TECHNICAL REPORT

Techniques of Prognostics for Condition-Based Maintenance in Different Types of Assets

Ángel Manuel Hernández Mejías
Diego Galar

LULEÅ UNIVERSITY OF TECHNOLOGY

Techniques of Prognostics for Condition-Based Maintenance in Different Types of Assets

Ángel Manuel Hernández Mejías
Diego Galar

2014

Technical Report in Maintenance Engineering

Luleå University of Technology
Department of Civil, Environmental and Natural Resources Engineer

Printed by Luleå University of Technology, Graphic Production 2014

ISSN 1402-1536
ISBN 978-91-7439-980-6 (print)
ISBN 978-91-7439-981-3 (pdf)

Luleå 2014

www.ltu.se

ACKNOWLEDGEMENTS

The research presented in this work was done between February 2013 and July 2014. The investigation was developed at the Division of Operation and Maintenance Engineering at Luleå Railway Research Centre at Luleå University of Technology, within the European Union (EU); the project was based on prognosis techniques for the condition-based maintenance of different industrial assets. First of all, I thank God for allowing me to stay here and giving me the health and fitness necessary to finalise this Master's Thesis.

Then, principally, I would like to thank Professor Diego Galar for his specialised technical advice and recommendations, which were a great help in completing this thesis. Besides, I must thank the Vicente Macian and Bernardo Tormos professors, for giving their knowledge to support this work. I also want to thank my colleagues at the Department of Operation and Maintenance Engineering; Emilio Rodriguez, Victor Simon, Roberto Villarejo, Numan Perales and Carl Johansson gave me their support and collaborated with me to realise this project.

Finally, I thank my parents, Manuel Hernandez and Yamilesca Mejias, who are always with me despite the distance. In like manner I want to thank my beloved wife Katyuska Poliandri for her support and for always being on my side. In one way or another, they have participated in preparing and performing this research.

Angel Manuel Hernandez Mejias
July, 2014, Luleå, Sweden

ABSTRACT

Today, many maintenance programs in the industrial sector rely on condition-based maintenance (CBM). This type of program helps improve maintenance tasks because machines or equipment are continuously monitored. Condition-based maintenance recommends maintenance decisions based on information collected through condition monitoring. It consists of three main steps: data acquisition, data processing and maintenance decision-making.

Prognostics is a key feature of today's maintenance strategies; it prevents inopportune maintenance spending, because with prognostics, we can estimate the remaining useful life (RUL) and minimise maintenance tasks.

Real prognostic systems are scarce in industry. For one thing, it is difficult to choose an efficient technology, as there are many possible approaches: model based, data driven and experience based. The applicability of each is dependent on industrial constraints. Thus, the general purpose of the present work is to review the various techniques of prognosis for different industrial assets. It investigates each approach to determine which techniques are applicable to different assets (rotating machines, esructutas and complex systems). Finally, it compares the approachs and their respective techniques in a table.

Keywords: Condition-based maintenance (CBM), Prognosis, Remaining useful life (RUL), Maintenance.

Table of contents

Part I
SUMMARY

Introduction 1

1.1. Background.

The maintenance of all types of assets is increasingly important in both the industrial and the scientific sectors. Industries want to improve their maintenance techniques in order to increase the lifetime of their equipment. All equipment deteriorates over time, as it operates under a certain voltage or load in the real environment, thus generating maintenance activities more often.

Maintenance activity combines various methods, tools and techniques in a bit to reduce maintenance costs while increasing reliability, availability and security of equipment [2]. The most common types of maintenance are corrective maintenance (also called unplanned maintenance, or run-to-failure maintenance), which takes place only at breakdowns. Time-based preventive maintenance (also called planned maintenance) sets a periodic interval to perform preventive maintenance regardless of the health status of a physical asset. With the rapid development of modern technology, products have become increasingly complex while better quality and higher reliability are required. This raises the cost of preventive maintenance. Preventive maintenance has become a major expense of many industrial companies [3]. Therefore, many are now implementing condition-based maintenance (CBM). CBM is designed to avoid unnecessary maintenance tasks. Techniques for monitoring condition include:

- ✓ Temperature Control: use of contact thermometers, infrared, thermographic;
- ✓ Dynamic Monitoring: control the energy emitted by mechanical equipment, such as Vibration Analysis, etc.;
- ✓ Analysis of Oils: test the quality of any type of oil, whatever its function, i.e. lubricating oils, hydraulic oils, insulating oils;
- ✓ Performance Supervision: compare nominal data to real time flows, pressures, times, temperatures, voltage.

A very important element of the maintenance strategy is called prognostics. Prognostics is of great interest to both industry and research centres as it can significantly improve the efficiency of a CBM program. Prognostics basically tries to predict how much time remains before a fault or failure will occur, given the current situation of the asset and the operation. In other words, prognostics is based on predicting the remaining useful life (RUL) of a system before a failure occurs.

This topic is of great interest since with prognostics, industries can potentially reduce the costs of both preventive and corrective maintenance. If they can predict when a machine or asset may fail, they will be able to prevent or limit maintenance activities.

Prognostics is mainly based on mathematical models for predicting the remaining useful life. These models are built using one of four approaches: experience based, data based, model based and hybrid methodology (discussed later) Figure 1 illustrates the hierarchy of possible approaches in relation to their applicability and relative costs.

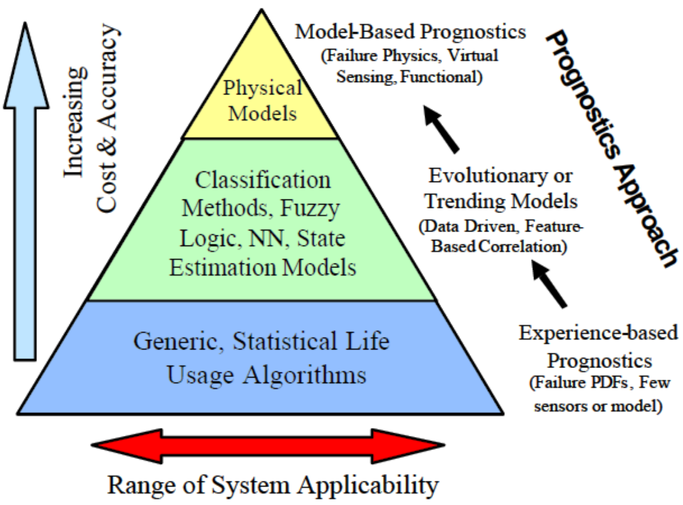


Figure 1. Hierarchy of Prognostics Approaches. [13]

A typical example in the application of prognostics is the use of models for modelling fatigue initiation and propagation of cracks in structural components. In this case, a model-based approach would be appropriate. Other examples include the study of the estimation of RUL for bridge structures using a system based on experience or a study of vibrations in rotating machines using a data-driven approach. The various approaches and their application in indusry will be explained in more detail in the following chapters.

1.2. Research Goal.

The objective of this work is to compare the different techniques and models used to estimate the remaining useful life (RUL) of different assets. It seeks to determine if a technique used in an asset in a particular area may be also used in other assets. It conductes a review of the relevent literature on prognostics techniques their use in the industrial sphere.

Part II
THEORETICAL FOUNDATION

2.1. History of Industrial Maintenance.

Since the industrial revolution, maintenance has gone through a number of different stages. At the beginning of the industrial revolution, the workers were responsible for the repair of equipment. As machines became more complex and the amount of repair work increased, the first maintenance departments, distinct from production departments, were established. Tasks in both eras were basically remedial, with efforts devoted to addressing failures.

The concept of reliability appeared in the 20^{th} century after the end of World War II. Maintenance departments now sought not only to fix the faults but to prevent them. This involvee creating a new figure in maintenance departments: staff whose purpose was to study what maintenance should be performed to avoid failures. With the advent of the computer age, maintenance changed again, with the introduction of new techonolgies. For example, Reliability Based Maintenance (RCM) is based on the analysis of failure modes and the application of statistical techniques and detection technology. It is basically a philosophy of technology maintenance. Total Productive Maintenance (TPM) is another recent concept; it refers to tasks normally performed by maintenance personnel that are now performed by production workers [4]. Finally, condition based maintenance (CBM) is a widely used concept which we will discuss later [5].

Clearly, maintenance techniques have evolved. Not all maintenance models apply to all companies. Some models interact; others do not. But all have been adapted to new uses in industry. Today, the specific needs of each team and each industry will determine the most appropriate maintenance model to optimise resources and needs.

2.2. Types of Maintenance.

2.2.1. Corrective Maintenance.

Corrective maintenance is used to repair damage that has already occurred. Usually, when this type of maintenance is performed, the manufacturing process is stopped, decreasing production and increasing costs. Repair time cannot be predicted, nor can the expenses resulting from the breakdown and consequent disturbances on the production line.

Therefore, corrective maintenance is applied on assets with low criticality, whose faults do not involve large temporal or economic problems. It is often used for specific equipment where other techniques would be more costly.

2.2.2. Preventive Maintenance.

Preventative maintenance is planned in a time horizon and aims to prevent breakdowns. Unlike corrective maintenance, because it is planned, it is not done during production time.

The intention of this type of maintenance is to reduce the number of corrective interventions, performing periodic reviews and replacing worn components.

It is a demanding type of maintenance, as it requires strict supervision and development of a plan to be carried out by qualified personnel. In addition, as it involves routine tasks, personnel may not be motivated. Furthermore, if it is not done correctly, there will be a cost overrun with no significant improvements in productivity.

2.2.3. Predictive Maintenance.

As its name suggests, predictive maintenance is done before faults appear. It requires the application of tools or techniques to detect certain measurable elements prior to failure, such as wear. The goal is to do the right maintenance at the right time. Predictive maintenance uses technology based on indicators to measure the variables that allow the machine to operate; it also requires personnel trained in the interpretation of the data[1].

2.3. Condition-Based Maintenance (CBM).

Condition-based maintenance (CBM) aims to determine the condition of equipment, so that operation remains safe, efficient and economic. Monitoring techniques are aimed at measuring physical variables which indicate the condition of the machine and to compare these with normal values to determine if they are in good condition or are deteriorating. CBM assumes there are measurable and observable characteristics that are indicators of the condition of the machine.

Condition monitoring studies the evolution of selected time-dependent parameters; it identifies trends indicating the existence of a fault, its severity and the likely time to failure. Timely decision-making avoids the occurrence of faults and eliminates the possibility of catastrophic failure. CBM can be performed while the machine is running [6].

CBM consists of three key steps (see Figure 2):

1. Data acquisition (information collecting) to obtain data relevant to system health.

2. Data processing (information handling) to handle and analyse the data or signals collected in step 1 for better understanding and interpretation of the data.

3. Maintenance decision-making (decision-making), to recommend efficient maintenance policies.

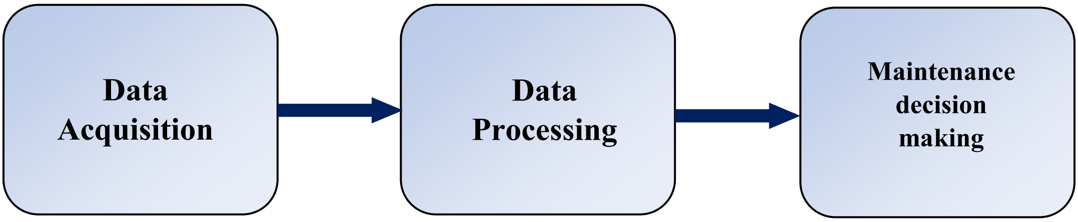


Figure 2. Three Steps in CBM.

2.3.1. Steps to Implement CBM.

2.3.1.1. Data Acquisition.

Data acquisition is a process of collecting and storing useful data (information) from targeted physical assets for the purpose of CBM. This is an essential step in implementing a CBM program for machinery faults (or failure, usually caused by one or more faults). Data collected in a CBM program can be categorised into two main types: event data and condition monitoring data. Event data include information on what happened (e.g., installation, breakdown, overhaul, etc., along with the causes) and/or what was done (e.g., minor repair, preventive maintenance, oil change, etc.) to the targeted physical asset. Condition monitoring data are measurements related to the health condition/state of the physical asset. These are very versatile and can include vibration, acoustic, oil analysis, temperature, pressure, moisture, humidity, weather or environment data, etc.

2.3.1.2. Data Processing.

Data processing consists of two stages. The first is data cleansing; this step is important because usually the data are entered manually. This leads to frequent errors, requiring data cleansing to increase the probability that the data are clean (no errors). The second step is the analysis of the data. There are a variety of models, algorithms and tools for analysis; selection depends on the types of data collected.

2.3.1.3. Maintenance Decision Support.

The last step of a CBM program is maintenance decision-making. Sufficient and efficient decision support is crucial for determining maintenance actions. Techniques for maintenance decision support in a CBM program can be divided into two main categories: diagnostics and prognostics. Fault diagnostics focuses on detection, isolation and identification of faults when they occur. Prognostics attempts to predict faults or failures before they occur. Obviously, prognostics is superior to diagnostics in the sense that prognostics can either prevent faults or failures, or be ready (with spare parts and human resources) for the problems, thus reducing the costs of unplanned maintenance. Nevertheless, prognostics cannot completely replace diagnostics since, in practice, there are always some faults and failures which are not predictable. In addition, prognostics, like any other prediction technique, cannot be 100% accurate. In the case of unsuccessful prediction, diagnostics can be a complementary tool for maintenance decision-making. Diagnostics are also helpful for improving prognostics; diagnostic information can result in more accurate event data, and a better CBM model can be built for prognostics. Furthermore, diagnostic information can be used as feedback information for system redesign.

The techniques listed above are used today to assist in making decisions on when and how to do maintenance on a machine or asset class to improve the planning of maintenance and reduce the cost [3] .

2.3.2. Types of CBM.

There are two ways to implement CBM: first, timely diagnosis (offline) and monitoring (online); second, diagnosis with guarantees of operation. The first requires rigorous planning; a specific program is carried out in a timely manner on each piece of equipment or machinery to be monitored, a history is established, and the necessary corrections are made. The second introduces diagnostic equipment for continuous condition monitoring. This means that while machines and equipment are functioning all parameters can be observed in real time [6].

2.3.3 Techniques Applied in the CBM.

A range of techniques have been developed for CMB, as described below.

2.3.3.1. Vibration Monitoring:

Vibration monitoring techniques can be used to detect fatigue, wear, imbalance, misalignment, loosened assemblies, turbulence, etc. in systems with rotational or reciprocating parts, such as bearings, gear boxes, shafts, pumps, motors, engines and turbines. The operation of such mechanical systems releases energy in the form of vibration with frequency components which can be traced to specific parts in the system. The amplitude of each distinct vibration component will remain constant unless there is a change in the operating dynamics of the system.

Vibration can be characterised in terms of three parameters: amplitude, velocity and acceleration. The sensitivity of senors used for measuring these parameters varies with the frequency of the vibration. The general selection guideline is to use amplitude senors to pick up low frequency signals, velocity senors in the middle ranges, and accelerometers at higher frequencies. In one form of vibration monitoring, readings of the overall vibration energy between 10 to 10,000Hz are taken from selected points on a machine. These data are compared to baseline readings taken from a new machine. Alarm limits are established on the basis of the baseline readings. A fault diagnosis will be triggered when a reading exceeds its alarm limit. Alternatively, vibration readings are compared to vibration severity charts to determine the relative condition of the machine. This approach is known as broadband vibration trending, and it monitors only the overall machine conditions. The common microprocessor-based instrumentation for this procedure monitors the rootmean-square (RMS) level of the vibration.

In the narrowband trending technique, the total energy across a specific bandwidth of vibration frequencies is tracked to monitor the health condition of specific machine components or failure modes. The process of scanning vibration signals across a bandwidth captures vibration data in the time domain. Such data can be transformed into the frequency domain so that the vibration at each frequency component can be measured. The frequency plot providing a visual representation of each frequency component generated by a machine is called the machine's vibration signature. When the vibration signatures of a machine at different times are arranged in chronological order and shown in a cascading manner on a three-dimensional plot, a waterfall plot of the machine is formed. Anomalies in the machine's condition can be easily detected by noting that the vibration signatures have changed with time.

2.3.3.2. Thermography:

Thermography uses instrumentation designed to measure emissions of infrared energy to determine the operating condition of plant machinery. Anomalies of thermal conditions, such as equipment being hotter or colder than it should be, are taken as alarm signals of potential problems within the system. Thermographic techniques are most appropriate to detect problems found in systems which rely on heat transfer or retention.

Infrared thermometers are designed to measure the surface temperature at a single point on a machine surface. They can be used to monitor the temperature of critical parts of plant machinery, such as bearing caps and motor windings, and to spot check process piping systems. When the infrared emission profile of a large area needs to be scanned within a short period of time, infrared imaging is the applicable technique. The imaging system functions much like a video camera, and the thermal profile of the observed area can be viewed through the instrument's optics.

The measurement of infrared emissions is very sensitive to variations of ambient conditions, such as the amount of airborne particles. Therefore, extra care must be taken to compensate for the effect of such factors when capturing thethermal data.

