentiometer. This set-up allows the actual sum of signals from the inputs, and has a gain easily calculated by the ratio of feedback resistors with negative input. The circuit requires a symmetrical DC power of 12 V and has a variable gain. The amplifier circuit has a voltage gain in the range from 1 to 6 (0 to 15 dB) and DC voltage of 0 to 4 V (its use is necessary because the signal generated by the piezoelectric sensor which is centered in neutral and symmetrical point), contains positive and negative components, and the amplitude is proportional to the deformation. Depending on the excitation source, the signal may be too low for good detail in the acquisition. Further to that, the signal must be conditioned to have the peak voltage at the maximum analog input voltage of the microcontroller (5.5 V), and its midpoint at half this voltage (providing that no

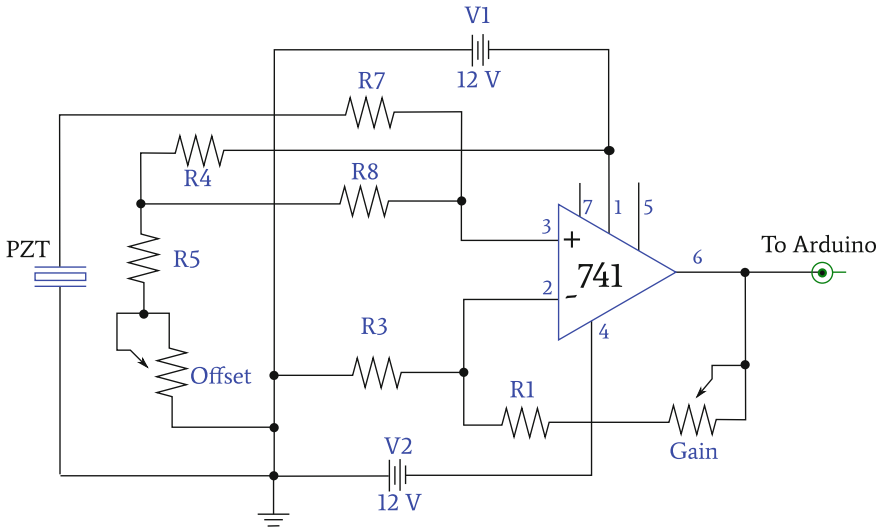


Fig. 9 Signal conditioning for acquisition via Arduino microcontroller

signal component is negative) as the microcontroller accepts only positive signals. This process is depicted in Fig. 9.

3.1.1 Resolution

Arduino UNO has 10 bit resolution which means that it has (2^{10}) 1023 divisions of the reference voltage (which is by default 5 V). This provides a voltage resolution of 4.9 mV. The reference voltage can also be adjusted internally to 1.1 V or other voltage using an external reference. If 1.1 V is used, a resolution of 1.1 mV can be achieved (Fig. 10).

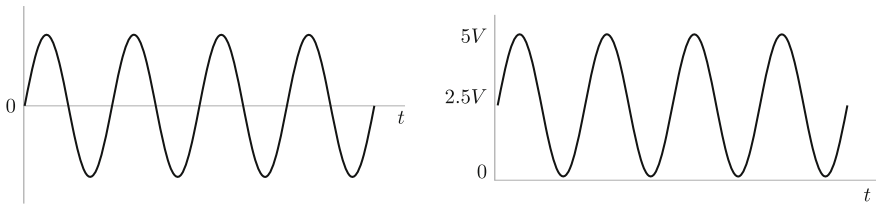


Fig. 10 Original voltage signal from sensor, and signal entering the Arduino

3.1.2 Data Sampling

Commercial DAQs acquire data at given time intervals, so if, for instance, a commercial data acquisition is set to acquire 1000 samples per second (1 kHz), there is no need to capture the time each sample was acquired since it is known that every sample was captured at a constant interval. This is not true when using Arduino as an acquisition system, as the system does not sample at constant time intervals because other Arduino tasks share the same CPU. A solution for this is to print a time stamp for every sample.

4 Experimental Results

In this section, the results acquired by the piezoelectric sensor using Arduino are compared to the results acquired by the accelerometer with a commercial data acquisition system. The signals after structural modifications are used. To perform the comparison between the two sensors, the measurements from the piezoelectric sensor and accelerometer were normalized.

4.1 Baseline Signals

Figure 11 shows the baseline signal using the piezo sensor and accelerometer as RMS of measured signal as function of voltage applied to the motor.

4.2 Damage Emulations

In order to introduce a modification of the stiffness of the structure, an electromagnet was placed under the beam and a neodymium magnet was attached to the beam (as shown in Fig. 12). With this set-up, the equivalent bending stiffness of the beam (with the attached magnet) was increased. The electromagnet was used in two levels, 6 and 12 V, the last being its maximum input voltage. Figure 13 shows the signals from baseline and electromagnet at 6 V acquired with piezo sensor and with the accelerometer. As a result of the increase in stiffness due to the magnetic interaction, the eigenfrequencies were higher than the baseline (as can be observed in the two resonance peaks).

Another modification to the original beam was performed by addition of mass at different locations, as shown in Fig. 14. As a result of this modification, the eigenfrequencies of the beam are reduced. Figure 15 shows the response of the system measured in three configurations: baseline, additional mass attached at 35 mm from the right edge, and additional mass attached at 160 mm from the right edge. We found

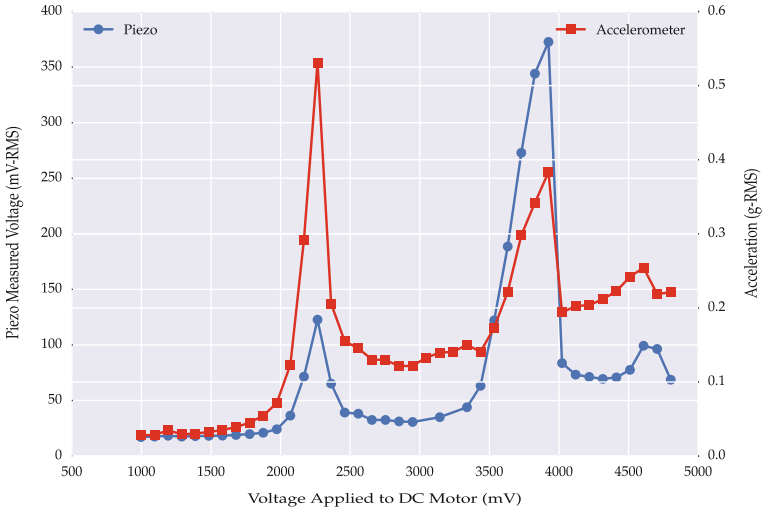


Fig. 11 Root Mean Squared (RMS) as a function of the voltage applied to the motor. This is a comparison of the signal acquired using the piezo sensor (*blue dotted line*) and the signal acquired using the accelerometer (*red squared line*) (color figure online)

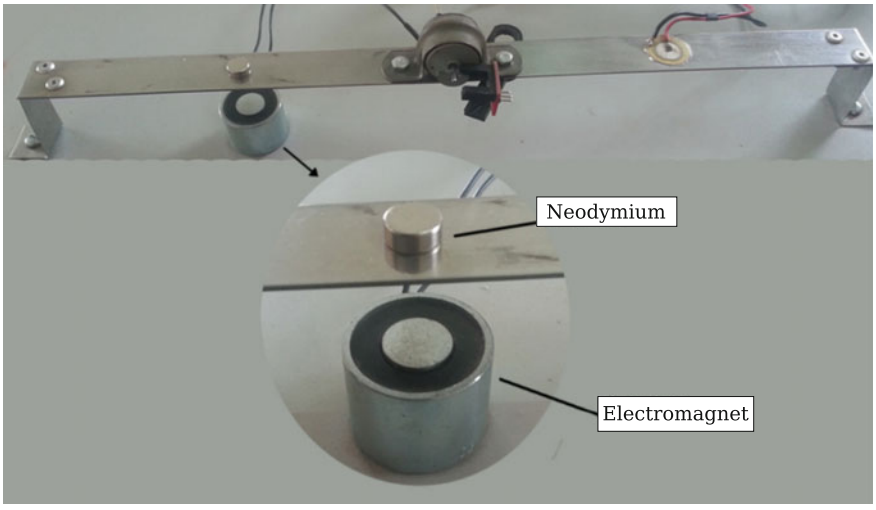


Fig. 12 Detail of the electromagnet coupled to the horizontal beam to emulate damage

that the first modification (35 mm) strongly influenced the second resonance capture while the second modification influenced both resonance captures.

The third and fourth central moments were calculated for the previous signals. These moments give an indication of asymmetry (skewness) and flattening (kurtosis) of the signals, and are shown in Figs. 16 and 17 for the mass modifications, and in

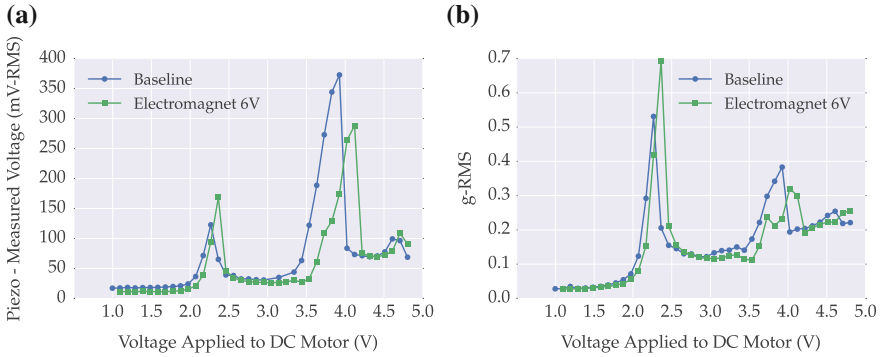


Fig. 13 Root Mean Squared (RMS) as a function of the voltage applied to the motor with the signal acquired for the baseline configuration (*blue dotted line*) and the system modified by the electromagnet (*green squared line*). **a** Piezo, **b** Accelerometer (color figure online)

Fig. 14 Detail of placement of additional mass to emulate damage

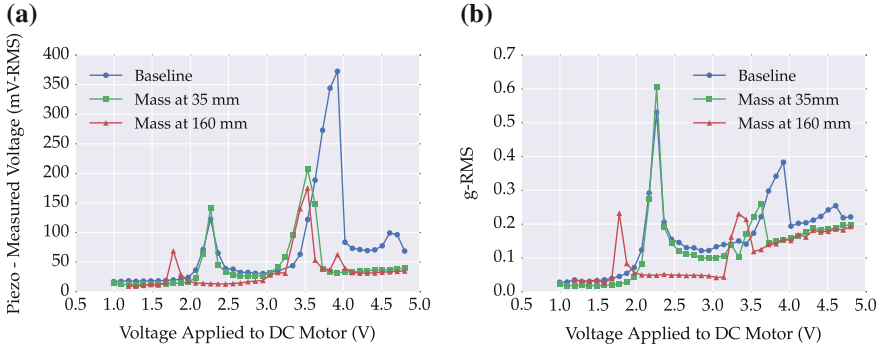


Fig. 15 Root Mean Squared (RMS) as a function of the voltage applied to the motor: signals acquired for the baseline configuration (*blue dotted line*), system modified—mass at 35 mm (*green square line*) and mass at 160 mm (*red triangle line*). **a** Piezo, **b** Accelerometer (color figure online)

Figs. 18 and 19 for the stiffness modifications. Positive skewness indicates that the distribution is asymmetric to the right, which was the case of addition of mass, while negative skewness indicates that the distribution is asymmetric to the left, which was the case of magnetic interaction.

The kurtosis curve indicates that the increase in stiffness exacerbates the peak values at the first resonance and diminishes at the second resonance. Regarding the

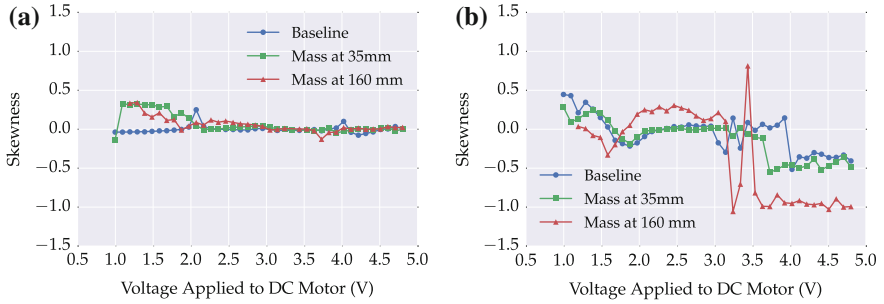


Fig. 16 Skweness mass. **a** Piezo, **b** Accelerometer

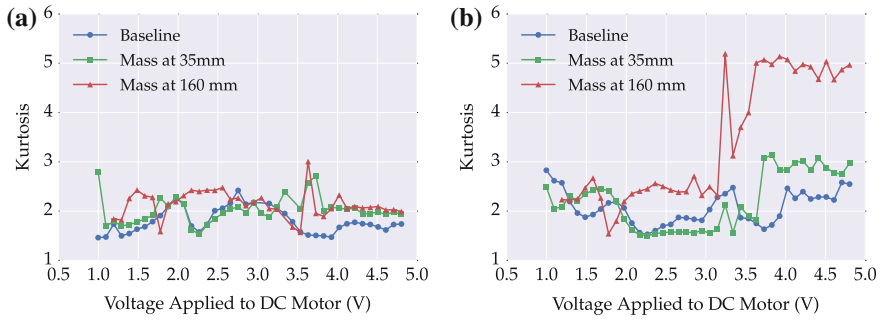


Fig. 17 Kurtosis mass. **a** Piezo, **b** Accelerometer

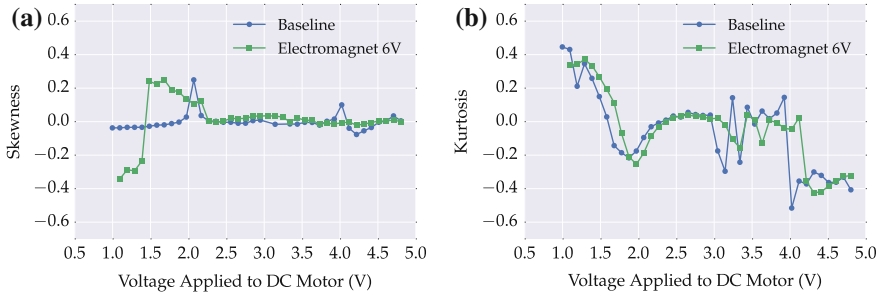


Fig. 18 Skewness. **a** Piezo, **b** Accelerometer

modifications of mass, the first resonance peak is shifted to the right, and this shift is more pronounced as the added mass is closer to the edge of the beam. The addition of mass shifts the second resonance peak to left, and this shift is more pronounced as the added mass is closer to the edge.