2.3.3.3 Tribology:

Tribology is the field of study relating to the interface between sliding surfaces. Three tribology techniques are used in condition-based maintenance: lubricating oil analysis, wear particle analysis and ferrography. These techniques are relatively slow and expensive because the analysis requires the use of laboratory facilities such as spectrometers and scanning electron microscopes (SEM). In lubricating oil analysis, samples of lubricating, hydraulic, and dielectric oils are analysed at regular intervals to determine if they still meet the lubricating requirements of their application. When the oil condition reaches an unacceptable state, it will be replaced to maintain satisfactory system operation. Results of the analysis may also form the basis for decisions to change the type of oil to improve performance or reduce variety. Lubricating oil analysis involves the use of spectrographic techniques to analyse the elements contained in the oil sample. However, it must be supplemented with other diagnostic procedures to identify the specific failure mode which may have caused the observed degradation of the oil.

Wear particle analysis provides direct information about the wearing condition of the machine. This information is derived from the study of particle shape, composition, size and quantity. There are two basic types of wear particle analysis. In the first type, the solids content of the machine lubricant, the quantity, size, and composition of particulate matter in the lubricating oil, is routinely monitored to detect changes over time. In the second, particulate matters in each lubricating oil sample are analysed to identify the type of wear (rubbing wear, cutting wear, rolling fatigue, combined rolling and sliding wear, or severe sliding wear) found in the sample. Wear particle analysis using spectrographic techniques is limited to the study of particulate matters with a size not exceeding 10mm [7].

2.3.4. Advantages and Disadvantages of CBM.

2.3.4.1. Advantages:

- ✓ Reduces outages due to failures and breakdowns.
- ✓ Reduces maintenance costs by allowing achieve maximum life cycle of equipment and machinery, with no unnecessary parts required.
- ✓ Substantially increases indicator reliability of equipment and machinery.
- ✓ Reduces stress caused by constant emergencies.

2.3.4.2. Disadvantages:

- ✓ High initial cost of implementing this system.
- ✓ High initial cost of equipment and diagnostic tools.

- ✓ Specialised personnel required to operate equipment and instruments.
- ✓ Constant training required in techniques of interpretation and diagnosis[6].

2.3.5. Typical Degradation Process of Equipment.

Following a period of normal operation when an item has been running smoothly, a change may occur that affects its performance. The earliest time to detect the degradation is called the P-point, or potential failure point, as after this time, the item can potentially fail at any time. This change occurs gradually, or rapidly in some cases, and worsens to the point where the equipment cannot reliably and safely do its job. At this stage, the item has functionally failed, i.e. it is not delivering its required performance. This is called the functional failure point or F-point. If the item continues to operate, the part will fail and work will cease (see Figure 3) [8].

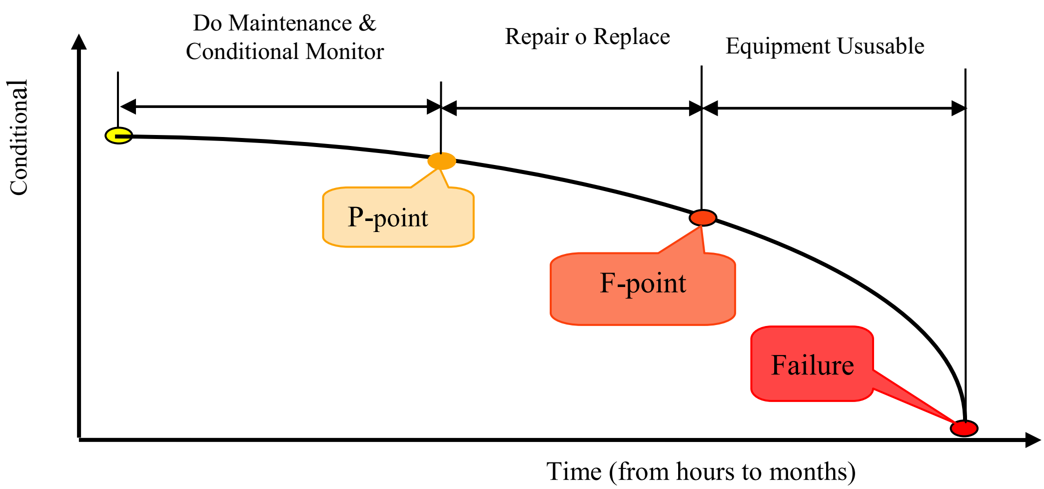


Figure 3. Degradation Process Experienced By Equipment.

2.4. Prognostics.

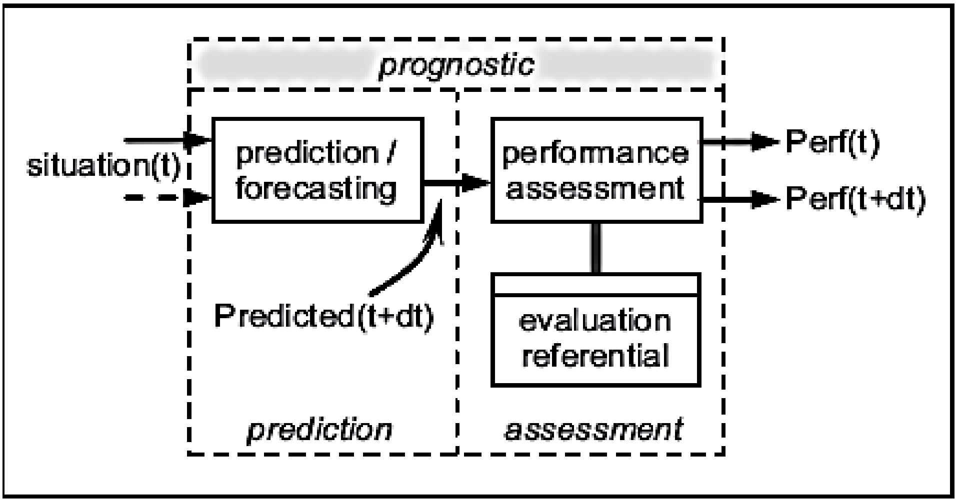
Maintenance activity combines various methods, tools and techniques to reduce maintenance costs while increasing reliability, availability and security of equipments. We usually speak about fault detection, failure diagnostics, response development (choice of management actions, i.e. preventive and/or corrective) and scheduling actions. Briefly, these steps correspond to the need, first, to "perceive" phenomena, second, to "understand" them, and third, to "act" correctly. However, rather than understanding a phenomenon which appears as a failure (*a posteriori*), it may be convenient to "anticipate" its manifestation to quickly resort to protective actions. This could be defined as the "prognostic process" [2].

Prognostics is a very promising maintenance activity as it should permit plants to improve safety, plan and schedule successful maintenance, and reduce maintenance cost and down time [9].

2.4.1. Concept of Prognostics.

Prognostics is an engineering discipline focused on predicting the time at which a system or a component will no longer perform its intended function. This lack of performance is most often a failure beyond which the system can no longer be used to meet its desired goals. Thus, the predicted time becomes the Remaining Useful Life (RUL) of the component, an important concept in decision-making for contingency mitigation. Prognostics predicts future performance by assessing the extent of deviation or degradation of a system from its expected normal operating conditions. The science of prognostics is based on the analysis of failure modes, detection of early signs of wear and aging, and fault conditions. These signs are correlated in a damage propagation model. Prognostics can potentially be used in condition-based maintenance [10].

Prognostics should be based on assessment criteria; its limits depend on the system itself and on performance objectives. Prognostics can be split into two sub-activities: first, to predict the evolution of a situation at a given time; second,



to assess this predicted situation using an evaluation referential [11] (see Figure 4).

Figure 4. Prognostics as a Prediction and Assessment Process. [11]

2.4.2. Remaining Useful Life (RUL).

A main concept of prognostics is the RUL, the output generated by a prognostic algorithm describing the distribution in time of likely equipment failure times. Figure 5 illustrates the key concepts of a RUL. At time $\mathbf{t_P}$, a prediction is made and an estimate of the RUL is generated. Once the RUL has been generated, the next question is when to carry out corrective maintenance. Ideally, the time chosen for maintenance action will avoid equipment failure and maximise the useful life of the equipment. However, these are conicting requirements and, as a consequence, selecting when to perform maintenance is typically an exercise in risk management.

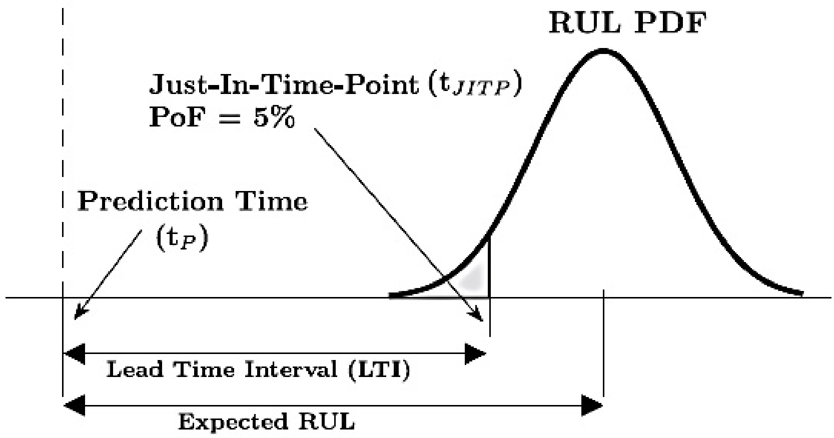


Figure 5. The Remaining Useful Life. [14]

To develop the requirements for a prognostic algorithm, the maximum allowable probability of failure (PoF) must be considered. This value defines the maximum acceptable level of risk of equipment failure, beyond which equipment can no longer be operated as the risk of equipment failure is deemed excessive. Using the defined maximum allowable PoF and the estimated RUL, an important value known as the just-in time-point (JITP) can be identified. The JITP defines the latest point in time before which corrective maintenance actions must be carried out to avoid operating equipment beyond the maximum allowable PoF.

In a real-life applications, selecting the maximum allowable PoF usually requires the consideration of a number of factors, including safety, criticality and economic concerns. In certain scenarios, where safety considerations are primary, the requirement might be to avoid as many in-service failures as possible; thus, a conservative value for the maximum allowable PoF might be chosen. Alternatively, a plant operator may accept a higher maximum allowable PoF value in situations where maximising the useful life of expensive equipment/components might be more economical than avoiding an occasional failure. An example of such a scenario might be the use of a diamond headed cutting tool.

In Figure 5, a maximum allowable PoF value of 5% is assumed for illustrative purposes. Once the JITP has been identified, he lead-time interval (LTI). Can be computed. The LTI is defined as the time interval between the time the prediction is generated, $\mathbf{t_P}$, and the JITP $\mathbf{t_{JITP}}$, so that[15]

$$t_{LTI} = t_{JITP} - tp \tag{2.1}$$

Another way to estimate the RUL is as follows:

$$T - \frac{t}{T} > t, Z(t), \tag{2.2}$$

where T denotes the random variable of time to failure, t is the current age and $Z(t)$ is the past condition profile up to the current time. Since RUL is a random variable, the distribution of RUL would be of interest to fully understand the RUL. In the literature, ‘‘remaining useful life estimate” (RULE) has two possible meanings. In some cases, it means finding the distribution of RUL. In other cases, it simply means the expectation of RUL [3]:

$$E\left[T - \frac{t}{T} > t, Z(t), \right] \tag{2.3}$$

2.4.3. Technical Approaches.

Technical approaches to building forecasting models can be classified broadly into data-based approaches and model-based approaches.

2.4.3.1. Model-Based Approaches.

Model-based prognostics attempts to incorporate physical understanding (physical models) of the system into the estimation of the RUL. Modelling can be accomplished at different levels, for example, micro and macro levels. At the micro level (also called material level), physical models are based on a series of dynamic equations that define relationships at a given time or load cycle, between damage (or degradation) of a system/component and environmental and operational conditions under which the system/component are operated. The micro-level models are often called damage propagation models.

Macro-level models are mathematical models at the system level; they define the relationship among system input variables, system state variables, and system measures variables/outputs. The model is often a somewhat simplified representation of the system, for example, a lumped parameter model [12].

Generally, model-based methods assume an accurate mathematical model can be constructed from first principles. These methods often use residuals as features; the residuals are the outcomes of consistency checks between the sensed measurements of a real system and the outputs of a mathematical model. The premise is that the residuals are large in the presence of malfunctions, and small in the presence of normal disturbances, noise and modelling errors [2].

Advantages and Drawbacks.

The main advantage of model-based approaches is their ability to incorporate physical understanding of the monitored system. In addition, in many situations, the changes in feature vectors are closely related to model parameters. and a functional mapping between the drifting parameters and the selected prognostic features can be established. Moreover, if the understanding of the system degradation improves, the models can be adapted to increase its accuracy and address subtle performance problems. Consequently, they can significantly outperform data-driven approaches (see next section). But this closed relation with a mathematical model may also be a weakness: it can be difficult, even impossible to catch the system's behaviour. Further, some authors think the monitoring and prognostic tools must evolve as the system does [14].

Figure 6 shows a schematic representation of physical-based prognostics (model-based approach).

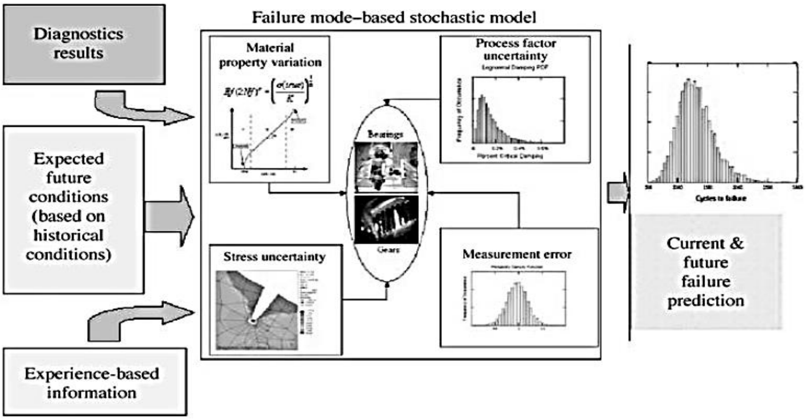


Figure 6. Physics-Based Modelling Approach.[14]

2.4.3.2. Data-Driven Approaches.

In many situations, the complexity of the systems under observation makes it impossible to derive robust and accurate models which can be used for prognostic purposes. However, historical data which capture the signal behaviour of measured signals or extracted features from the incipient fault stage to equipment failure are often available. In such cases, data-driven methods which model how such signals and features evolve can be utilised to generate predictions of the RUL.

Data-driven prognostic approaches typically adopt one of two strategies. The first is a two-stage process. Appropriate dimensionality reduction, feature extraction, or pattern matching techniques are employed to map system signals or features onto a single dimension damage, degradation, or health index. Technically, this first step falls under the realm of fault diagnostics since it is concerned with posterior event analysis. Once the current level of degradation is identified, it is extrapolated into the future until a predefined critical threshold limit is exceeded. A range of techniques can be applied in both steps. The second strategy is to directly model the relationship between monitored signals or features, and the remaining life of the system. In this situation, the remaining life of the system is the output generated by the models.

The next section presents a brief overview of some data-based techniques which have been applied to prognostic problems.

Time Series Approaches: The simplest data-driven approaches to prognostics rely on projection methods which project the current level of degradation into the future. This task is essentially a time series prediction problem and within the realm of prognostics has been addressed by a variety of approaches, including autoregressive models and exponential smoothing techniques.

The autoregressive integrated moving average model (ARIMA) is part of a general class of linear models that have historically seen wide use in modelling and forecasting time series. Unsurprisingly, such approaches have also been applied to prognostic problems which are similar, in many respects, to forecasting problems. ARIMA models are derived from the more common autoregressive moving average (ARMA) model which models a time series using two parts, an autoregressive (AR) part and a moving average (MA) part. However, since ARMA models can only be used to model stationary processes, ARIMA models are often employed, as these can be used to model non-stationary time series signals. Examples of ARIMA models applied to prognostic problems can be found in[15].

Artificial Neural Networks: Perhaps the most common data-driven technique applied to prognostic problems is an artificial neural network (ANN). ANNs model relationships between input and output variables with a model structure inspired by the neural structure of the brain. The network weights and biases, which define the interconnections between the neurons, are adapted during a training process to maximise the fit between the input and output data on which the models are trained. ANNs have been applied in a number of ways in prognostics. Their most common use is in time series prediction, where the current degradation state is predicted into the future until it exceeds a threshold value. Typically, in a feed-forward ANN, previous values of the degradation index are used as the inputs to generate a one step ahead prediction. The generated output is then fed back as an input to the next iteration, to generate long-term predictions. ANNs can also employed to estimate the current degradation index, using system features as inputs. The degradation index can then be predicted into the future using ANNs again, or via alternative prediction methods. A dynamic wavelet neural network (DWNN) is used to predict the fault propagation process into the future and estimate the RUL. The first ANN used, the WNN, is a static feed-forward neural network used to derive a static relationship between inputs and ouputs. The second ANN used, the DWNN, is a recurrent neural network (RNN), which incorporates feedback within the network structure to predict the time series evolution. ANNs have also been used to directly model the relationship between system features and the RUL. ANNs learn by example and, as such, require sufficient instances of historical failure examples for training, and are typically data hungry. As a result, ANNs can generate poor prediction performance when future failure examples on which they have not been trained do not exhibit behaviour similar to the examples in the training set.

While the intrinsic ability of GPR and RVM to generate confidence limits on generated predictions makes such approaches attractive, ANN approaches, in general, do not provide confidence limits associated with predictions. The availability of confidence limits associated with RUL predictions are highly desirable and provide a means for uncertainty management.

Other data-based techniques which have been applied to prognostic problems include hidden Markov models and Neuro-Fuzzy networks [17], but in many situations, the complexity of the systems under observation makes it impossible to derive robust and accurate models for prognostic purposes.

Advantages of Data-Driven Approaches.

- Relatively simple to implement and faster:
  - ✓ Variety of generic data-mining and machine learning techniques are available.

- Help gain understanding of physical behaviours from large amounts of data:
  - ✓ These represent facts about what actually happened which may not be apparent from theory.

Disadvantages of Data-Driven Approaches.

- Physical cause-effect relationships are not used:
  - ✓ e.g. different fault growth regimes, effects of overloads or changing environmental conditions.
- Difficult to balance between generalisation and learning specific trends in data:
  - ✓ Learning what happened to several units on average may not be good enough to predict for a specific unit under test.
- Require large amounts of data.
  - ✓ We never know if we have enough data or even how much is enough.

2.4.3.3. Experienced-Based Prognostics.

If a physical model of a subsystem or component is absent and there is an insufficient sensor network to assess the condition, an experienced-based prognostic model may be the only alternative. This model is the least complex and requires the failure history or "by-design" recommendations of the component under similar operation. Typically, failure and/or inspection data are compiled from legacy systems and a Weibull distribution or other statistical distribution is fitted to the data. An example of these types of distributions is given in Figure 7. Although simplistic, an experienced-based prognostic distribution can be used to drive interval-based maintenance practices that can be updated at regular intervals. An example may be the maintenance scheduling for a low criticality component with few or no sensed parameters. In this case, the prognosis of when the component will fail or degrade to an unacceptable condition must be based solely on analysis of past experience or OEM recommendations. Depending on the maintenance complexity and criticality associated with the component, the prognostics system may be set up for a maintenance interval (i.e. replace every 1000+/-20 Effective Operating Hrs) and updated as more data become available. Having an automated maintenance database is important for the application of experience-based prognostics [14].

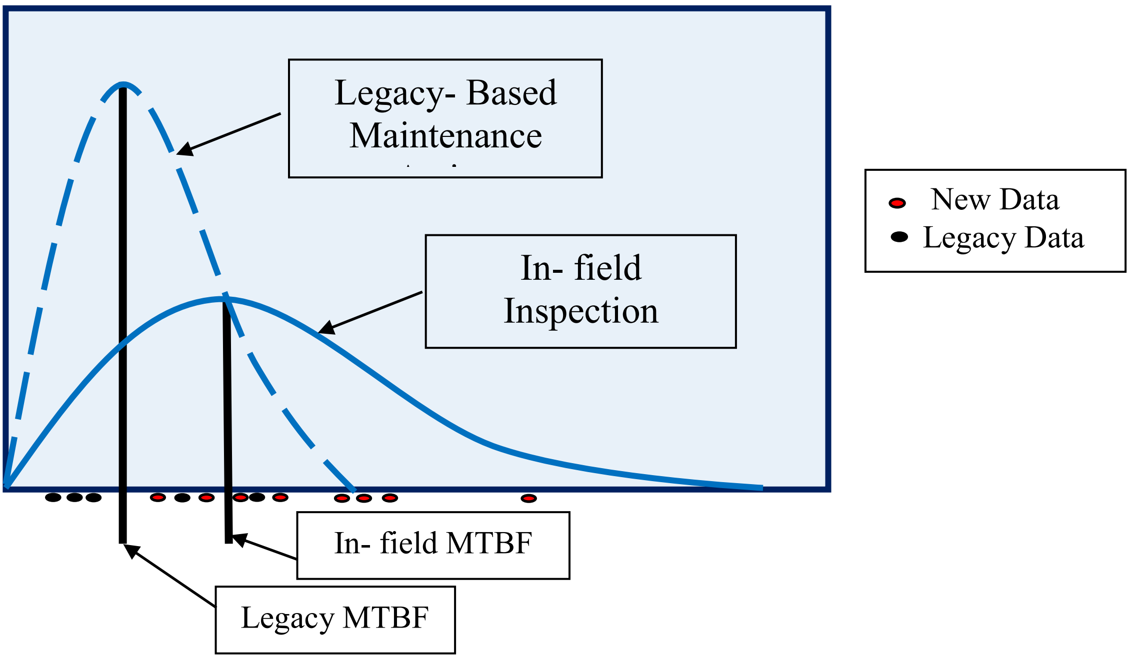


Figure 7. Experienced-Based Approach.

The simplest prognostic approaches rely on the collection of statistical information which examines historical failure rates of systems or components. Such data can be used to develop life-usage models showing distributions of failure rates over time. Such approaches are helpful in the development of preventative maintenance schedules in which maintenance is performed on the basis of mean time between failures (MBTF), as derived from life-usage models. However, such approaches lack predictive capability and, as such, cannot be described as truly predictive prognostic techniques. At the same time, they have wide applicability in systems or components with low criticality or cost or in situations where sensor data which can be used to infer condition are not available [15].

2.4.3.4. Hybrid approaches.

The hybrid approach attempts to harness the power of both data-based approaches and model-based approaches. In fact, we rarely see a case where only one model is used. Hybrid approaches can be classified into two types, pre-estimate and post-estimate fusion.

Pre-estimate fusion is usually applied when there are no data on the ground. This can occur in situations where the diagnosis does a good job detecting faults before a system failure occurs; as a result, there is almost no information on failure. But for maintenance, it is always useful to know when a system will make better use of the remaining useful life while avoiding unscheduled maintenance (usually more costly). Pre-estimate fusion can also be done by a process outside the combined online process line: This requires the use of a simulation model based on physics to understand the relationships of sensor responses to error states in the online mode. First, data can be used to identify the current status of the damage. The data can be then tracked to determine the spread of damage. Finally a propagation model based on individual data for remaining life prediction can be applied.

Post-estimate fusion can be used to reduce the uncertainty ranges of approaches based on data or models while improving accuracy. The underlying notion is that multiple sources of information can improve the performance of an estimator. This principle has been successfully applied in the context of the fusion classifier; the output of multiple classifiers is used to achieve a better result than any single classifier. In the context of forecasts, fusion can be performed by assigning quality assessments to individual estimates based on a variety of inputs, e.g., heuristics, performance known *a priori*, horizon prediction, and robustness of the prediction [18].

- *Advantages of Hybrid Approach*

✓ Does not necessarily require high fidelity models or large volumes of data: works in a complementary fashion.

✓ Retains intuitiveness of a model but explains observed data.

✓ Helps in uncertainty management.

✓ Flexible.

- *Disadvantages.*

✓ Needs both data and models.

✓ An incorrect model or noisy data may bias the approach.

Figure 8 shows a sample of hybrid approaches to prognosis; here, the estimation of the RUL equation requires I Tando-based data such as those data used in model-based approaches.

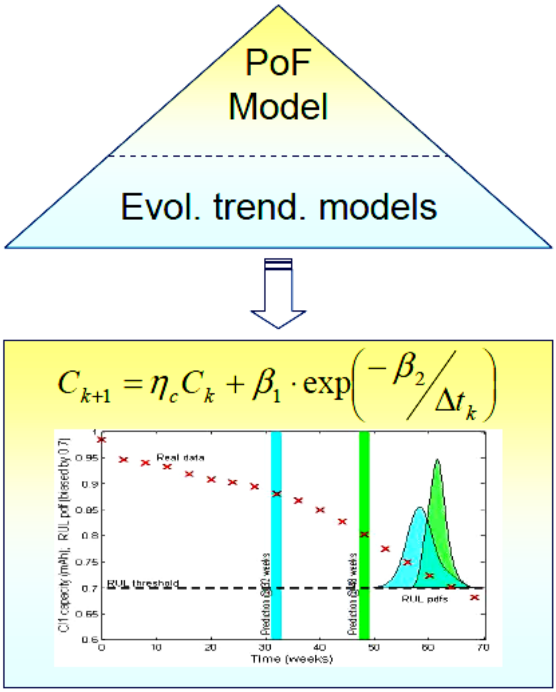


Figure 8. Hybrid approaches. [19]

2.4.3.5. Approaches and Prognostics Techniques.

The following diagram illustrates the approaches and their respective techniques used in prognostics.

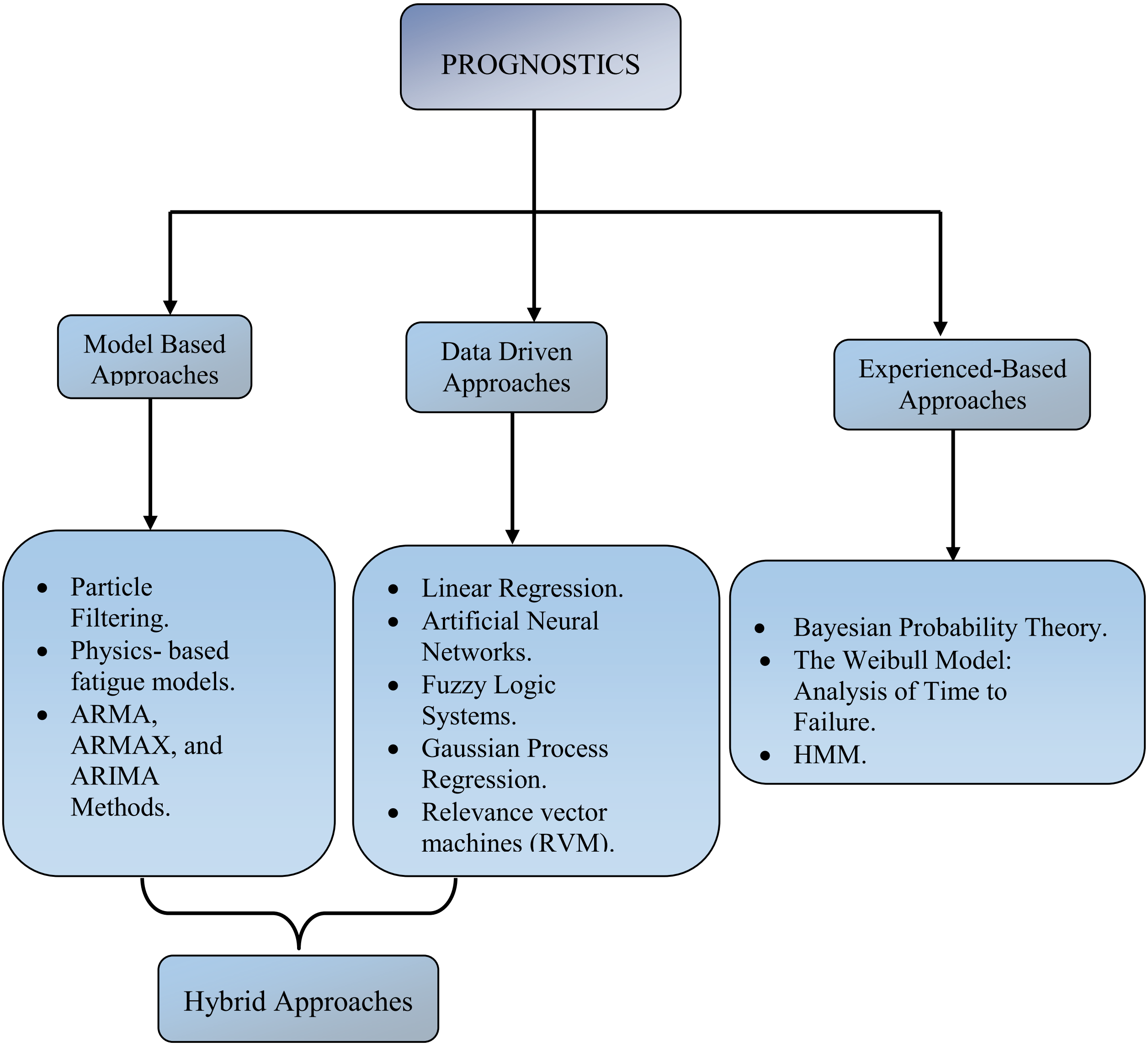


Diagram 1. Approaches to and Techniques of Prognostics.

Condition Monitoring and Prognostics Techniques 3

This chapter introduces the procedure used to implement prognostics in a particular asset to estimate the remaining useful life (RUL), and explains the prognostics techniques used for the estimation.

3.1. Implementing Prognostics.

Figure 9 shows the steps used to estimate the remaining useful life (RUL) of an asset. The steps are explained in more detail in the following subsections.

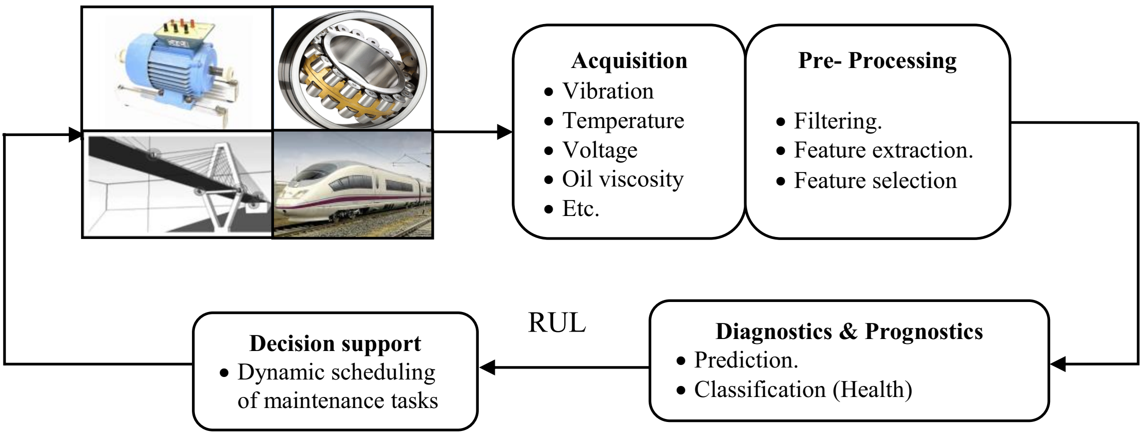
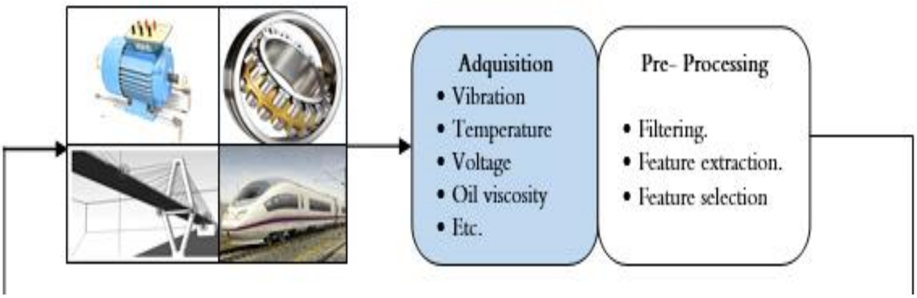


Figure 9. Implementing Prognostics.

3.1.1. Data Acquisition and Data-Processing

3.1.1.1. Data Acquisition:



Once the machine or asset has been selected for analysis, depending on the type of asset, we must select the most appropriate monitoring technique. , Possible techniques include vibration analysis, thermography, tribology, oil analysis etc. This will enable the collection of data on the physical variables (temperature, vibration , viscosity, voltage, acoustic waves etc.) to be analysed to estimate useful life.

Data are collected using sensors placed on specific points on the assets; the types of sensors represent the dependent variable and can be analysed.

Vibration sensors:

- ✓ Displacement transducers or contact (how vibrates, distance);
- ✓ Speed sensors (speed of vibration);
- ✓ Laser vibrometer;
- ✓ Accelerometers (relationship between internal forces);
- ✓ Force transducer.

Temperature sensors:

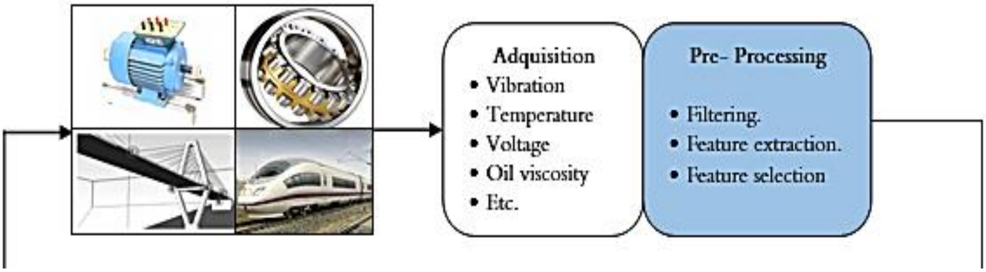
- Mechanical:
  - ✓ Expansion of liquid bulb thermometer;
  - ✓ Expansion of solids: prints bimetallic.