When comparing the signal from the piezo sensor to the commercial accelerometer, a difference in peak values can be observed in most curves. The piezo sensor has a maximum reading as the system goes through the second resonance peak, while

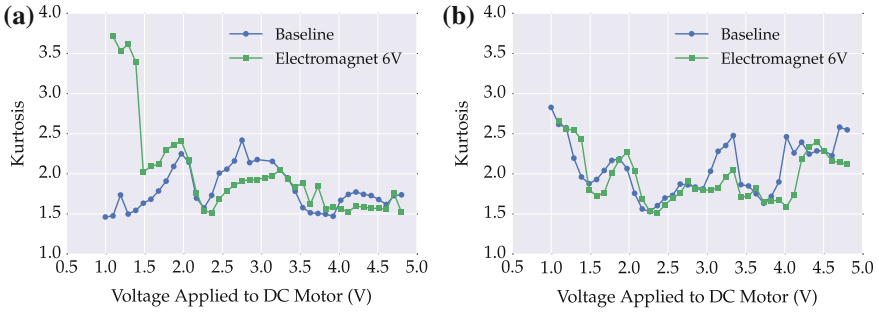


Fig. 19 Kurtosis. **a** Piezo, **b** Accelerometer

the maximum reading of the accelerometer occurs as the system goes through the first resonance.

The accelerometer, as used, senses the vertical vibrations of the beam, and so is more susceptible to the first bending mode. The piezo sensor output depends on the local strains deforming the crystal structure which are greater on the second vibration mode. A possible solution to make both signals more similar is to place the piezoelectric sensor on the vertical supports of the horizontal beam, which would match the strain orientation of the piezo to the movement direction of the accelerometer.

The skewness coefficient was effective in identifying the structural modifications related to the vibration modes to which the sensors had less sensibility, which were the first bending mode for the piezoelectric sensor, and the second bending mode for the accelerometer. This occurs as the shifts related to stiffness or mass modifications are larger relative to the signal amplitude. A similar effect was observed for kurtosis, with the piezo sensor showing changes of stiffness more clearly at the first resonance and the accelerometer at the second resonance.

Apart from this difference in maximum points, both signals were very similar and made possible the distinction among all cases tested, baseline, stiffness and mass modifications. The double resonance capture (Sommerfeld effect) is also clear from all signals. Considering the cost difference between the systems (2–3 orders of magnitude), this makes a low-cost system based on piezoelectric sensors and multifunction microcontroller a very interesting possibility.

5 Conclusions

This work has proposed a method for monitoring structural changes in systems using low cost equipments (such as piezoelectric sensors and microcontrollers). A system considered as an example consisted of an unbalanced DC motor supported by a flexible structure. This is a situation that occurs in many common appliances and could be used as an example for health monitoring of devices connected in the Internet of Things.

In terms of the analysis of such systems, power limited motors can lead to the phenomenon of resonance capture (Sommerfeld effect), and this must be considered when analyzing structural changes in structures with coupled motors.

The low cost system was able to capture the dynamic behavior of the system with a similar capability of an existing commercial system. The results showed that mass and stiffness modifications induce significant changes in the signals. Using a piezoelectric sensor to acquire the signal, such changes were possible to be identified in the frequency response and also via skewness and kurtosis analysis. Further to that, the double resonance capture (Sommerfeld effect) was clear from all signals.

As the piezoelectric sensor measurement is related to local strain, and the accelerometer measurement is related to local acceleration, the measurement values are quantitatively different. Qualitatively, however, two natural frequencies could be clearly observed, and the structural modifications were identified in all cases.

Position and orientation of the piezo sensors for optimizing its characteristics must be considered for better comparison with systems based on accelerometer signals. The piezo sensors have to be positioned on the region with the correct strain orientation relative to the movement direction which is being analyzed.

The results show that the resolution of currently available low-cost sensors (such as piezoelectric sensors) combined to robust statistics measures (such as skewness and kurtosis) embedded in versatile microcontrolers (such as Arduino) makes a very interesting solution to condition monitoring in engineering systems.

References

1. C.R. Farrar, K. Worden, An introduction to structural health monitoring. Philos. Trans. R. Soc. Lond. A Math. Phys. Eng. Sci. **365**, 303–315 (2007)
2. C.R. Farrar, N.A.J. Lieven, Damage prognosis: the future of structural health monitoring. Philos. Trans. R. Soc. Lond. A Math. Phys. Eng. Sci. **365**, 623–632 (2007)
3. A. Rytter, P.H. Kirkegaard, *Vibration Based Inspection of Civil Engineering Structures* (Aalborg Universitetsforlag, 1994)
4. Y. Narkis, Identification of crack location in vibrating simply supported beams. J. Sound Vib. **172**, 549–558 (1994)
5. S.W. Doebling, A summary review of vibration-based damage identification methods. Shock Vib. Dig. **30**, 91–105 (1998)
6. D.C. Zimmerman, M. Kaouk, Structural damage detection using a minimum rank update theory. ASME J. Vib. Acoust. **116**, 222–231 (1994)
7. H. Sohn, C.R. Farrar, F. Hemez, J. Czarnecki, *A Review of Structural Health Monitoring Literature: 1996–2001* (Los Alamos National Laboratory, 2004)
8. Y. Yan, L. Cheng, Z.Y. Wu, L.H. Yam, Development in vibration-based structural damage detection technique. Mech. Syst. Signal Process. **21**, 2198–2211 (2007)
9. B.P. Lathi, *Sinais e Sistemas Lineares* (Bookman, 2007)
10. M.R. Siegel, *Estatística* (McGraw-Hill, 1984)
11. L.J. Hadjileontiadis, E. Douka, A. Trochidis, Crack detection in beams using kurtosis. Comput. Struct. **83**, 909–919 (2005)
12. P.J.P. Gonçalves, M. Silveira, B.R. Pontes Jr., J.M. Balthazar, The dynamic behavior of a cantilever beam coupled to a non-ideal unbalanced motor through numerical and experimental analysis. J. Sound Vib. **333**, 5115–5129 (2014)