- Electrical:
  - ✓ Thermoelectric effect: thermocouples;
  - ✓ Variation of resistance;
  - ✓ Metals: resistance thermometers;
  - ✓ Semiconductors, thermistors.
  - ✓ *Sound sensors:*

Domain	*Examples*
Thermal	Temperature (ranges, cycles, gradients, ramp rates), heat flux, head dissipation.
Electrical	Voltage, current, resistance, inductance, capacitance, dielectric constant, charge, polarisation, electric field, frequency, power, impedance
Mechanical	Length, area, volume, acceleration, mass flow, force, torque, stress, density, stiffness, angular, direction, vibration.
Chemical	Chemical, species concentration, gradient, reactivity, mess, molecular weight.
Humidity	Relative humidity, absolute humidity.
Optical	Intensity, phase, wavelength, polarisation, reflectance, transmittance, refractive index, distance, vibration, amplitude and frequency.
Magnetic	Magnetic, field, flux density, magnetic, moment, permeability, direction, distance, position, flow

Figure 10. Physical Variables.

3.1.1.2. Data-Processing:



Once the data are collected from the monitoring techniques, we must filter the data, because most of the time, the data are contaminated by the environment where the asset is used. This contamination can be, for example, noise signals. These alter the collected data causing erroneous results.

In addition to filtering, the following must be done.

Feature extraction:

- ✓ Extract patterns from raw signals.
- ✓ Compute numeric or symbolic information from the observations: build features.
- ✓ Use techniques from signal/image processing, data-mining, statistics, etc.

Feature selection:

- ✓ Reduce dimensionality.
- ✓ Use techniques from data-mining, statistics, machine learning, genetic algorithms, etc. [19].

Ffigure 11 shows a example of feature extraction from a vibration signal.

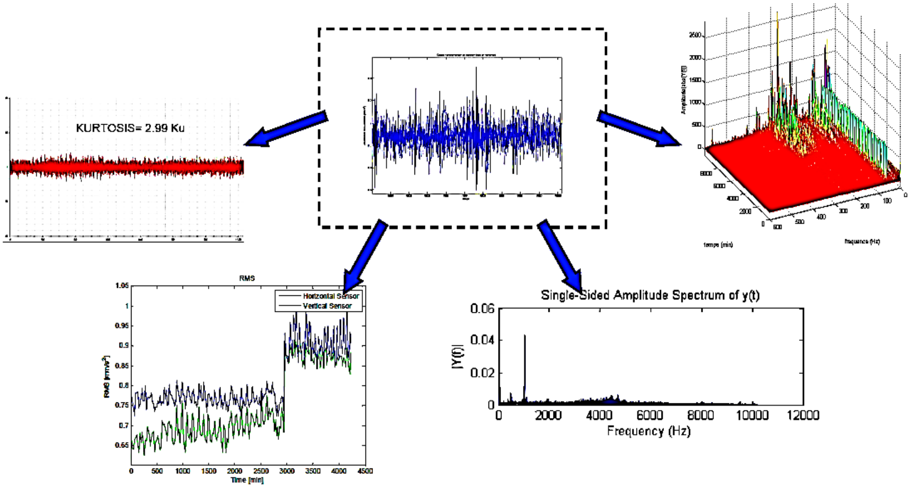
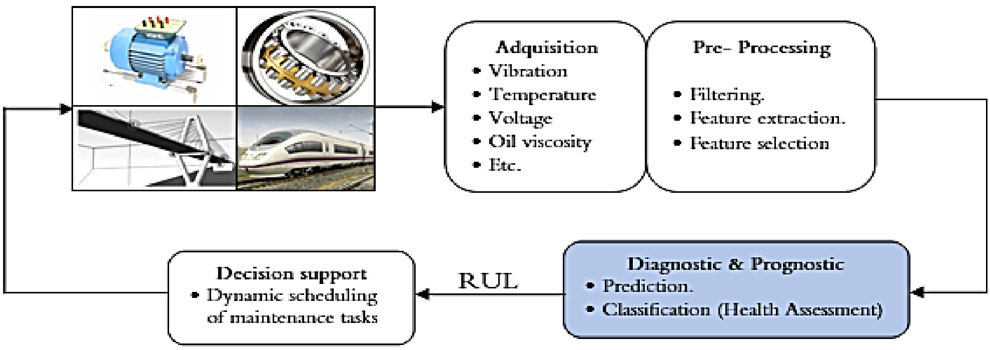


Figure 11. Feature Extraction From a Vibration Signal. [19]

3.1.2. Diagnostics & Prognostics.



3.1.2.1. Diagnostics.

After the data are monitored, analysed and processed, we proceed to make a diagnosis of the asset under analysis. The goal of diagnosis is to determine the state of the asset's health to identify whether an asset is failing or not operating correctly. The algorithms' fault diagnosis are designed to detect system performance, control levels of degradation, and identify failures (faults) based on changes in physical properties, through detectable phenomena. Such systems also identify the subsystem or component which is failing, and the specific mechanism of failure that has occurred [15].

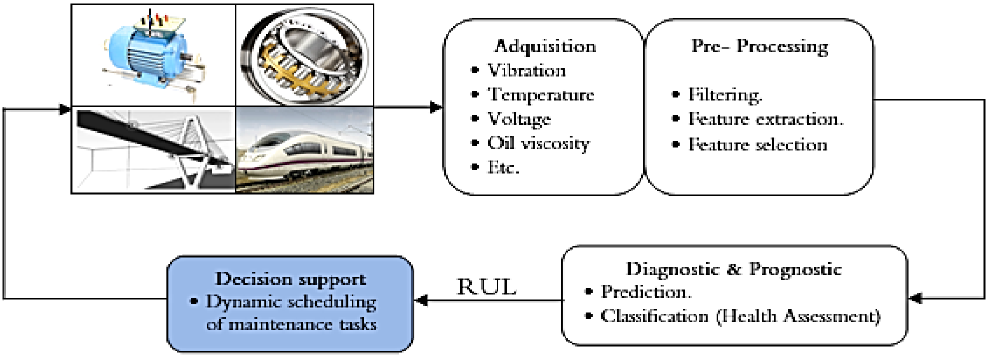
After making the diagnosis and determining the health of the asset mostmaintainers want to know when stop using it; this requires an estimation of the RUL which is always done after the diagnosis.

Note: This work does not explain data processing techniques in depth as this is beyond the present scope of analysis.

3.1.2.2. Prognostics to Estimate the RUL:

As noted above, prognostics can be used to estimate the remaining useful life (RUL). The approach can be model-based, data-driven, or experience-based approaches. A hybrid approach fuses elements of the others. Each can be implemented using the techniques shown in Diagram 1. These will be explained later.

3.1.3. Decision Support.



After estimating the remaining useful life of the asset or the analysed systems, a series of decisions must be made, depending on the type of asset or system. For example, we must decide whether to perform maintenance activities and whether to intervene in the system to perform those activities. Clearly, such decisions are very important to save costs. Hence, a good estimate of RUL is crucial. Should a maintenance activity be performed now, or can we wait for the date given by the RUL?

industry.
The procedure to implement prognostics is shown in Figure 12.

1. ***Sensor Module*** : provides system access to digitized sensor or transducer data
2. ***Signal Processing***: performs signal transformations and CBM feature extraction
3. ***Condition Monitor***: compares features against expected values and generates alerts
4. ***Diagnostic Processing*** : generates a diagnostic record (fault conditions and confidences)
5. ***Prognostic Processing***: projects the current health state into the future and estimates

Figure 12. Implemeting prognostics.

3.2. Prognostics Techniques.

As explained above, prognistics is based on three types of approaches, alonf with an approach fusing two of the three. The techniques used by these approaches to estimate the remaining useful life of an asset or system are explained below.

3.2.1. Techniques of Model-Based Approaches.

3.2.1.1.Particle Filtering for Prognostics.

One way to estimate the remaining useful life (RUL) of a component failure or degrading system is to use a Bayesian technique. This technique uses a dynamic state model and a measurement model to predict the density function posterior probability of the state, i.e., to predict the time evolution of a fault.

A particulate filter is a recursive Bayesian estimation technique and is used to avoid the assumption of linearity and Gaussian noise. Kalman filtering provides a solid framework for long term prognosis. A methodology for accurate and precise prediction, it presents a failing component on the basis of particle filtering strategies. We use fatigue failure as an example here.

Prediction of the evolution of a fault or fault indicator entails large-grain uncertainty. Accurate and precise prognosis of the time to failure of a failing component/subsystem must consider critical-state variables such as crack length, corrosion pitting, etc. as random variables with associated probability distribution vectors. Once the probability distribution of the failure is estimated, other important prognosis attributes such as confidence intervals can be computed.

A possible solution to the prognosis problem is the use of recursive Bayesian estimation techniques that combine both the information from fault-growth models and online data obtained from sensors monitoring key fault parameters (observations). Prognosis or long-term prediction for failure evolution is based on both an accurate estimation of the current state and a model describing the fault progression. If the incipient failure is detected and isolated at the early stages of the fault initiation, it is reasonable to assume that sensor data will be available for a certain time window, allowing corrective measures to be taken. Thus, there will be improvements in model parameter estimates so that prognosis will provide accurate and precise prediction of the time to failure. At the end of the observation window, the prediction outcome is passed on to the user (operator, maintainer); additional adjustments are not feasible because corrective action must be taken to avoid a catastrophic event.

Figure 13 depicts a conceptual schematic of a particle-filtering framework aimed at addressing the fault prognosis problem. CBM sensors and the feature-extraction module provide the sequential observation (or measurement) data of the fault growth process $\mathbf{Z_k}$ at time instant $\mathbf{k}$. We assume the fault progression can be explained through the state-evolution model and the measurement model:

$$X_k = f_k(X_{k-1}, \omega_k) \leftrightarrow p(X_k \mid X_{k-1}) \quad (3.1)$$

$$Z_k = h_k(X_k, v_k) \leftrightarrow p(Z_k \mid X_k) \quad (3.2)$$

where:

- ✓ $\mathbf{X_k}$ is the state of the fault dimension (such as crack size)
- ✓ $\boldsymbol{f_k}$: $\mathbb{R}^{n_x}\ x\ \mathbb{R}^{n_\omega} \rightarrow \mathbb{R}^{n_x}$ is the state transition function (possibly nonlinear) which describes the evolution of the system state, where nx and n! are the dimensions of the state and process noise vectors respectively.
- ✓ $\boldsymbol{h_k}$: $\mathbb{R}^{n_z}\ x\ \mathbb{R}^{n_v} \rightarrow \mathbb{R}^{n_z}$ is the measurement/observation function which describes the sequence of measurements $\mathbf{Z_k}$ collected at successive time steps $\mathbf{t_k}$.
- ✓ $\boldsymbol{\omega_k} \in \mathbb{R}^{n_\omega}$ is an independent identically distributed (i.i.d.) process noise sequence of known distribution.
- ✓ $\boldsymbol{v_k} \in \mathbb{R}^{n_v}$ is an i.i.d. measurement noise sequence of known distribution.

The first part is state estimation, that is, estimating the current fault dimension as well as other important changing parameters in the environment. The second part is long-term prediction based on the current estimate of the fault dimension and the fault growth model with parameters refined in the posteriori state estimation. A novel recursive integration process based on both importance sampling and PDF approximation through Kernel functions is then applied to generate state predictions from (k+1) to (k + p) as the following:

$$P(X_{k+p} \mid Z_{1:k}) = \int P(X_{k+p} \mid Z_{0:k}) \prod_{j=k+1}^{k+p} P(x_j \mid x_{j-1})\, dx_{k:k+p-1} \quad (3.3)$$

$$= \sum_{i=1\$}^{N} w_k \int \ldots \int P(x_{k+1} \mid x_k^{(i)} \prod_{j=k+2}^{k+p} P(x_j \mid x_{j-1})\, dx_{k+1:k+p-1} \quad (3.4)$$

Long-term predictions can be used to estimate the probability of failure in a process, given a hazard zone defined by its lower and upper bounds (Hlb and Hup, respectively). The prognosis confidence interval, as well as the expected time to failure (TTF), can be deduced from the TTF PDF as:

$$P_{TTF}(TTF) = \sum_{i=1}^{N} \Pr\left\{ H_{lb} \leq x\overset{(i)}{TTF} \leq H_{up} \right\} w_{TTF} \quad (3.5)$$

The procedure described in Sections 3.3 and 3.4 form the basis for the determination of the optimal Bayesian solution. However, the recursive computation of the posterior state PDF is more conceptual than practical and, in general, cannot be determined analytically. In a restricted set of cases, such as when using linear Gaussian state space models, the optimal solution leads to the well-known Kalman filter (mentioned above). However, in the presence of a nonlinear process model and/or non-Gaussian noise processes, an alternative approach must be considered. A common approach is to use particle filtering methods, which approximate the optimal Bayesian solution.

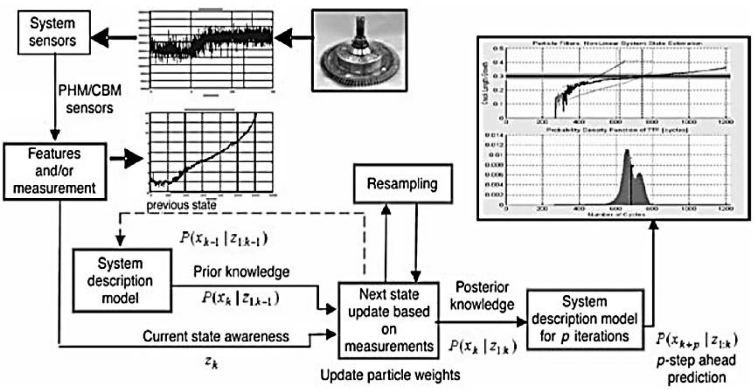


Figure 13. Fault Prognosis Based on Particle Filtering.

3.2.1.2. Physics-Based Fatigue Models.

Physical models quantitatively characterise the behaviour of a failure mode using physical laws. This means we must have deep knowledge about the behaviour of the system being studied. Physical models perform an estimate of the remaining useful life of an asset or system by solving equations or a deterministic set of equations derived from empirical data collected in the data acquisition. Some of these data are obtained using common scientific and engineering knowledge, while others are collected through specific laboratory or field experimentation.

Physical models for the estimation of the remaining useful life of a particular system must identify one or more specific parameters of the system (e.g., exact physical properties, corrosion rates, equation constants). These models are generally described in two ways: first, using dynamic Lagrangian or Hamiltonian dynamic ordinary differential or partial equations (approximation methods applied to partial differential equations, distributed models and other techniques); second, using methods of state space (i.e., no differential equations) and resolved accordingly[22]. Once a physical model is available, the sensor measurements are compared to actual process model outputs. The difference between reality and the model is called waste; large residuals are assumed to indicate a fault while small residues occur in normal conditions, noise and modelling errors. The residuals are calculated using parameter estimation, state-space methods or parity equations .[23]

Remaining useful life estimates are based on projecting degradation behaviours into expected future operating conditions. To construct a physics of failure model, the following features and associated levels of certainty must be characterised:

- ✓ Set of likely initiating failure modes for which behavioural models are required.
- ✓ Process behaviour across possible/typical operating ranges.
- ✓ Degradation behaviour under aforementioned process conditions.
- ✓ Relationship between process measurements and degradation behaviour(s).
- ✓ Process and measurement noise.[24]

Generally, failure mechanisms by which physical models have been developed include fatigue, overload failure, corrosion and ductile to brittle transitions. When discussing effects, failure mechanisms can be divided into two categories. The first can be considered failures upon effort. These occur when the load exceeds the material strength; these is no long term effect once the load has been removed. Examples include brittle fracture, ductile fracture, and buckling performance. The second category is failures; these are characterised by irreversible, cumulative damage that does not disappear when the load is removed, such as fatigue, corrosion, stress corrosion cracking, wear and creep. Once the mechanism allowed for the particular damage tolerance is exhausted, the normal operating loads will exceed the strength the asset has left and a failure will occur upon effort [25].

In a physical forecasting model, if we want to implement the failure modes, we must be able to track the damage and its rate of progression under any or all conditions of operation. However, when talking about failures upon effort, we only need to monitor the current state to determine the immediate risk of failure. Similarly, a stochastic analysis of force and distributed load must be carried out [26]. In practice, when discussing behaviour, forecasting applies only when we have identified the relevant failures. The main advantage of this group of models is the ability to incorporate the current understanding of the physical mechanisms of failure; in many cases, this has been the subject of extensive and exhaustive empirical testing. Crack propagation models, of clear interest to prognosis, include deterministic and stochastic models. Growth of fatigue cracks in the components of typical equipment, such as bearings, gears, shafts and aircraft wings, is affected by a variety of factors, including the states of stress, material properties, temperature, lubrication and other environmental effects. A variation of empirical and deterministic models can model fatigue crack propagation based on Paris 's formula [22]. Fatigue crack growth can be expressed as:

$$\frac{d\alpha}{dn} = C_0(\Delta K)^n \tag{3.6}$$

where:

α : Instantaneous length of dominant crack.