13. P.J.P. Gonçalves, M. Silveira, E.A. Petrocino, J.M. Balthazar, Double resonance capture of a two-degree-of-freedom oscillator coupled to a non-ideal motor. Meccanica (2015)
14. J.M. Balthazar, D.T. Mook, H.I. Weber, R.M.L.R.F. Brasil, A. Fenili, D. Belato, J. Felix, An overview on non-ideal vibrations. Meccanica **38**, 613–621 (2003)
15. A. Sommerfeld, Beitrage zum dynamischen ausbau der festigkeislehre. Zeitschrift für Physik A Hadrons and Nuclei **46**, 391–394 (1902)
16. M. Zukovic, L. Cveticanin, Chaotic responses in a stable duffing system of non-ideal type. J. Vib. Control **13**, 751–767 (2007)
17. M. Zukovic, L. Cveticanin, Chaos in non-ideal mechanical system with clearance. J. Vib. Control **15**, 1229–1246 (2009)
18. L. Cveticanin, Dynamics of the non-ideal mechanical systems: a review. J. Serbi. Soc. Comput. Mech. **4**, 75–86 (2010)
19. M. Eckert, The Sommerfeld effect: theory and history of a remarkable resonance phenomenon. Eur. J. Phys. **17**, 285–289 (1996)
20. I.I. Blekhman, D.A. Indeitsev, A.L. Fradkov, Slow motions in systems with inertial excitation of vibrations. J. Mach. Manuf. Reliab. **37**, 21–27 (2008)
21. M. Dimentberg, L. McGovern, R. Norton, J. Chapdelaine, R. Harrison, Dynamics of an unbalanced shaft interacting with a limited power supply. Nonlinear Dyn. **13**, 171–187 (1997)
22. K.A. Castão, L.C. Goes, J.M. Balthazar, A note on the attenuation of the sommerfeld effect of a non-ideal system taking into account a MR damper and the complete model of a DC motor. J. Vib. Control **17**, 1112–1118 (2011)
23. M. Tsuchida, K.L. Guilherme, J.M. Balthazar, On chaotic vibrations of a non-ideal system with two degrees of freedom: resonance and Sommerfeld effect. J. Sound Vib. **282**, 1201–1207 (2005)
24. F.H. Moraes, B.R. Pontes Jr., M. Silveira, J.M. Balthazar, R.M.L.R.F. Brasil, Influence of ideal and non-ideal excitation sources on the dynamics of a nonlinear vibro-impact system. J. Theor. Appl. Mech. **51**, 763–774 (2013)
25. J. Palacios, J.M. Balthazar, R.M.L.R.F. Brasil, On non-ideal and non-linear portal frame dynamics analysis using Bogoliubov averaging method. J. Braz. Soc. Mech. Sci. Eng. **24**, 257–265 (2002)
26. D. Quinn, R. Rand, J. Bridge, The dynamics of resonant capture. Nonlinear Dyn. **8**, 1–20 (1995)
27. G. Kerschen, D.M. McFarland, J.J. Kowtko, Y.S. Lee, L.A. Bergman, A.F. Vakakis, Experimental demonstration of transient resonance capture in a system of two coupled oscillators with essential stiffness nonlinearity. J. Sound Vib. **299**, 822–838 (2007)
28. Y.S. Lee, G. Kerschen, A.F. Vakakis, P. Panagopoulos, L. Bergman, D.M. McFarland, Complicated dynamics of a linear oscillator with a light, essentially nonlinear attachment. Phys. D Nonlinear Phenom. **204**, 41–69 (2005)
29. S.R. Bishop, U. Galvanetto, The behaviour of nonlinear oscillators subjected to ramped forcing. Meccanica **28**, 249–256 (1993)
30. J.L. Felix, J.M. Balthazar, R.M.L.R.F. Brasil, On tuned liquid column dampers mounted on a structural frame under a non-ideal excitation. J. Sound Vib. **282**, 1285–1292 (2005)
31. J.L.P. Felix, J.M. Balthazar, Comments on nonlinear dynamics of a non-ideal duffing-rayleigh oscillator: numerical and analytical approaches. J. Sound Vib. **319**, 1136–1149 (2009)
32. T. Krasnopolskaya, Chaos in acoustic subspace raised by the Sommerfeld-Kononenko effect. Meccanica **41**, 299–310 (2006)
33. J.M. Ko, C.W. Wong, H.F. Lam, Damage detection in steel framed structures by vibration measurement approach, in *Proceedings of 12th International Modal Analysis Conference* (1994), pp. 280–286
34. J. Curie, P. Curie, Développement, par pression, de lélectricité polaire dans les cristaux hémièdres á faces inclinèes. Comptes Rendus **91**, 294–295 (1880)
35. G. Lippmann, Principe de la conservation de l'álectricité, ou second principe de la théorie des phénomènes électriques. J. de Physique Théorique et Appliquée **10**, 381–394 (1881)
36. Imran Patel, *Ceramic Based Intelligent Piezoelectric Energy Harvesting Device*, Advances in Ceramics—Electric and Magnetic Ceramics, Bioceramics, Ceramics and Environment (InTech, 2011)

37. G. Park, C.R. Farrar, F.L. Scalea, S. Coccia, Performance assessment and validation of piezoelectric active-sensors in structural health monitoring. Smart Mater. Struct. **15**, 1673–1683 (2006)
38. D. Wang, Health monitoring of reinforced concrete structures based on PZT admittance signal, in *Proceedings of the SPIE*, vol. 7493 (2009)
39. A. Guechaichia, I. Trendafilova, A simple method for enhanced vibration-based structural health monitoring. J. Phys. Conf. Ser. **305**, 012073 (2011)

Maintenance Management and Case Studies in the Luís Carlos Prestes Thermoelectric Power Plant

Bernardo Botamede, Leonardo Leucas and Marcelo Pelegrini

Abstract Operating on the growing Brazilian thermoelectricity market since 2002, Petrobras S.A. is today the largest thermoelectric generation company nationwide and the seventh on the overall Brazilian energy market. During the past few years, Petrobras has increased its generation capacity reaching 6,885 GW of installed capacity. In order to achieve the maximum availability and performance of its machines, Petrobras has built global performance and conditioning monitoring systems which are applied to support complementary monitoring strategies and predictive maintenance tasks at Luís Carlos Prestes (LCP) thermoelectric power plant. Specific models are under development with the purpose of further enhancing turbomachinery performance and reliability.

Keywords Asset management · Machine diagnosis · Maintenance tasks

1 Introduction

Every day and since the new market demands have become apparent, maintenance management has sought new approaches to asset reliability optimization problems [1]. For example, the introduction of ISO 55.000 standards offers an integrated view of the asset management process, which then has a direct impact on maintenance policies.

One of the most important tools for optimization and active cost reduction is to implement effective and comprehensive usage of predictive maintenance techniques, which may be defined as an intervention methodology in equipment based on the verification and analysis of the condition or performance parameters and which follows a pre-defined methodology [2].

B. Botamede · L. Leucas · M. Pelegrini (✉)
Petrobras S.A. – Luís Carlos Prestes Thermoelectric Powerplant, Três Lagoas, MS, Brazil
e-mail: mpelegrini@petrobras.com.br

S. Ekwaro-Osire et al. (eds.), *Probabilistic Prognostics and Health Management of Energy Systems*, DOI 10.1007/978-3-319-55852-3_16

Furthermore, enhancement of the operation and maintenance of thermal power plants can be achieved through integrated monitoring systems that continuously observe the intrinsic variables of the various processes in order to reveal the true condition of specialized components and systems. Additionally, any degradation of a thermal power plant can be monitored and recorded continuously using standard computer systems and plant instrumentation.