N : Running cycles.

C_O : Material dependent constants.

ΔK: Range of stress-intensity factor over one loading cycle.

$\Delta K = Y(\alpha)(\Delta_S)\sqrt{\pi\alpha}$ where: is a stress intensity factor, Y(α) is related to crack geometry, Δs is the stress range.

Growth models of stochastic cracks consider all parameters as random quantities. Thus, the resulting equation is a growth stochastic differential equation. Here, to estimate the distribution parameters, we use Monte Carlo simulations, probabilistic neural networks, and others [22].

3.2.1.3. ARMA, ARMAX, and ARIMA Methods.

Parameter estimation prognosis is another model possibility when the physics-based process model is unavailable or too complicated for implementation. In such cases, some of the system identification procedures must be used. Based on the knowledge about the system, a model structure and complexity can be proposed. System identification reduces the estimation of the unknown model parameters vector using the observed input and output sequences.

As explained at greater length below, autoregressive moving average models or ARMA (p, q) capture the entire process when a simple model AR (p) or MA (q) cannot. Autoregressive integrated moving average models or ARIMA (p, d, q) solve the problem where the test range is not stationary due to the existence of a trend. Finally, multiplicative seasonal ARIMA models (p, d, q) x (p, d, Q)s solve the problem where the test range is not stationary due to the existence of a seasonality and / or a trend.

- *Autoregressive Moving Average Models or ARMA (p, q).*

An autoregressive moving average model combines the two basic models AR (p) and MA (q) to model partially and partially autoregressive series of moving averages. These combined models are abbreviated by the acronym

ARMA (p, q), where p is the order of the autoregressive part and q is the order of moving average.

The general expression for an autoregressive moving average model ARMA (p,q) is:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q} + \mu \quad (3.7)$$

As the model is a combination of the two basic models, the value of the series ***Yt*** is a linear combination of the most recent disturbance ***q*** and ***p***'s most recent observations.

- *Autoregressive Integrated Moving Average Models or ARIMA (p, d, q).*

A time series ***(Yt)*** follows an autoregressive integrated moving average model if the tenth differentiation (where d is an integer) $Wt = \nabla^d Y_t$ is a stationary ARMA process. If ***(Wt)*** follows an ARMA (p, q), we can say that ***(Yt)*** is an ARIMA process (p, d, q). For practical purposes, we usually use a ***d*** of 1 or 2.

The differential equation of the form of an integrated moving average ARMA (p, d, q) autoregressive model is:

$$Y_t = (1+\phi_1)Y_{t-1} + (\phi_2 - \phi_1)Y_{t-2} + \ldots + (\phi_p - \phi_{p-1})Y_{t-p} + a_t + \theta_1 a_{t-1} + \ldots +$$
$$\theta_2 a_{t-2} + \ldots + \theta_q a_{t-q} + \mu$$

(3.8)

- *Multiplicative Seasonal ARIMA Models (p, d, q) x (p, d, q)s.*

A time series ***(Yt)*** is a multiplicative seasonal ARIMA model with no seasonal orders (regular) ***p, d*** and ***q***, seasonal orders ***P***, ***D*** and ***Q***, and a seasonal period s if the series differentiated $Wt = \nabla^d \nabla_s{}^D Y_t$ satisfies an ARMA model (p, q) x (P, Q)s with seasonal period ***s***. Thus, (Yt) process is an ARIMA (p, d, q) x (P, D, Q)s with seasonal period ***s***.

In its most general form the ARIMA model (p, d, q) x (P, D, Q) s can be written as:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_{ps+p+Ds+d} Y_{t-ps-p-Ds-d} + \ldots + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_{qs+q} a_{t-qs-q} + \mu \tag{3.9}$$

This expression defines which is the most generic of the family forming the ARIMA model; all the others are special cases of it.

When we want to do prognostics using the ARIMA model we must take into account the need to understand the equation that defines the delay model to a set in ***t + l***, where ***l*** is the moment we want to predict. As can be seen, no delay is actually calculated, and if the time ***t*** is considered "current", the predicted time is at a distance ***l*** in the future.

The formula to predict the value of the variable for a moment ***l*** for a general ARIMA (p, d, q) x (P, D, Q)s is given as:

$$\hat{Y}(l) = \phi_1 \hat{Y}_t(l-1) + \phi_2 Y_t + \phi_{l-1} Y_t(l) + \ldots + \phi_{l+1} Y_{t-1} + \ldots + \phi_{Ps+p+Ds+d} Y_{t+l-ps-p-Ds-d} + \theta_l a_t + \theta_{l+1} a_{t-1} + \theta_{Qs+q} a_{t+l-Qs-q} + \mu \tag{3.10}$$

Importantly, the following aspects of prognostics follow this formulation:

- ✓ Prognostics values are calculated sequentially. To enable the implementation of the autoregressive part of a model in periods other than the first prediction, the predicted value is taken in the immediately preceding $\hat{Y}_t(l-1)$.
- ✓ All known random perturbations in the sample period are considered to have the character of white noise in prognostics periods.
- ✓ By the necessary condition of being at the optimal predictor (the expected value of the series in the period of prognostics is equal to that predicted, if optimal), random perturbations used in the prediction are those of previous periods, i.e. :

$$a_{t+j} \quad \forall j \leq 0 \tag{3.11}$$

Thus, in the prognostics period, random perturbations will have no effect. To make subsequent predictions ***Yt (l)*** we must have values for ***at, at-1, ... t + lq***, calculated as:

$$a_{t+j} = Y_{t+j} - \hat{Y}_{t+j-l}(l) \quad \forall j \leq 0 \tag{3.12}$$

- ✓ ARIMA prognostics are adaptive and the results obtained for ***(t +l),*** with the information available at time ***t***, are the same as those obtained for the same period taking as baseline information to ***t-1***, and adding a term error [22].

3.2.2. Data Driven Approaches.

3.2.2.1. Linear Regression.

This subsection provides an overview of the concepts and techniques associated with regression analysis. Regression analysis uses the existing data and determines the relationships, if any, between the measurable outcome and the variables contributing to that outcome (e.g. life expectancy is the outcome and exercise and diet are the variables contributing to that outcome.) Neter et al. [27] present a framework for the statistical relation in order to predict machine remnant life. Their general linear regression model is given by

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_j, \; i = 1,2,\ldots n \tag{3.13}$$

where $\boldsymbol{Y_i}$ is a random variable denoting the value of the i[th] trial's response, bo,b1,b2, are estimated parameters $\boldsymbol{X_{I,I}}$, $\boldsymbol{X_{i,2}, \ldots X_{i,p-1}}$ are the values of the prediction, or contributing variables, and $\boldsymbol{\varepsilon_i}$ is the random error with mean = 0, variance = $\boldsymbol{\sigma^2}$, and covariance = 0. Regression analysis seeks to estimate the parameters of the regression function, $\beta_0, \beta_1, \beta_2, \beta_{P-1}$, in order to find a representative model by using the method of least squares. This method defines a variable $\boldsymbol{Q}$ where

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i,1} - \beta_2 X_{i,2} - \ldots \beta_{p-1} X_{i,p-1})^2 \tag{3.14}$$

and attempts to find estimates for $\beta_0, \beta_{1,}\beta_2, \beta_{P-1}$ denoted by ***b_o, b_1, b_2…., b_{p-1}***, which minimise ***Q*** for the observations (X_1, Y_1), (X_2, Y_2), …(X_n, Y_n). The simultaneous solution to the equations formed by taking the derivative of ***Q*** with respect to $\beta_0, \beta_{1,}\beta_2, \beta_{P-1}$, provides the least squares estimates, ***b_o, b_1, b_2…., b_{p-1},***. Least squares estimates are desired because they are unbiased and have minimum variance, thus resulting in

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_{p-1} X_{p-1} \qquad (3.15)$$

The method of maximum likelihood can also be used to estimate ***b_o, b_1, b_2…., b_{p-1},*** if the probability distribution of the error terms is known [27].

3.2.2.2. Neural Network Model.

In computer programs, neural networks (NNs) are designed to work in a way similar to how the human brain processes information. By using the concept of learning through experience, NNs pool knowledge to identify patterns and relationships in data. The NN computational model contains hundreds of artificial neurons and connects with known coefficients as weights. Figure 14 shows the model structure of NN. Input signals, ***X_1, X_2,…X_n*** are propagated through the network with weights. ***W_1,W_2,…W_n*** connect input and hidden neurons. The combination of input signals and weights is passed through an activation function to produce the output value of the neuron ***y_k*** [28].

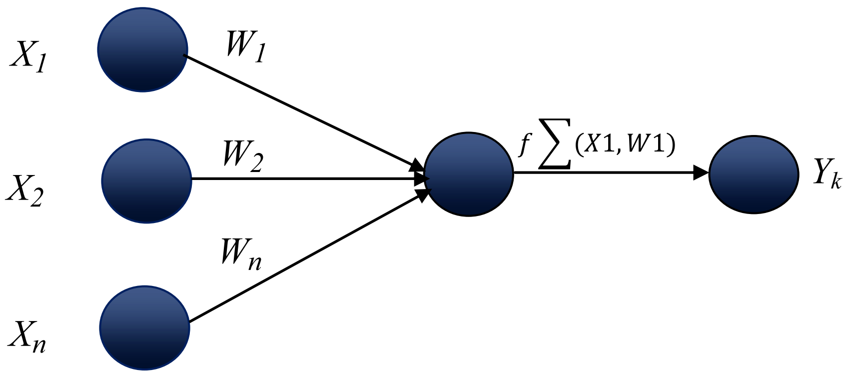


Figure 14. Neural network structure.

NNs have been widely applied in time series predictions. The time series prediction model uses a feedforward neural network (FNN) and employs a sliding window over the input sequence. To construct the multi-step prediction, the model uses previous predicted value to forecast the future values iteratively until the expected future values are obtained. This multi-step prediction is illustrated in Figure 15 [29].

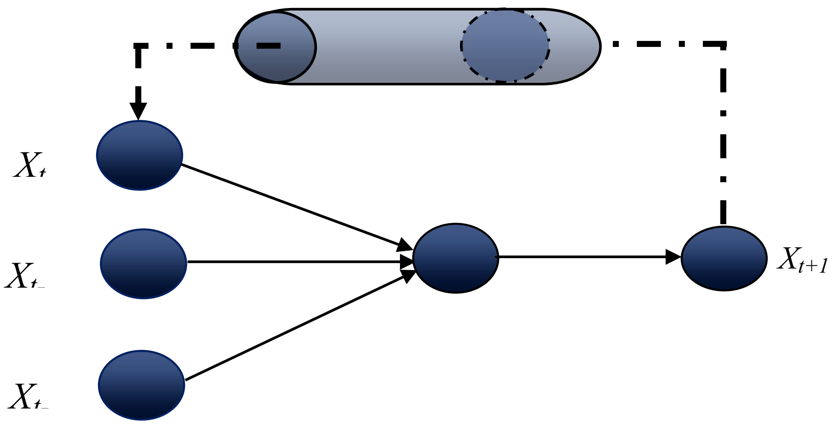


Figure 15. The concept of multi-step prediction.

This multi-step prediction technique uses the previous values to forecast iteratively the future d time step values. Given the observation Yt = [Xt-r+1, Xt-r-2,…Xt], the first future value can be predicted by using

$$\hat{y}_{t+1} = f(y_t) = f(x_{t-r+1}, x_{t-r+2} \dots x_t) \tag{3.16}$$

where r denotes the number of inputs or the size of sliding window dimension. To predict the next value, the model can be given as:

$$\hat{y}_{t+2} = f(x_{t-r+2}, x_{t-r+3} \dots \hat{y}_{t+1}) \tag{3.17}$$

Then, the procedure repeats recursively depending on the required number of time series *d* [30] as shown in:

$$\hat{y}_{t+d} = f(x_{t-r+d}, x_{t-r+d+1} \dots \hat{y}_{t+d-1}) \tag{3.18}$$

The process of prognostics is accomplished by predicting and extrapolating the dynamic FPs over time from the performance degradation model using FNN. In order to use FNN, three other key parameters need to be considered: the number of hidden layers, the choice of activation functions and the number of neurons. As a single hidden layer can compute a uniform approximation of any continuous function [31], the proposed FNN architecture is composed of an input layer, a hidden layer and an output layer with one output neuron. To introduce nonlinearity into the network model, nonlinear activation functions are needed. The logistic and tanh functions can be conbined and used as an activation function from input layer to the output layer.

When applying neural networks, deciding the number of input hidden neurons has always been an issue. Having a smaller number of hidden neurons tends to lead to ineffective performance, while having too many neurons may increase the risk of over-fitting the data and impede generalisation.

Ultimately, the selection of the architecture of a neural network comes down to trial and error [32]. There is a selection method for the process of trial and error to determine the number of input neurons and hidden neurons to create the final combination of activation functions. Here, the training data are iteratively trained with the increase in the number of input and output activation functions. These are changed until the error produced by the network is minimal based on the root mean square error (RMSE), calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x}_t)^2} \qquad (3.19)$$

where $\boldsymbol{n}$ is length of time series data, ***xt*** represents the target values and $\bar{\boldsymbol{X}}_{\boldsymbol{t}}$ represents actual values.

After the FNN architecture is identified, the training dataset is used to train the network in adjusting the synaptic weights. Once the network is completely trained, the weights are frozen, and the network is ready to predict and extrapolate failure probability. To validate the predicting performance of the network, a set of validation data compares the network output with the predicted output.[30]

Is important to note that the NN technique is frequently used in the medical realm, especially to estimate the risk of relapse in cancer patients. Many institutions and medical laboratories are currently using this type of prognostics, as doctors can estimate the time when the patient may have a relapse. This extends the use of prognostics techniques beyond engineering to other fields.

3.2.2.3. Fuzzy Logic Systems.

Neuro-fuzzy systems (NFS) are data-driven methods that have been employed successfully in the prediction of machine condition degradation. The prediction performance of the NFS has been shown to outperform conventional neural-network-based predictors such as the feedforward-neural-network, radialbasis-funtion, and recurrent-neural network-based models. Through off-line training using available data sets, the NFS is used to model machine dynamics to make accurate predictions of machine health conditions. However, since the machine dynamics in real applications change with time, the trained NFS cannot carry out accurate predictions if the new dynamics/states are not taken into account during the prediction process. Since Bayesian algorithms can update system states in real-time via new data, the NFS is integrated with Bayesian algorithms so that on-line data can be used to improve the prediction accuracy [33].

The NFS predictor is, in essence, a fuzzy logic system, where the system parameters are optimised via neural network training. The architecture of the NFS is schematically shown in Figure 16. The NFS consists of five layers: an input layer, membership function (MF) layer, rule layer, normalised layer and output layer. There are $\boldsymbol{l}$ input nodes in the input layer, and each input node is related to $\boldsymbol{m}$ term nodes in the MF layer. Thus, the number of nodes in the MF layer is $\boldsymbol{l} \times \boldsymbol{m}$, where m denotes the rule number [34]. The signal propagation in the NFS proceeds as follows:

Layer 1 (Input layer): The input values are transmitted directly to the next layer without any computation. The outputs of this layer can be expressed by

$$O_i^{(1)} = X_i^{(1)} \qquad (3.20)$$

Layer 2 (MF layer): Each node in this layer performs a membership function calculation. Sigmoid membership functions are used here, as shown below:

$$u_{ij}^{(2)} = \frac{1}{1+\exp(-b_{ij}^{(2)}(O_i^{(1)} - m_{ij}^{(2)}))} \qquad (3.21)$$

where $u_{ij}^{(2)}$ is the MF layer's output associated the $\boldsymbol{j^{th}}$ term of the $\boldsymbol{i^{th}}$ input $O_i^{(1)}, b_{ij}^{(2)}$ and $m_{ij}^{(2)}$ are the parameters of the sigmoid function.

Layer 3 (Rule layer): The following max-product operation is carried out in this layer:

$$O_i^{(3)} = \prod_i u_{ij}^{(2)} \qquad (3.22)$$

where the output $O_i^{(3)}$ represents the firing strength of the j^{th} fuzzy rule.

Layer 4 (Normalised layer): This layer performs the normalisation operation for all the rule firing strengths. The resulting output is given by:

$$O_i^{-(4)} = \frac{o_i^{(3)}}{\sum_i o_j^{(3)}} \qquad (3.23)$$

Layer 5 (Output layer): The output of the NFS is calculated by using the centroid defuzzification procedure, that is,

$$X_{t+r} = O_k^{(5)} = \sum_j O_j^{-(4)} W_{JK}^{(5)} \qquad (3.24)$$

where $W_{jk}^{(5)}$ is the $\boldsymbol{k^{th}}$ estimated output associated with the $\boldsymbol{j^{th}}$ rule [35].

Figure 16 gives an example of the architecture of a fuzzy system login (NFS).

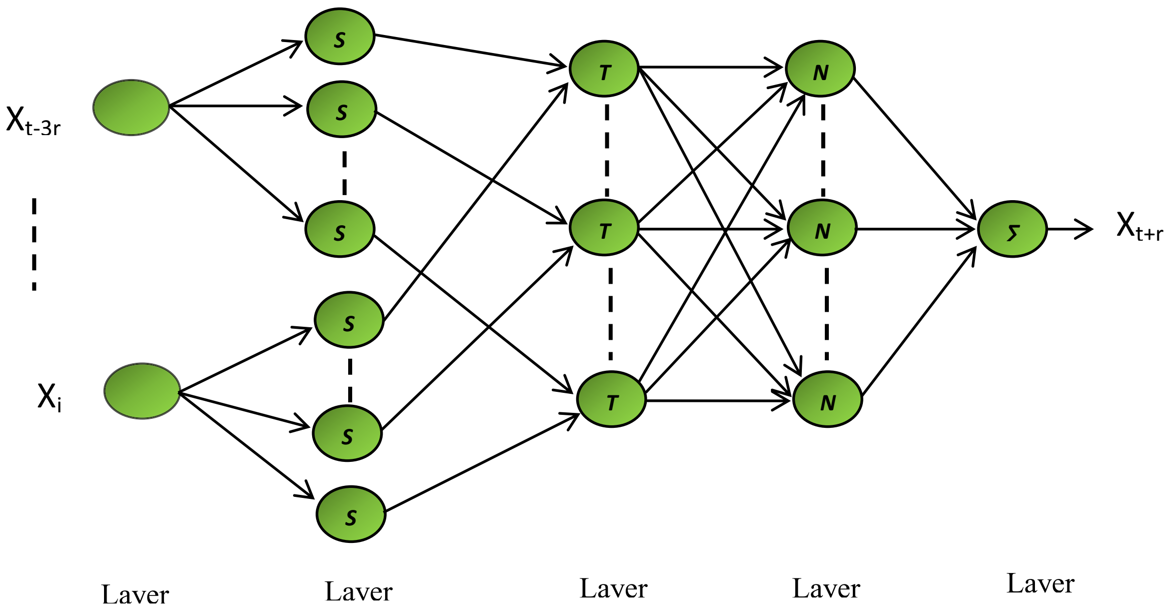


Figure.16. Architecture of the NFS predictor; S is a sigmoid function; T means max-product operation; N means normalisation operation.

Another way to apply data-driven prognostics is to use recurrent neuro-fuzzy systems (RNFS). The RNFS predictor (Figure 17) possesses additional feedback links added in Layer 2 (MF layer). Each node in Layer 2 functions as a memory unit that performs the following operation:

$$u_{ij}^{(2)} = \frac{1}{1+\exp(-b_{ij}^{(2)}(x_{ij}^{(2)} - m_{ij}^{(2)}))} \tag{3.25}$$

where $u_{ij}^{(2)}$ is the MF layer's output associated with the $\boldsymbol{j^{th}}$ term of the $\boldsymbol{i^{th}}$ input $x_{ij}^{(2)}, b_{ij}^{(2)}$ and $m_{ij}^{(2)}$ are the parameters of the sigmoid function. Note that the inputs of this layer involve the feedback components

$$X_{ij}^{(2)}(t) = O_i^{(1)}(t) + \theta_{ij}^{(2)} u_{ij}^{(2)}(t-1) \tag{3.26}$$

where $\theta_{ij}^{(2)}$ is the feedback link weight. It is clear that the activation responses $u_{ij}^{(2)}(t-1)$ at the previous time step are utilised used as one part of the current input values. This allows the RNFS predictor to memorise past information so it can deal with temporal issues. More information on recurrent neuro-fuzzy systems can be found in [36].

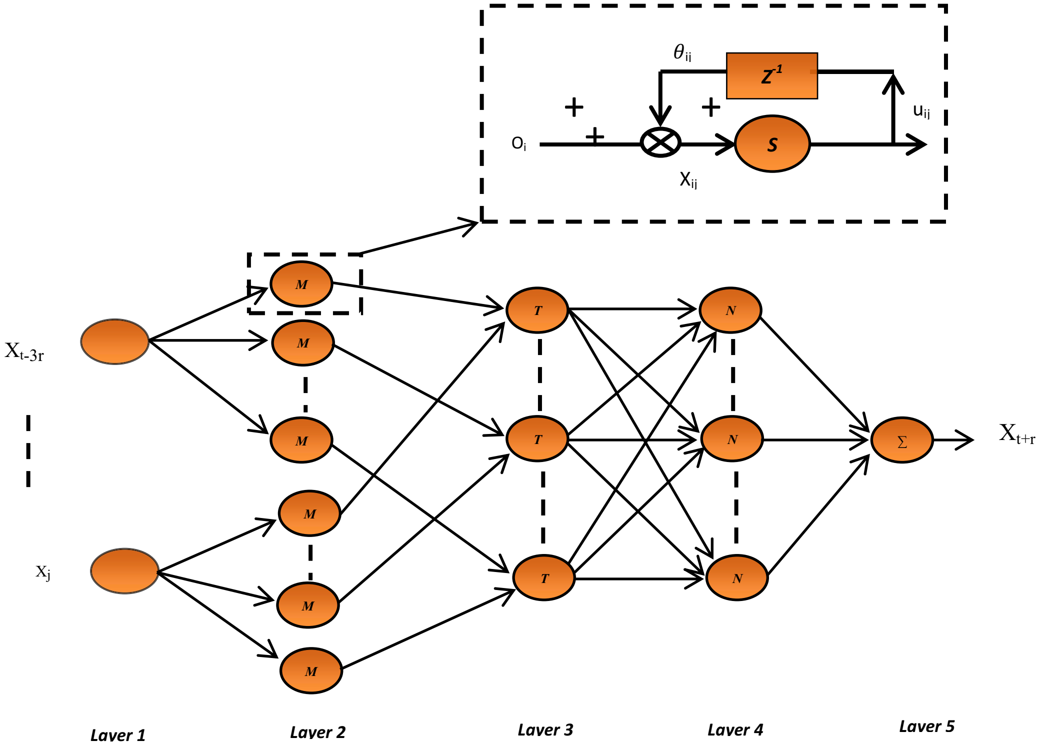


Figure. 17. Architecture of the RNFS predictor; Z-1 is a unit delay operator; S is a sigmoid function; T is max-product operation; N is normalisation operation.

3.2.2.4. Gaussian Process Regression.

A Gaussian Process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution. A real GP ***f(x)*** is specified by its mean function ***m(x)*** and co-variance function ***k(x,x')*** defined as:

$$m(x) = E[f(x)],$$
$$k(x,x') = E[(f(x) - m(x))(f(x') - m(x'))],$$
$$f(x) \approx GP(m(x), k(x,x')).$$
(3.27)

The index set $\boldsymbol{X \in \Re}$ is the set of possible inputs, which need not necessarily be a time vector. Given prior information about the GP and a set of training points $\{(x_i, f_i)\ i=l,\ldots,n\}$, the posterior distribution over functions is derived by imposing a restriction on prior joint distribution to contain only those functions that agree with the observed data points [37]. These functions can be assumed to be noisy, as in real world situations we have access to only noisy observations rather than exact function values; i.e. $y_i = f(x) + \varepsilon$, where ε is additive IID $N(0,\ \sigma^2_n)$. Once we have a posterior distribution, it can be used to assess predictive values for the test data points. The following equations describe the predictive distribution for GPR [38]:

PRIOR

$$\begin{bmatrix} y \\ f_{test} \end{bmatrix} \approx N(0, \begin{bmatrix} K(X,X) + \sigma_n^2 & K(X, X_{test}) \\ K(X_{test}, X) & K(X_{test}, X_{test}) \end{bmatrix})$$

$$f_{test} \mid X, y, X_{test} \approx N(\overline{f}_{test}, \mathrm{cov}(f_{test}))$$

(3.28)

POSTERIOR

$$\overline{f}_{test} \equiv E[f_{test} \mid X, y, X_{test}] = K(X, X_{test})[K(X,X) + \sigma_n^2 I]^{-1}$$
$$\mathrm{cov}(f_{test}) = K(X_{test}, X_{test}) - K(X_{test}, X) + \sigma_n^2 I]^{-1} K(X, X_{test}).$$
((3.29)

where X (Inputs), y (targets), K (convariance function), σ_n^2 (Noise level), and X_{test} (test inputs).

A crucial ingredient in a Gaussian process predictor is the covariance function that encodes assumptions about the functions to be learnt by defining the relationship between data points. The covariance structure also incorporates prior beliefs of the underlying system noise. A GPR requires prior knowledge of the form of covariance function, which must be derived from the context if possible. Furthermore, covariance functions consist of various hyper-parameters that define their properties. Setting right values of such hyper-parameters is another challenge in learning the desired functions. Although the choice of covariance function must be specified by the user, corresponding hyper-parameters can be learned from the training data using a gradient based optimiser, such as maximising the marginal likelihood of the observed data with respect to hyper-parameters [39].

3.2.2.5. Relevance Vector Machine.

The relevance vector machine (RVM) is a Bayesian form representing a generalised linear model of identical functional form of the support vector machine (SVM). Although SVM is a state-of-the-art technique for classification and regression, it suffers from a number of disadvantages, one of which is the lack of probabilistic outputs that make more sense in health monitoring applications. The RVM attempts to address these issues in a Bayesian framework. Besides the probabilistic interpretation of its output, it uses fewer kernel functions for comparable generalisation performance [40].

This type of supervised machine learning starts with a set of input vectors $\{x_n\}$, $n = 1,\ldots, N$, and their corresponding targets $\{t_n\}$. The aim is to learn a model of the dependency of the targets on the inputs in order to make accurate predictions of t for unseen values of x.Typically, the predictions are based on some function $F(x)$ defined over the input space, and learning is the process of inferring the parameters of this function. The targets are assumed to be samples from the model with additive noise:

$$t_n = F(x_n; w) + \varepsilon_n \qquad (3.28)$$

where ε_n are independent samples from some noise process (Gaussian with mean 0 and variance σ^2). Assuming the independence of tn, the likelihood of the complete data set can be written as:

$$p(t \mid w, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2} \| t - \Phi W \|^2\right\} \qquad (3.31)$$

where $w = (w_1, w_2,\ldots, w_M)^T$ is a weight vector and Φ is the N x (N+1) design matrix with $\Phi = [\o(t_1), \o(t_2),\ldots, \o(T_n)]^T$; $\o(t_N) = [1, K(x_n,x_1), K(x_n,x_2),\ldots,K(x_n,x_N)]^T$, and $K(x,x_i)$ is a kernel function.

To prevent over-fitting, a preference for smoother functions is encoded by choosing a zero-mean Gaussian prior distribution over w parameterised by the hyperparameter vector η. To complete the specification of this hierarchical prior, the hyperpriors over η and the noise variance σ^2 are approximated as delta functions at their most probable values η_{MP} and σ^2_{MP}. Predictions for new data are then made according to:

$$p(t^* \mid t) = \int p(t^* \mid w, \sigma^2_{MP})\, p(w \mid \eta_{MP}, \sigma^2_{MP})\, dw \tag{3.32}$$

3.2.3. Experienced-Based Approaches.

3.2.3.1. Bayesian Probability Theory.

Various methods of confronting uncertainty include Bayesian methods, the Dempster-Shafer theory, and fuzzy logic. Probability methods are generally based on Bayes' theorem (Lewis, 1986), which says that

$$f_{X,Y}(x, y) = f_{X/Y}(x/y) f_Y(y) = f_{Y/X}(y/x) f_x(x) \tag{3.33}$$

where $f_{X,Y}(x, y)$ is the joint probability density function (PDF), the marginal PDFs are $f_X(x)$ and $f_Y(y)$, and the conditional PDFs are $f_{X/Y}(x/y)$ and $f_{Y/X}(y/x)$. We compute the marginal PDFs from the joint PDF by using.

$$f_X(x) = \int f_{X,Y}(x, y)\, dy \qquad f_Y(y) = \int f_{X,Y}(x, y)\, dx \tag{3.34}$$

where the integrals are over the entire region of definition of the PDF. Therefore, we can write

$$f_{Y/X}(y/x) = \frac{f_{X,Y}(x, y)}{\int f_{X,Y}(x, y)\, dy} \tag{3.35}$$

Ssimilarly for ***fX/Y(x/y)***, the mean or expected value of a random variable is given in terms of the PDF by

$$\bar{x} = E(x) = \int x f_x(x)\,dx \qquad (3.36)$$

and the conditional mean is given by

$$E(Y/X) = \int y f_{Y/X}(y/x)\,dy = \frac{\int y f_{X,Y}(x,y)\,dy}{\int f_{X,Y}(x,y)\,dy} \qquad (3.37)$$

Clearly, all these quantities may be computed from the joint PDF. Therefore, in applications, the computation of the joint PDF, given the available data, is of high priority. Figure 18 shows the geometric meaning of the manipulations involved in obtaining the marginal and conditional PDF from the joint PDF [22].

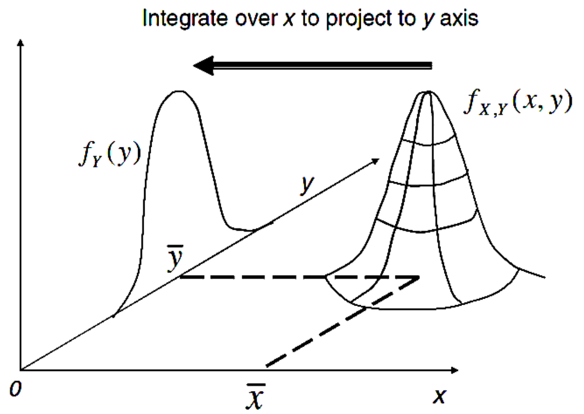


Figure. 18 Finding marginal PDFs from joint PDF.[22]

3.2.3.2. The Weibull Model: Analysis of Time to Failure.

In general, a typical Weibull probability distribution function (PDF) is defined by

$$f(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-\left(\frac{t}{\eta}\right)^{\beta}} \qquad (3.38)$$

where $t \geq 0$ represents time, $\beta > 0$ is the shape or slope parameter and $\eta > 0$ is the scale parameter of the distribution. Equation (3.38) is usually referred to as a 2-parameter Weibull distribution. Of the two parameters, the slope of the Weibull distribution, β, is very important as it determines which member of the family of Weibull failure distributions best fits or describes the data. It also indicates the class of failures in the "bathtub curve" failure modes as shown in Figure 19. The Weibull shape parameter β indicates whether the failure rate is increasing, constant or decreasing. If $\beta<1$, the product has a decreasing failure rate. This scenario is typical of "infant mortality" and indicates that the product is failing during its "burn-in" period. If $\beta=1$, there is a constant failure rate. Components that have survived burn-in will frequently exhibit a constant failure rate. If $\beta>1$, there is an increasing failure rate. This is typical for products that are wearing out. To summarise:

- ✓ $\beta < 1$ indicates infant mortality;
- ✓ $\beta = 1$ means random failures (i.e. independent of time);
- ✓ $\beta > 1$ indicates wear-out failures.

The information about the $\boldsymbol{\beta}$ value is extremely useful for reliability centred maintenance planning and product life cycle management because it can provide a clue to the physics of the failures and tell the analyst whether scheduled inspections and overhauls are needed. For instance, if $\boldsymbol{\beta}$ is less than or equal to one, overhauls are not cost effective. With $\boldsymbol{\beta}$ greater than one, the overhaul period or scheduled inspection interval can be read directly from the plot at an acceptable or allowable probability of failures. For wear-out failure modes, if the cost of an unplanned failure is much greater than the cost of a planned replacement, there will be an optimum replacement interval for minimum cost [41].

On the other hand, the scale parameter, or spread, η, sometimes called the characteristic life, represents the typical time-to-failure in Weibull analysis and is related to the mean-time-to-failure (MTTF). In Weibull analysis, η is defined as the time at which 63.2% of the products will have failed [42].

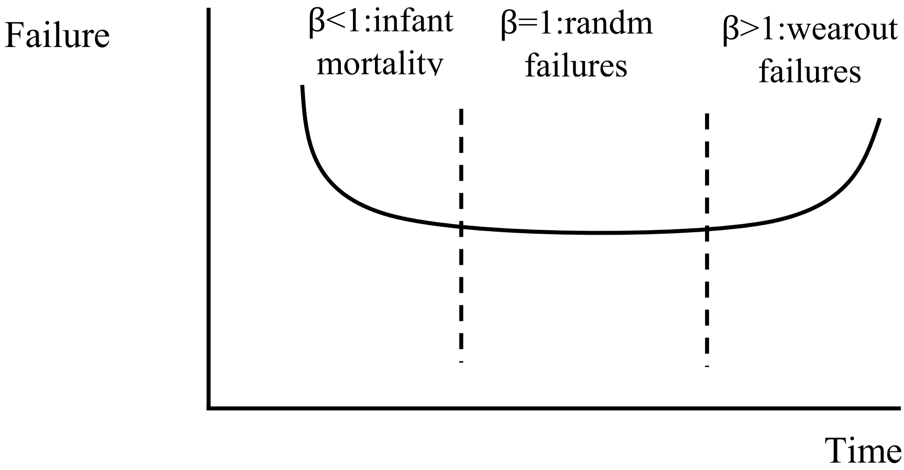


Figure 19. The “bathtub curve” failure modes.