It is also imperative that all of this information must be available to all parties involved in the maintenance process (including the technicians and other stakeholders) to support their final decisions and then to further evaluate their actions.

2 Objective

The object of this paper is to present a general maintenance philosophy adopted in Luís Carlos Prestes (LCP) power plant in order to achieve the best reliability and performance for its machinery, including:

- Description of the predictive tasks performed in the power plant;
- General explanation of two integrated monitor systems applied in Petrobras;
- Present three case studies that illustrate how monitoring systems and predictive maintenance can be used by a local team for effective tracking of turbomachinery and then, if action is required, to support any further corrective intervention decisions.

3 Predictive Maintenance Strategy

Ongoing and effective maintenance management policies in LCP power plant rely on increased use of predictive maintenance techniques and, when combined with the concepts of Total Productive Maintenance (TPM), facilitate the achievement of power plant goals that should be the combination of highest availability and reliability rates with the lowest costs.

Predictive maintenance techniques applied in LCP power plant include:

- Vibration analysis for general rotary equipment, performed by a local technical team;
- Vibration analysis on turbomachinery, performed by subcontractors;
- Oil analysis (as applied to turbomachinery, gear boxes and special pumps systems);
- Thermography and ultrasound tests applied to electric equipment.

4 Integrated Monitoring Systems [3, 4]

Petrobras has implemented integrated monitoring systems to support several industrial facilities, such as the Plant Information Management System (PIMS) which is applied to both power plants and refineries units. PIMS is used to gather general data available on machinery that can be used for further historical analysis or to supply raw data for post-processing.

Other specific tool available in Petrobras power plants is the Diagnosis and Monitoring Center (or CMD) that receives data from critical equipment in power plants. CMD has tools designed for data processing and real time analysis which can support condition monitoring and provide the following data for a more complete picture:

- Variable trends and variable data exportation;
- Equipment status reports and early alarms of defects or process deviations;
- Balance calculations and combined cycle efficiency;
- Customized monitor models for each equipment;
- Data comparison of similar machinery and arrangements among power plants;
- Fault trees for equipment failure;
- Database with failure rates of the monitored equipment.

All of the above-mentioned integrated systems receive raw data from digital control systems via local network servers, and the access to these systems for local users is granted through remote applications. Figure 1 illustrates the LCP power plant maintenance philosophy based on both predictive tasks and the integrated monitoring systems.

5 Case Studies

The case studies presented in this section work that have been performed (or is under development) by the Petrobras engineering team and is related to predictive maintenance and prognostics models for turbomachinery system optimization and further reliability enhancement.

5.1 Case Study #1: Detecting Damage on Turbine Bearing

After a shutdown of one GE 6FA gas turbine, it was noted that the temperature indicated for one of its bearings had started to increase and then begin to reach the alarm level approximately one month of operation after the event. Then, an

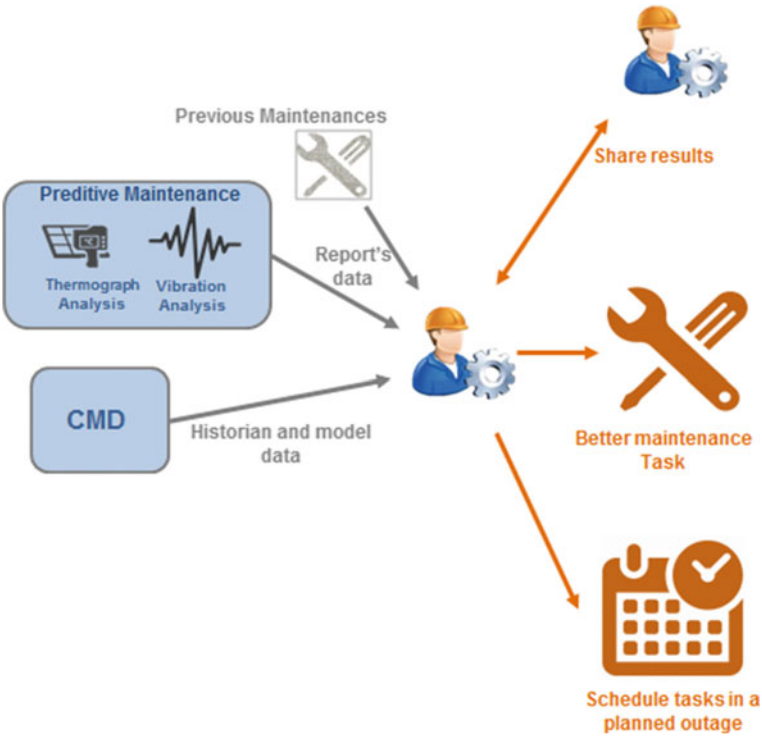


Fig. 1 LCP maintenance diagram

investigation process of all operational parameters was initiated to support any further inspection on the machine. After a careful analysis of the available data, it was found that some vibration levels of the bearings had been increased slightly during the continued operation of the machine, and then by further small increments after each shutdown/start up. However, such vibration levels were still far below the alarm limits, as shown in Fig. 2.

With the information provided by the monitoring systems, the engineering team decided to proceed with an advanced vibration analysis during shutdown, which detected an issue in a specific bearing. The damaged bearing was finally replaced in a planned outage, allowing the equipment to keep operating safely and avoiding any major damage or change in productivity.

This case study analysis demonstrates that it is possible to utilize data as an early signal and as a valuable alarm in fore seeing potential problems, even if only a function of a very simple variable pattern such as rate of change over a long time. Other complex pattern alarms are possible but will need further developments.

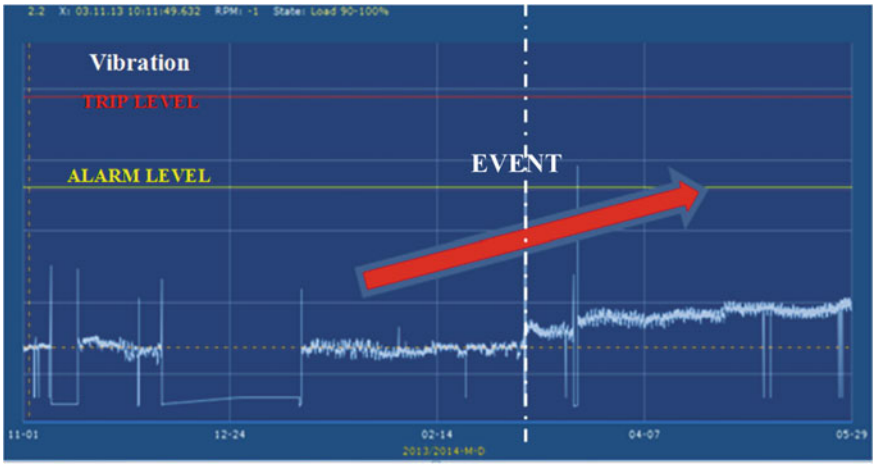


Fig. 2 Temperature and vibration analysis on CMD [3]

5.2 *Case Study #2: Gas Turbine Offline Washing Time Optimization*

The deposition of fouling on the compressor blades is one of the major mechanisms responsible for gas turbine degradation. It is caused mostly by the presence of air contaminants, such as, dust, sand, salt, moisture, oil, and many others and has negative impact on airflow and compressor efficiency. In order to minimize such negative impacts and to recover turbine performance, a suitable maintenance procedure is to wash the compressor blades during compressor offline periods.