There are basically two fitting methods for parameter estimation in widespread use in reliability analysis: the maximum likelihood estimation (MLE) and regression methods. The former involves developing a likelihood function based on the available data and finding the values of the parameter estimates that maximise the likelihood function. The latter generally works best with datasets with smaller sample sizes that contain only complete data (i.e., data in which all units under consideration have been run or tested to failure). This failure-only data is best analysed with rank regression on time, as it is preferable to regress in the direction of uncertainty. In Weibull analysis, the median rank regression (MRR) method which uses median ranking for regression fitting is often deployed to determine the shape and scale parameters for complete life data [43].

The probability of failure at time t, also called the Weibull distribution or the cumulative distribution function (CDF), can be derived from Equation (3.38) and expressed as:

$$F(t) = 1 - e^{-\left(\frac{t}{\eta}\right)^{\beta}} \qquad (3.39)$$

Thus, the Weibull reliability at time ***t***, which is ***1 – F (t) = R(t),*** is defined as:

$$R(t) = 1 - F(t) = e^{-\left(\frac{t}{\eta}\right)^{\beta}} \qquad (3.40)$$

This can be written as:

$$\frac{1}{1-F(t)} = e^{\left(\frac{t}{\eta}\right)^{\beta}} \qquad (3.41)$$

Taking two times the natural logarithms of both sides gives an equation of a straight line as in:

$$\ln\ln\left(\frac{1}{1-F(t)}\right) = \beta\ln t - \beta\ln\eta \qquad (3.42)$$

Equation (3.42) represents a straight line in the form of "***y* = *ax* + *b***" on ***log/log(Y)*** versus ***log(X),*** where the slope of the straight line in the plot is ***β***, namely, the shape parameter of Weibull distribution. Through the above transformation, the life data samples can be fitted into the Weibull model and the two Weibull parameters can be estimated.

The mean of the Weibull PDF, T , which is the MTTF in Weibull analysis, is given by:

$$\overline{T} = \eta \,.\, \Gamma\left(\frac{1}{\beta} + 1\right) \qquad (3.43)$$

where ***Γ*** is the gamma function.

Note that when ***β=1***, MTTF is equal to η. In fact, as a rough approximation, in practices of Weibull analysis where ***β*** is equal to or slightly larger than 1, the characteristic life can be approximated as MTTF. However, for ***β*** that is much larger than 1, MTTF should be calculated using Equation (3.43) [41].

3.2.3.3. *Hidden Markov Models (HMM).*

An HMM is a statistical model used to represent stochastic processes when no states where this is completely defined by the following parameters are directly observed:

- ✓ N: number of states in the model.
- ✓ M: the number of distinct observations for each state.
- ✓ A: the state transition probability distribution.
- ✓ B: the observation probability distribution of each state.
- ✓ π: the initial state distribution π.

For simplicity and clarity of presentation, a compact notation ***(λ= π, A, B)*** is used for each HMM. In practice, HMMs are used to solve three typical problems: a detection problem, a decoding problem and a learning problem.

Discrete HMMs usually consider the observations as discrete symbols and use discrete probability densities to model the transition and the observation probabilities. The problem with this approach is that in condition monitoring, the observations are typically continuous signals. In order to use a continuous observation density, some restrictions are required to insure the parameters of the probability density function can be re-estimated, as expressed by:

$$b_i\,(o) = \sum_{m=1}^{M} C_{jm}\,\xi(o, \mu_{jm}, U_{jm}), 1 \leq j \leq N \qquad (3.44)$$

In Equation (3.44), O is the observation vector, $\boldsymbol{C_{jm}}$ is the mixture coefficient for the m^{th} mixture in state $\boldsymbol{i}$ and $\boldsymbol{\xi}$ is any log concave or elliptically symmetric density with mean vector µjm and covariance matrix Ujm for the mth mixture component in state ***j***. A Gaussian density is usually used for ***ξ***; the corresponding model is called a MoG-HMM and is completely defined by: the A matrix, the ***B*** matrix and the initial probability ***π***. For an MoG-HMM, the observation matrix ***B*** is modelled by a Gaussian density with a mean ***µ***, a standard deviation ***σ*** and a mixture matrix M [44]. Figure 20 shows an example.

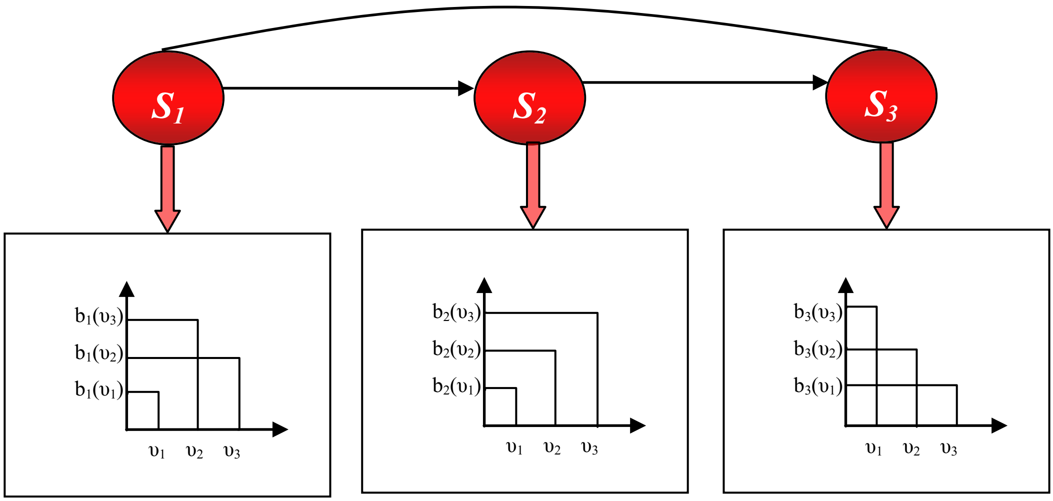


Figure 20. A three state left-to-right HMM.

Failure Prognostic Methods Based on HMMs.

There are three types of HMM-based prognostics: traditional HMMs, those based on hidden Markov models (HSMM Semi) and those using Mog-HMM.

- *The HMM case.*

The learning phase and the exploitation phase for traditional HMMs to perform diagnostics and prognostics include the following:

- *Learning phase:* the extracted features from complete monitoring histories (from the normal condition to the failure state) are transformed into HMMs by using the well known Baum-Welch algorithm. Thus, an HMM model is created for each failure. The model is then stored in a model base containing all the HMMs with a diagnostic label associated with each learned history.

- *Exploitation phase*: the online data are used as inputs to the learned models to make a prognosis of the health state. The online extracted features are used in a first step to find the model that best fits the actual observation sequence by computing the probability $P(O|\lambda)$. In a second step, the parameters of the selected model are used to assess the current health state and to estimate the RUL.

The Chapman- Kolmogorov equation (3.45) is used to re-estimate the health state after n iterations. When the predicted probability of being in the last state reaches a predefined limit ε, the RUL can be calculated (3.46) [45] as:

$$\hat{P}(n) = \pi_t A^n \qquad (3.45)$$

$$RUL = n \Leftrightarrow \hat{P}(s = s_N \mid n) = \varepsilon \qquad (3.46)$$

The HSMM case.

The problem with traditional HMMs is that the durational behaviour is usually characterised by a geometrically decaying function. This assumption is a source of inaccurate duration modelling because most real-life applications do not obey this function. TO solve this problem a model with explicit time durations, called a hidden semi Markov model, has been proposed. The learning and exploitation phases using this model are performed as follows:

- *Learning phase:* like a traditional HMM, the parameters of the HSMM are defined by using the history data and the Baum-Welch algorithm. For an HSMM, the shape is constrained to a left-to-right model. In addition, for each state, the stay durations D(si) are learned by using the Viterbi algorithm. The idea is to use the learned parameters (π, A, B) and the history data to obtain the whole observation sequence. Then, by taking into account the transition instant between the states, the duration D can be defined.

Finally, by assuming the sojourn time follows a Gaussian distribution, the mean time duration μ(D(si)) and the standard deviation σ(D(si)) of the same state concerning different histories of the same fault can be estimated ((3.47) and (3.48)):

$$\mu(D(s_i)) = \frac{1}{H}\sum_{h=1}^{H} D(s_i)_h \qquad (3.47)$$

$$\sigma(D(s_i))=\sqrt{\frac{1}{H}\sum_{h=1}^{H}\left[D(s_i)_h-\mu(D(s_i))\right]^2}$$

(3.48)

In Equations (3.47) and (3.48), D(·) stands for the stay duration, *i* is the state index, *h* is the history index and *H* is the total number of histories from a particular fault state [46].

- Exploitation phase: this phase uses the learned models and their associated stay duration. First, a competitive model selection is performed by using the raw data and the forward-backward algorithm to compute the probability P(O|λ). The label of the winner model is used to perform a prognosis of the monitored system; the actual health state is defined in the same way as in traditional HMMs.

The selected model is also used to estimate the RUL. An HSMM permits to estimate the RUL in two ways:

- By using the expression where the stay durations are merged with the state probability transitions, we get:

$$RUL_l = t_{lc} + a_{l,l+1}(RUL_{l+1}] \qquad (3.49)$$

where *l* is the actual state index, *tlc* is the state changing point and *al,l+1* is the probability transition to the next state.

- By using a more intuitive expression, adding the state duration from the current state until the last state and subtracting the time spent in the actual state, we get:

$$RUL_l = \sum_{i=l}^{N} \mu(D(s_i)) - t_l \qquad (3.50)$$

Furthermore, a confidence interval with different recovering values can be easily estimated by using the Bonferroni method, as in:

$$RUL_{\lim its} = \left[\sum_{i=l}^{N} \mu(D(s_i)) \pm cf \,.\sigma(D(s_i)) \right] - t_l \qquad (3.51)$$

In Equations (3.50) and (3.51), l is the actual state, i is the state index, N is the total number of states, μ(D(si)) is the mean time duration in the state i, σ(D(si)) is the standard deviation, tl is the time spent in the actual state, α is the confidence interval between [0,1] and cf is the confidence factor defined by using Φ, which is the cumulative distribution function of a Gaussian probability distribution [47].

The MoG-HMM based method

The originality of this method is that raw signals are processed using the Wavelet Packet Decomposition (WPD) to extract the relevant information to learn the behaviour models. Also in the generated MoG-HMM, the states' stay durations are not assumed to be a geometrically decaying function as in the HSMM case, but are learned from the monitoring data (note that multiple continuous signals are considered as observations for both learning and simulation phases, instead of the traditional mono-observation approach). Moreover, the proposed method has no limitations on the type of generated MoG-HMM (the model can be ergodic, left-right or parallel left-right).

- Learning phase: in this first phase, which is executed off-line, the raw data recorded by the sensors are processed to extract the energy of each node at the last decomposition level by using the WPD technique. These features are then used to learn several behaviour models (in the form of MoG-HMMs) corresponding to different histories related to several initial states and/or operating conditions of the component. One global left-to-right MoG-HMM is learned for each type of fault. There are X states, i.e., different asset health states. Each raw data history corresponding to a given component's condition is transformed into a feature matrix F, by using the WPD. In this matrix, each column vector (C features at time t) corresponds to a snapshot of the raw signal, and each cell fct represents the node c of the last WPD level at time t, as expressed by:

$$Raw signal \xrightarrow{WPD} F = \begin{pmatrix} f_{11} & \cdots & f_{lt} & \\ & . & .. & .. \\ & f_{cl} & \cdots & f_{ct} \end{pmatrix} \qquad (3.52)$$

with $1 \leq t \leq T$ and $1 \leq c \leq C$.

The nodal energy (features) are then used to estimate the parameters (π, A, B) and the temporal parameters (stay duration in each state) of the MoG-HMMs. This algorithm permits us to obtain the state sequence and t compute the length of time the component has been in each state of the corresponding MoG-HMM. Thus, by assuming that the state duration in each state follows a normal law, it is possible to estimate the mean duration (3.53) and the corresponding standard deviation (3.54) by computing the duration and the number of visits in each state. The Viterbi algorithm also permits us to identify the final state which represents the physical component's failure state, as given by:

$$\mu(D(s_i)) = \frac{1}{\Omega}\sum_{\omega=1}^{\Omega} D(s_{i\omega}) \qquad (3.53)$$

$$\sigma(D(s_i)) = \sqrt{\frac{1}{\Omega}\sum_{\omega=1}^{\Omega}\left[D(s_{i\omega})_h - \mu(D(s_i))\right]^2} \qquad (3.54)$$

In equations (3.53) and 3.54), D(·) stands for the visit duration, i is the state index, ω is the visit index and Ω corresponds to the total of visits. A compact representation of each learned MoGHMM used to perform diagnostic and prognostic is given by the following expression:

$$\lambda = (\pi, A, B, \mu(D(s_i)), \sigma(D(s_i)), S_{final}) \qquad (3.55)$$

where λ is the fully defined model and Sfinal is the final state (corresponding to the end of the considered condition monitoring history).

- Exploitation phase: this phase, which is performed online, consists of exploiting the learned models to detect the component's current condition (using the Viterbi algorithm) and to compute the corresponding RUL. The processed data and the extracted nodal energy using the Wavelet toolbox from Matlab® are continuously fed to the learned models to determine the global MoG-HMM which best fits the observed sequence. The diagnosis is then made, and the current health state is defined. The selection process is based on the calculation of a likelihood $P(O|\lambda)$ of the model over the observations (HMMs problem 1). Using this last calculation and the current model state, along with the stay durations learned in the off-line phase, the component's RUL and its associated confidence value can be estimated.

  The generated MoG-HMMs are used during the online phase to estimate the RUL and the associated confidence value of the physical component by using a dedicated procedure with the following Steps:

  a) Detection of the appropriate global left-to-right general MoG-HMM that best fits and represents the online observed sequence of nodal energy. The diagnostic label of the selected model is used to diagnose the current condition.

  b) Choice of the nearest RUL model knowing the active failure mode.

c) Identification of the current state of the selected RUL model.

d) Identification of the critical path, from the current state to the end state. The idea is to identify all the non-zero probabilities in the transition matrix as potential transitions and then to choose the minimal path amongst all the possible ones (Figure. 21) with only one visit per state.

e) Estimation of the RUL by using the temporal parameters of the stay duration in each state. In addition, a confidence value over the RUL is calculated based on the standard deviation values of the stay durations and the Bonferroni confidence interval [48]. This is expressed as:

$$RUL_{lower} = \left[\sum_{i=current\,state}^{N} \mu(D(s_i)) \pm cf \,.\sigma(D(s_i)) \right] \quad (3.56)$$

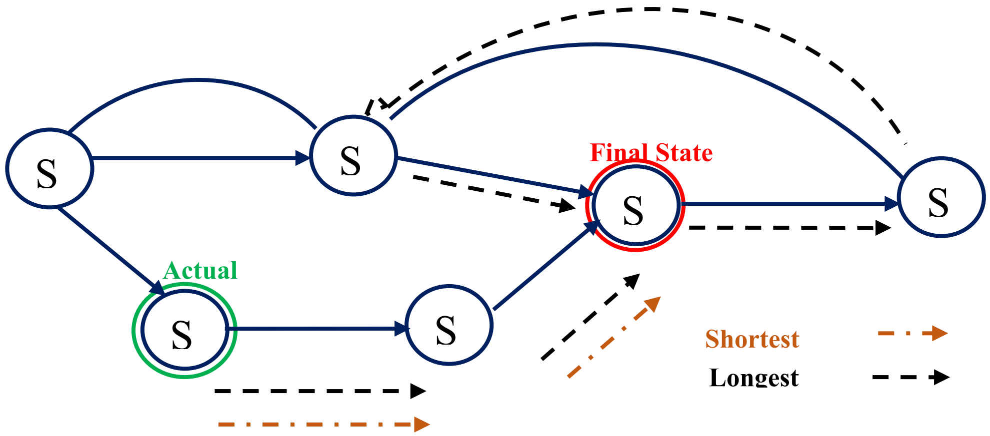


Figure 21. Path Estimation

Part III

METHODOLOGY

For purposes of this paper only discuss the calculation of the remaining useful life of assets into three types: rotating machines (compressors, pumps,

turbines, etc.), structures (bridges, buildings, etc.) and complex systems (cars, planes, etc.). It goes to compares the various techniques used to estimate the RUL of each asset; it determines if the RUL for that asset can be estimated by using one specific technique or several different techniques.

4.1. Rotating Machines.

Rotating machinery is a major and critical component of many mechanical systems found in industrial plants, air and ground transportation vehicles and many other applications. Rotating elements have unique but predictable characteristics in both performance and acoustics. For example, they exhibit high harmonic oscillations when on the verge of failure. Thus, fault diagnosis is commonly employed as a maintenance methodology for rotating machinery.

Recently, maintenance methodology has shifted toward Condition-Based Maintenance (CBM). CBM employs a continuous maintenance methodology to maximise availability of equipment. It is accomplished by continuously monitoring equipment conditions and performing maintenance actions only when faults exist. At the operational level, CBM alerts operators to evolving faults and allows maintenance actions to be taken before catastrophic failure occurs.