One major issue is to determine the best time to execute offline washing, in order to achieve best turbine average performance throughout each operation period and to guarantee minimal compressor surge margin for the machine safe operation. With the monitoring systems available, Petrobras teams are able to model and monitor compressor fouling degradation and should define the best moment to schedule gas turbine outage and to execute offline washing, as exemplified on Fig. 3.

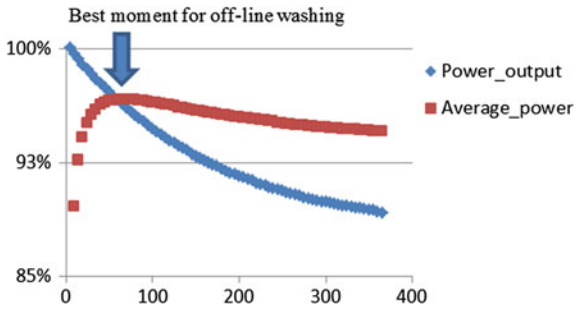


Fig. 3 Curves for turbine power output after time period between offline washings and turbine average power within the same period of time. Data is shown as percentage of initial values in function of time period between offline washings

5.3 Case Study #3: Gas Turbine Compressor Monitoring

The occurrence of a failure on the compressor blades represents a major risk in gas turbine machinery, resulting in downstream damage throughout the compressor and turbine sections. Many events of this nature have been reported across the world during the last decade, with some of them reporting generalized damage throughout the whole compressor. Consequently, several different monitoring approaches have been developed in order to enhance gas turbines compressor reliability [5].

Along with market available solutions, which are under technical-economic viability analysis for implementation, Petrobras engineering team objective is to develop mathematical models based on additional vibration probes (not provided by equipment manufacturer) located on compressor casing. The aim is to identify vibration patterns and develop a monitoring tool that should be capable of detecting major compressor issues, including:

- compressor clashing between stator and rotor blades,
- compressor rubbing between rotor blades and casing, and
- foreign or domestic object damage (FOD/DOD).

6 Conclusions

Integrated monitoring systems allow for the sharing of findings and experiences between interested parties providing a common data base that may be consulted at any time by all Petrobras power plants. Processed data can also be used (along with predictive tasks) to help in many decisions taken by local staff (e.g. machine fault diagnosis, continuous improvement of maintenance tasks and so on) which can help to streamline such a process.

The CMD tool is a large platform that can provide solid data to a better understanding of the machines behavior. As more people use it, the historical database will become larger (and thus more useful) and this can then provide more reliable data in the ongoing decision making process.

The case studies previously presented have been developed by Petrobras and show the advantages of applying an integrated monitoring strategy and predictive maintenance for effective machine maintenance in power plants. Other models that use health-monitoring techniques (including early fault detection and advanced real time diagnosis) are currently in development to provide technicians key information to improve turbomachinery and the overall combined cycle performance.

References

1. J. Moubray, *Introdução à Manutenção Centrada na Confiabilidade* (Aladon, São Paulo, Brazil, 1996)
2. A. Kardec, J. Nascif, T. Baroni, *Gestão Estratégica e Técnicas Preditivas* (Editora Quality Mark, Rio de Janeiro, 2002)
3. R.C.D. Oliveira, B.Q. Lima, Monitoramento e Diagnóstico Remoto de Usinas Termelétricas, in *Proceedings of XIII Encontro para Debates de Assuntos de Operação* (Belo Horizonte, Brazil, 2014)
4. H. Andersen, B. Lima, Real-time fleet conditioning and performance monitoring, Power Generation Brazil (2014)
5. Electric Power Research Institute, Assessment and development of gas turbine compressor health monitoring technologies, Report 2014

Stiffness Nonlinearity in Structural Dynamics: Our Friend or Enemy?

Michael John Brennan

Abstract The effects of nonlinearity, particularly stiffness nonlinearity, has been of concern to structural engineers for many years. Primarily this has been because this type of nonlinearity can cause unpredictable dynamics, and considerable effort is necessary to analyze nonlinear structures. Due to the constant drive to improve the performance and efficiency of structures and mechanical devices, engineers have recently started to investigate whether nonlinearity can be incorporated into structures to provide some benefit. This chapter discusses some of the problems that stiffness nonlinearity can cause, and gives three examples where this type of nonlinearity can be put to good use. The examples are in vibration isolators, vibration absorbers and energy harvesters. If the nonlinearity is introduced in an appropriate way then it should not have an adverse impact on probabilistic prognostics and health management of energy systems.

Keywords Nonlinear vibrations · Vibration isolation · Energy harvesting

1 Introduction

For many years, engineers have sought to eliminate nonlinearity in stiffness elements, at least from the point of view of structural dynamics, mainly because this type of nonlinearity can cause unpredictable dynamics, and nonlinear structures can be difficult to analyze [1]. In recent years, however, there have been attempts to harvest the beneficial effects of stiffness nonlinearity, for example [2]. This has been driven by the need to improve the performance of structures by making them compact, and without adding weight. Greater understanding of the effects of nonlinearity and improved prediction methods have facilitated this. In this chapter,

M.J. Brennan (✉)
Department of Mechanical Engineering, São Paulo State University (UNESP), Ilha Solteira, SP, Brazil
e-mail: mjbrennan0@btinternet.com

S. Ekwaro-Osire et al. (eds.), *Probabilistic Prognostics and Health Management of Energy Systems*, DOI 10.1007/978-3-319-55852-3_17

three systems in which nonlinearity in the stiffness is deliberately introduced, are discussed. The systems involve vibration isolation [3], vibration control using absorbers/neutralizers [4], and energy harvesting from ambient vibrations [5].

2 Stiffness Nonlinearity

The stiffness of a system or structure is often the cause of nonlinearity, and this can take the form of a softening or hardening stiffness, a bilinear stiffness, a clearance, or a saturation [1]. The type of stiffness considered here is the hardening type as this occurs often in practice and it is relatively easy to design a structure with this type of nonlinearity. The displacement frequency response curve for a simple oscillator that has a softening, linear and hardening stiffness is shown in Fig. 1.

In Fig. 1, the non-dimensional frequency is the excitation frequency divided by the natural frequency of the linear system. It can be seen that for the linear system, for each excitation frequency there is a unique displacement response (i.e., it is single-valued), with a clear resonant peak. However, for the softening and the hardening systems, the frequency response curve bends to the left and to the right respectively, resulting in the displacement response being multi-valued (three values) in certain frequency regions (the dashed lines in the figure denote unstable solutions that cannot be reached in practice). The net effect of this behavior is that large jumps in the vibration can occur at specific frequencies [6]. These jumps, which occur as frequency is either increased from low to high frequency or decreased from high to low frequency, are denoted by black arrows in Fig. 1, and can be dangerous or can cause damage. Furthermore, unpredictable chaotic behavior can occur in such systems [7]. It is for these reasons that engineers have tried to eliminate nonlinearity from structural design. However, incorporating

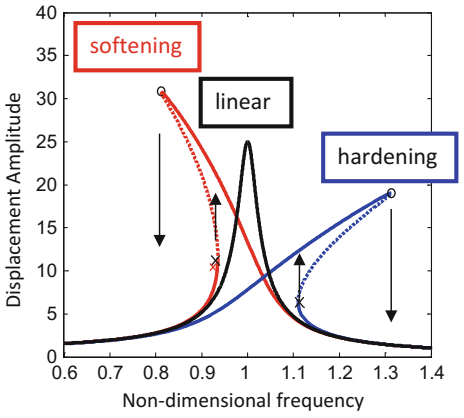


Fig. 1 Frequency response curve of a nonlinear oscillator. The *arrows* indicate nonlinear jumps

nonlinearity in some structures can offer advantages. The engineering challenge is to make use of these advantages while minimizing the undesirable dynamic effects discussed above.

3 Nonlinear Vibration Isolators

A recurrent problem in many engineering applications is that of preventing the transmission of vibrations using a vibration isolator [8]. Ideally, an isolator should have a high-static (HS) stiffness—capable of bearing the load with little static displacement, and a low-dynamic stiffness (LDS) so called HSLDS isolators. These provide the low natural frequency required for improved vibration isolation performance. Such a characteristic requires stiffness nonlinearity [9]. An example of an isolator and the force-deflection characteristics for several different models of the isolator (a bubble mount) can be seen in Fig. 2.

The static equilibrium position is marked in Fig. 2b. It can be seen that the slope of the graphs (the local or dynamic stiffness) is small at this position. Simple models of such an isolator and the optimum parameters to maximize the benefits of the nonlinearity and to minimize undesirable dynamic effects for these types of isolators have been studied by the author and co-workers [3, 10–12]. Further work in this area has involved the study of using magnetics in an isolator [13]. The asymmetry of the isolator about the static equilibrium position can be seen in Fig. 2b. This is often found in such isolators, and has been studied in [14].

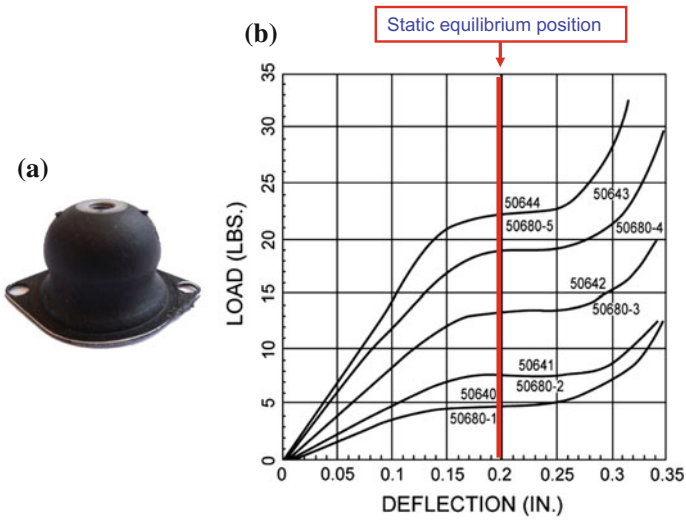


Fig. 2 A typical nonlinear isolator (http://www.novibes.com) **a** rubber isolator, **b** force-deflection characteristic

Essentially, the type of nonlinear isolator shown in Fig. 2 can increase the frequency range over which isolation occurs, extending it to very low frequencies. The undesirable dynamic effects can be avoided if the excitation forces are relatively small compared to the static loading on the isolator and the damping in the isolator is high enough to avoid the occurrence of nonlinear jumps.

4 Nonlinear Vibration Absorbers/Neutralizers

Vibration absorbers and vibration neutralizers are tuned mass-spring-damper devices that are attached to structures to reduce vibration in a specific frequency range. A vibration absorber is designed to reduce the resonance response of the host structure, and a vibration neutralizer is used to reduce the vibration of the host structure at a particular forcing frequency. Whilst it does not seem that there are significant advantages in using a nonlinear vibration absorber compared to a linear vibration absorber [15], there are some advantages in using a nonlinear vibration neutralizer [4]. For such a device with a hardening stiffness nonlinearity, a plot of the change in vibration level of a mass-like structure to which it is attached, as a function of frequency, is shown in Fig. 3. For comparison, the effect that a linear neutralizer would have is shown as a black dashed-dotted line. The frequency is normalized to the frequency at which the neutralizer is tuned to. It can be seen that close to the tuned frequency, the nonlinearity in the neutralizer has little effect. However the nonlinearity has a profound effect close to the resonance frequency. It has the beneficial effect of shifting this peak to much higher frequencies, away from the tuned frequency. In a linear neutralizer, this effect can only be achieved by adding mass to the neutralizer. Thus, the inclusion of nonlinearity, in this case, has the benefit of saving weight.

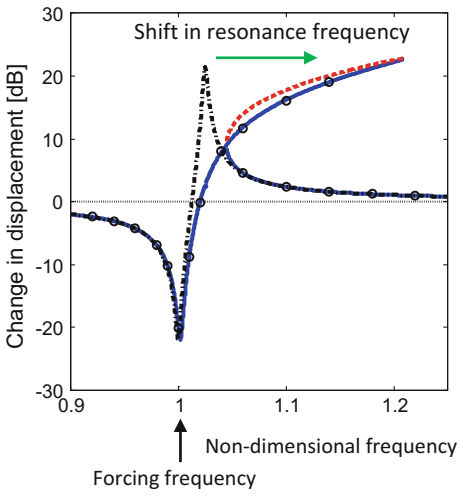


Fig. 3 The effect that a nonlinear neutralizer has on the change in displacement of a mass-like structure. The dashed-dotted line is for a linear neutralizer

5 Nonlinear Energy Harvesters

Harvesting energy from ambient sources has become an area of increasing interest within the last decade or so, particularly with the increase in the use of wireless sensors, which often require autonomous power supplies. Among the sources, ambient vibration has the potential to power these sensors in remote and hostile environments. Many energy harvesting devices are linear mass-spring-damper systems in which the devices are tuned so their natural frequencies coincide with particular forcing frequencies allowing maximum energy to be harvested [16]. However, the ambient frequency may not be tonal and could vary with time, which can degrade the performance of the device drastically. To overcome this limitation, nonlinear energy harvesters have been proposed [5]. Such devices can improve the bandwidth by using a hardening spring or by creating a bi-stable device [17]. An example of such a device is shown in Fig. 4a.

The positive stiffness in the system is provided by the steel beam and the magnets provide negative stiffness. The resulting potential energy characteristic of the system is shown in Fig. 3b. It can be seen that the shape of the potential energy

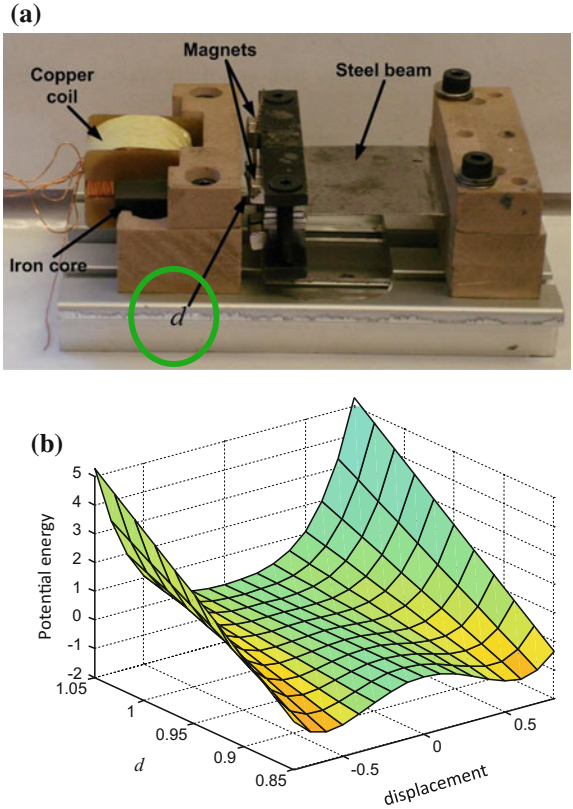


Fig. 4 Two-mode electro-dynamic energy harvester **a** Two-mode energy harvester. **b** Potential energy in the energy harvester

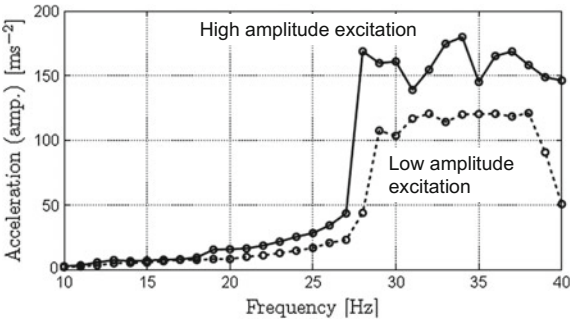


Fig. 5 Acceleration of the mass of the two-mode electro-dynamic energy harvester shown in Fig. 4 working in bi-stable mode

as a function of the displacement is governed by the gap between the magnets. If the gap is large then the device has a hardening stiffness and if the gap is small then the device is bi-stable. If the device is tuned so that it acts as an oscillator with a hardening stiffness, then the system will behave as shown in Fig. 1, with a frequency response curve that bends to higher frequencies. This can achieve a wider bandwidth of operation compared to a linear energy harvester, but there are issues in achieving this in practice. A better configuration is when the stiffness is adjusted so that the system behaves as a bi-stable system, which has recently attracted considerable attention in the literature [18]. Essentially, this device is useful when there is low frequency excitation and it is difficult to design a system with a low natural frequency to match this frequency. The device has the advantage that once there is a strong enough level of excitation to cause it to vibrate in its bi-stable mode, then it will continue to vibrate in this way over a wide range of frequencies. Thus, it does not suffer from tuning issues that affect many designs of energy harvesters. An example of the vibration output for a two levels of excitation for the device in Fig. 4 is shown in Fig. 5.

6 Conclusions

This chapter has described the concept of introducing nonlinearity to improve the performance of some mechanical situations. Three examples have been described: in vibration isolators, vibration absorbers and energy harvesters. In all three cases, it has been shown that the nonlinearity can have a beneficial effect. However, it can also have undesirable dynamic effects, and so it has to be implemented carefully to minimize or avoid these effects. If this is done correctly, then nonlinearity can be the engineer's friend rather than his enemy. Moreover, if these concerns are attended to then the nonlinearity incorporated into an energy system then it should not have an adverse impact on probabilistic prognostics and health management of such systems.

References

1. K. Worden, G.R. Tomlinson, *Nonlinearity in Structural Dynamics: Detection, Identification and Modelling* (Institute of Physics Publishing, Bristol and Philadelphia, 2001)
2. D. Wagg, S. Nield, *Nonlinear Vibration with Control, for Flexible and Adaptive Structures* (Springer, Dordrecht, 2010)
3. A. Carrella, M.J. Brennan, T.P. Waters, V. Lopes Jr., Force and displacement transmissibility of a nonlinear isolator with high-static-low-dynamic-stiffness. Int. J. Mech. Sci. **55**, 22–29 (2012)
4. G. Gatti, M.J. Brennan, The characteristics of a nonlinear vibration neutralizer. J. Sound Vib. **331**(13), 3158–3171 (2012)
5. R. Ramlan, M.J. Brennan, B.R. Mace, I. Kovacic, Potential benefits of a non-linear stiffness in an energy harvesting device. Non-linear Dyn. **59**(4), 545–558 (2010)
6. M.J. Brennan, I. Kovacic, A. Carrella, T.P. Waters, On the jump-up and the jump-down frequencies of the Duffing oscillator. J. Sound Vib. **318**(4-5), 1250–1261 (2008)
7. J.J. Thomsen, Vibrations and Stability, in *Advanced Theory, Analysis and Tools* (Springer, Germany, 2003)
8. E.I. Rivin, *Passive Vibration Isolation* (ASME Press, 2001)
9. R.A. Ibrahim, Recent advances in nonlinear passive vibration isolators. J. Sound Vib. **314**, 371–452 (2008)
10. A. Carrella, M.J. Brennan, T.P. Waters, Static analysis of a passive vibration isolator with quasi-zero stiffness characteristic. J. Sound Vib. **301**(3-5), 678–689 (2007)
11. I. Kovacic, M.J. Brennan, T.P. Waters, A study of a non-linear vibration isolator with quasi-zero stiffness characteristic. J. Sound Vib. **315**(3), 700–711 (2008)
12. A. Carrella, M.J. Brennan, I. Kovacic, I.T.P. Waters, On the force transmissibility of a vibration isolator with quasi-zero stiffness. J. Sound Vib. **322**(4-5), 707–717 (2009)
13. A. Carrella, M.J. Brennan, T.P. Waters, K. Shin, Introducing stiffness nonlinearity using magnets to improve vibration isolation. J. Sound Vib. **315**(3), 712–720 (2008)
14. A. Abolfathi, M.J. Brennan, T.P. Waters, B. Tang, On the effects of mistuning a force-excited system containing a quasi zero stiffness vibration isolator. Trans. ASME, J. Vib. Acoust. **137** (4), 044502 (6 pages) (2015)
15. N.A. Alexander, F. Schilder, Exploring the performance of a nonlinear tuned mass damper. J. Sound Vib. **319**(1-2), 445–462 (2009)
16. C.B. Williams, R.B. Yates, Analysis of a micro-generator for microsystems, in *Proceedings of 8th International Conference on Solid-State, Sensors and Actuators* (1996), pp. 8–11
17. R. Ramlan, M.J. Brennan, B.R. Mace, S.G. Burrow, On the performance of a dual-mode non-linear vibration energy harvesting device. J. Intell. Mater. Syst. Struct. **23**(13), 1423–1432 (2012)
18. R.L Harne, K.W. Wang, A review of the recent research on vibration energy harvesting via bi-stable systems. Smart Mater. Struct. **22**, 023001 (12 pp) (2013)