A schematic of a typical CBM system is shown in Figure. 22. The key components are machinery, sensors, fault detectors, fault classifiers, predictive models, model/data fusion, and outputs.

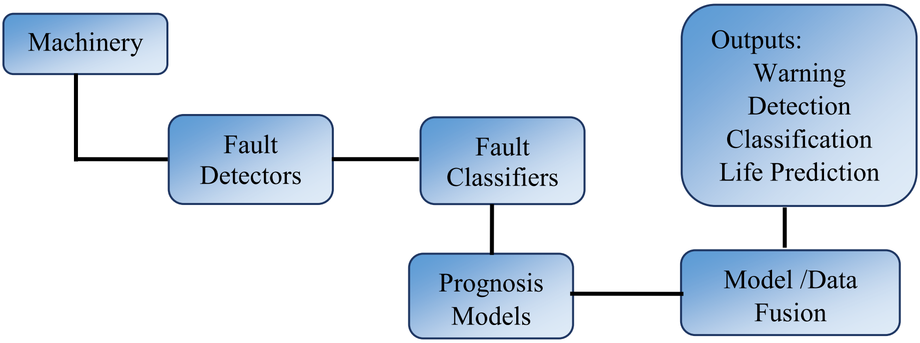


Figure 22. Key elements of condition-based maintenance system.

4.1.1. Sensing Techniques and Sensors.

- *Monitoring Parameters.*

The monitoring of machinery performance and its operating condition requires careful selection of measured parameters. The most common parameters are vibration and acoustic signatures, temperature, pressure, motor current, wear debris and electrostatic exhaust measurements. To detect failure inception, deterioration mechanisms are continuously monitored and the initiation of cracks, fractures, and other failures, such as shaft cracks, seal leakage, and corrosions are identified.

- *Sensing Techniques.*

Sensing techniques are constantly being developed. State-of-the-art techniques include: active interrogation approaches, nondestructive evaluations such as acoustic emission and stress wave monitoring, holographic imaging, oil debris analysis, chemical composition and analysis, use of in situ and embedded microsensors, electrostatic exhaust measurements, remote sensing, and electrical resistance measurements. Also important are the optimal placement of sensors and a sensing strategy to monitor small changes in state variables that could capture fault precursors and features. Virtual sensors using the simulation and modelling of signals based on limited data and measurements have the potential to reduce the number of sensors and, thus, the cost.

- *Sensors*

Recent advances in materials and sensor development, combined with new manufacturing techniques (e.g., nanofabrication), have allowed manufacturers to build rugged sensors that are both affordable and small. The new materials and processing include piezoelectric elements, piezofilm, piezocomposites, optical fibers, and micro-electromechanical systems (MEMS). In addition, biosensors are being developed to sense and identify chemical species; this will be useful for fault diagnostics [49].

4.1.2. Feature Extraction.

- *Detection and Identification.*

With the increasing need to identify and locate incipient failures, new signal processing techniques have been implemented to enhance detection and identification of fault precursors and features using signal enhancement or background noise removal and signal detection techniques. The signal processing techniques include cepstral analysis, time-frequency analysis, nonlinear dynamic techniques, and higher and lower order spectral analyses. Time-frequency analyses such as the Wigner distribution and wavelet transforms have been used and proved to detect and identify fault features. Sources of higher harmonic oscillations in rotating machinery have been found to be caused by imbalance, misalignment, and nonlinear excitation. In addition, nonlinear dynamic techniques and lower and higher order spectral (polyspectral) analyses techniques can be used to detect and identify higher harmonic oscillations, as well as fault features. In particular, alpha-stable distributions, a lower-order statistics method, can detect impulsive type features embedded in a non-Gaussian signal.

- *Feature Classification.*

Once fault features are detected, identified and extracted using signal processing techniques, the fault features are transformed to fault vector spaces where classification is performed. Classification techniques include conventional statistical methods, neural networks, fuzzy logic, and pattern recognition. Neural networks have been commonly applied to fault vectors to train and classify the various faults. Fuzzy logic, a rule-based reasoning approach, has been used successfully to classify faults. Efforts are ongoing to combine neural networks and fuzzy logic, the so-called neuro-fuzzy network approach, to classify fault precursors and features. Pattern recognition still remains as one of the classification techniques for faults.

4.1.3. Prognosis Model.

Failure analysis, which plays an important role in understanding how components fail, forms the basis of the failure predictive models. As mentioned previously, failure mode identification and classification is a technical issue. Though failure mechanisms have been studied for a long time, understanding how failures occur and theoretical developments in predicting mechanical failures are far from satisfactory. In particular, there is a need for physics-based models relating fundamentals with experimental data to predict the remaining useful (safe operating) life of structures and components. The challenge is to model fault initiation and propagation as a unified approach and then predict state variables relating to mechanical response or vibration. A statistical approach has been used to predict the fatigue life expectancy of structural components subject to cyclic loading. This approach deals with the statistical variability in such quantities as material properties and initial flaw sizes, as well as the statistical nature of the fatigue process. The estimation of the remaining useful life can be applied based on data and/or experience using ANNs, fuzzy logic and regression; for experience-based approaches as mentioned above the most widely used is Weibull.

4.1.4. Data/Model Fusion.

Data monitoring and predictive models can be fused to enhance decision-making and to provide reliable information and warning to operators. Occasionally, signals from multiple sensor inputs and models can be contradictory; therefore, selec tion of the proper data and integration or fusion of this information are critical in the reasoning and decision-making processes [50].

In summary, a number of approaches can be used to estimate the remaining useful life of rotating machines. In experience-based approaches, the most commonly used method is Weibull. Model-based approaches include Paris' law for crack growth modelling, fatigue spall initiation, and progression model. Time series predictions use neural networks, fuzzy neural networks, regression analysis and fuzzy logic.

4.2. Structures.

Today, many of the world's structures are aging and much of its civil infrastructure requires maintenance, rehabilitation or replacement. In practice, the main method used to identify structural deterioration is visual inspection. Inspections provide the engineer with data on the number of cracks and the rate of crack growth, but quantitative data are often necessary to distinguish between structures that can be kept in service and those requiring replacement or modernisations. There are a few studies on patterns of maintenance replacement for bridges and buildings. Most are based on ranking factors against failure (evaluation of capacity). One of the most critical types of structural deterioration is induced fatigue fracture; therefore, the maintenance strategy based on the remaining useful fatigue life is increasingly used by maintainers. This is usually based on a combination of load voltage, Miner's rule, and design fatigue curve.

4.2.1. Implementation of CBM to Instrumented Bridges y Buildings.

The CBM for a bridge can be initiated according to the degradation state of the structure, as monitored by various sensors. Once the degradation characteristic crosses a specified threshold, maintenance actions are triggered. In CBM, the degradation characteristic measures of the structure are obtained and updated with the use of real-time data from both structural and durability health monitoring systems. This system enables the early detection of degradation mechanisms by continuously updating information on the performance of the structure in load-carrying capacity and durability resistance, along with degradation effects due to aging and environments, thereby allowing the overall degradation mechanisms to be understood. The determination of the deterioration rates for bridges is essential to evaluate the effectiveness of maintenance/repair options. As the deterioration rate is a function of time, the conventional inspection approach, due to its ad hoc nature and lack of continuous time-history supporting data, is inadequate. Long-term real-time monitoring systems and analytical tools are required to estimate the deterioration rates at key locations/components of bridges under in-service condition.

Figure 23 illustrates various potential degradation mechanisms for bridges (reinforced concrete structures). The decision making on the maintenance of a reinforced concrete structural component depends on two factors: the deterioration status of its embedded steel tendons or steel reinforcements and its load-carrying capacity. It therefore requires both durability and structural health monitoring. For durability health monitoring, analytical tools have to be developed to evaluate the formation and propagation of cracking and the time for the spalling of concrete cover based on the monitoring data from corrosion sensors (corrosion potentials, corrosion currents, concrete resistivity, linear polarisation resistance, concrete temperature, concrete relative humidity, and gas concentration in deck and piers). After developing such analytical tools, a linkage must be established to transfer the monitoring/analysis results into an updatable and evolving rating system to facilitate the CBM decision making and to prioritise maintenance activities. The bridge rating system is generally developed in terms of the criticality and vulnerability indices that can be formulated/updated using the design information, inspection results and monitoring data. As the structural condition is obtained from the CBM, a new synthetic rating system (Rating Cubic) will be formulated as a rating matrix in terms of the criticality rating, vulnerability rating, and condition rating of the structural components. The action and prioritisation of condition-based inspection/maintenance for structural components will be based on the results (which are dynamically updated over time) of the synthetic rating (the rating results in Cube 1 correspond to the highest priority while the rating results in Cube 8 correspond to the lowest priority) [51].

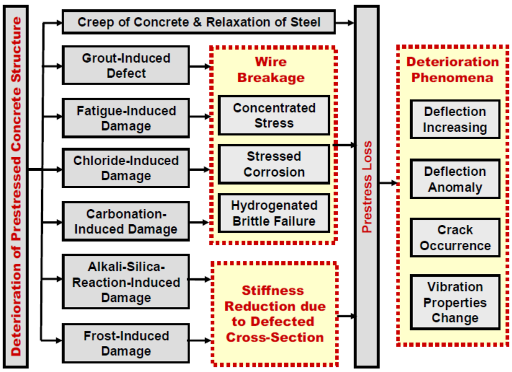


Figure 23. Degradation Mechanisms for Reinforced Concrete Structures

4.2.2. Remaining Fatigue Life Estimation of Members.

Suppose a component is subjected to a certain stress amplitude or stress range i σ for ni number of cycles at load level i and Ni is the fatigue life (failure number of cycles) corresponding to σ_i . Hence, the residual life at load level *i* can be obtained as (N_i-n_i). The stress σ_{eqi}) which corresponds to the failure life (Ni−ni) is named as i[th] level damage stress amplitude or stress range (otherwise introduced as stress amplitude or stress range relevant to the residual life). The new damage indicator, D_i is stated as,

$$D_i = \frac{\sigma_{(i)eq} - \sigma_i}{\sigma_u - \sigma_i} \qquad (4.57)$$

where σ_u is the intercept of the Wöhler curve with the ordinate at one-quarter of the first fatigue cycle. Furthermore, it can be stated that σ_u is the ultimate tensile strength amplitude or range for rotating bending test-based S-N curves and the ultimate shear strength amplitude or range for torsional fatigue test-based S-N curves.

In the first cycle, the damage stress amplitude or range $\sigma_{(i)eq}$ is equal to applied stress σ_i and the corresponding damage indicator becomes $D_i=0$. According to the proposed methodology, current damage must then be transformed to the next load level. Therefore, in the last cycle, the damage indicator becomes $D_i=1$ when $\sigma_{(i)eq}$ is equal to σ_u . Therefore, the damage indicator is normalised to one $(D_i=1)$ at the fatigue failure of the material, and the same procedure is followed until $D_i=1$. Here, the defined fatigue failure is the time taken for the occurrence of the first throughthickness crack at the location of maximum stress of the structural component. In the case of railway bridge components, it can probably be taken as the time taken for the initiation of a crack near a connection (rivet or bolt).

4.2.3. Identification of Critical Members/ Connections

The members with the lowest remaining fatigue life of each member set (set of members with the same load capacity) are called “critical members” in this study. Generally, they are to be given more attention in the member replacement based maintenance. From the previously obtained remaining fatigue lives, it is easy to identify these critical members.

The connections joining the previously obtained critical members are called “critical connections”.

4.2.4. Remaining Fatigue Life Estimation of Critical Connections.

The stress concentration effect in connections between the primary members of bridges is a main reason for fatigue damage. Most of these connections are subjected to multiaxial fatigue. To capture this effect at riveted connections or discontinuities, the detail class (BS 5400, 1980) of riveted connection based Wöhler curves is considered in a previous life estimation. However, the variation of real rotational fixity, clamping force and geometry at the connection may change the real stress distribution at the connections. Such changes may results in over or under predictions of the estimated fatigue life of the corresponding member. As a result, replacement of members based on previously determined remaining lives may not be an appropriate maintenance procedure. Replacement of members based on fatigue lives of critical connections is a more appropriate strategy. This section describes the methodology to estimate the remaining fatigue life of such connections.

Initially, all critical connections should be investigated non-destructively to determine the current condition using various tests, including X-ray, ultrasonic, magnetic particle, radiographic examinations etc. The connections which do not illustrate significant change from the initial state or condition are not subjected to any unexpected stress concentration. Other connections which have been subjected to significant change may need to have their remaining fatigue life assessed to determine the degree of criticality.

The remaining fatigue lives of the connections not subjected to significant deviation from the initial condition are finalised in the same way as the lowest remaining fatigue life of the member joined to a particular connection. The remaining fatigue lives of other connections, where the conditions have been significantly changed, must be evaluated based on the current geometric condition, secondary stress distribution etc.

This procedure is based on a newly proposed multiaxial fatigue model. Initially, the accumulated plastic strain per each stabilised cycle, expressed as

$$\varepsilon_S^{pc} = \frac{4}{\sqrt{3}} \frac{2k^* - k_{\max} - k_{\min}}{c} \quad , \qquad (4.58)$$

has to be obtained from the secondary stress analysis of the connection or part of the member (sub model) where c = b + 2η . The b and η are the mesoscopic linear hardening modulus and the shear modulus respectively. The k^* is the radius of the smallest hypersphere which contains the entire history of the macroscopic deviatoric stress amplitude of the stabilised cycle. The k_{max} and $k_{\min}$ are the maximum and the minimum values of mesoscopic yield stresses that can be reached during the loading cycle. The fatigue life is calculated from the new damage indicator as shown:

$$D_i = \frac{(\varepsilon_s^{pc})_{(i)eq} - (\varepsilon_s^{pc})_i}{(\varepsilon_s^{pc})_u - (\varepsilon_s^{pc})_i} \tag{4.59}$$

where $(\varepsilon_s^{pc})_i$ is accumulated plastic meso-strain per stabilised cycle of ni number of cycles at load level i. The N_i is the fatigue life (number of cycles to crack nucleation) corresponding to $(\varepsilon_s^{pc})_i$ and can be estimated from

$$N = A(\varepsilon_s^{pc})^{-\xi} \tag{4.60}$$

where ξ and A are material parameters to be determined from fatigue tests. The accumulated plastic meso-strain $(\varepsilon_s^{pc})_{(i)eq}$, which corresponds to the failure life (Ni−ni), is named as ith level damage accumulated plastic meso-strain. The $(\varepsilon_s^{pc})_u$ is the accumulated plastic meso-strain which corresponds to one-quarter of the first fatigue cycle. At this point, according to the proposed methodology, current damage has to be transformed to the next load level. As in the previous sequential law in Section 4.2.2, the damage indicator is normalised to one (Di=1) at the fatigue failure of the material, and the same procedure is followed until Di=1.

4.2.5. Member Replacement/ Strengthening Scheme.

The lowest remaining fatigue life of the connections describes the remaining fatigue life of the bridge. Once the age of the bridge reaches this value (when bridge life becomes zero), it is advisable to replace the corresponding critical member with a new member with longer fatigue life. At the same time, the associated connection should be strengthened. After this essential repair, a new sequence for future member replacement should be obtained by following a similar procedure. This type of maintenance strategy extends the service life (fatigue capacity) of the bridge in the safest manner.

In summary, the most widely used approach for estimating the remaining life of bridges (and buildings) is the model-based technique to estimate the lifetime of fatigue [52].

4.3. Complex System.

System maintenance is a challenge for manufacturers who produce complex systems, like aircraft or cars, as good maintenance can improve the reliability, security, safety and the final cost of their products. To improve the maintenance capabilities of a complex system, an onboard health monitoring system must be deployed; it should provide enough data to indicate what type of maintenance operation is required. These types of systems are assemblages of heterogeneous components (continuous, discrete, hybrid components) that require many different types of techniques to monitor their health state. Before a global health monitoring system can be built, there are many difficulties to deal with.

First of all, in classical maintenance, the objective is to replace components (also called Line Replacement Units, LRU for short) of the system that are faulty in the sense that they do not perform the set of functions they are supposed to perform. The monitoring system requires diagnostic capabilities to determine online the faulty components [22].

The second difficulty is the optimality of the maintenance. To decrease maintenance costs, it is necessary to perform preventive maintenance which requires embedding prognostic capabilities in the health monitoring system to determine the ageing state of the system

The third difficulty concerns the description of the system itself. Without a good level of abstraction to describe the system knowledge to be used in the health monitoring system, it is impossible to guarantee the global consistency of the health monitoring architecture. The objective is, thus, to get an abstracted description that is homogeneous [53].

We must take into account that complex systems, as mentioned above, must have a continuously monitored system (i.e., CBM). This is done by sensors able to capture all information related to the health status of the entire system; these data are then used to make a prognosis and estimate the remaining useful life. The prognostic methods commonly used to explain complex systems are explained in the following subsections.

4.3.1. Prognosis Methods.

As mentioned previously, prognostics basically consists of estimating at time $t_{prog} \geq t_j$ the time $t_{j+1} \geq t_{prog}$ of the future mode change relying on the ageing models available in the system mode m_j^{Σ}. A set of ageing models (or life models) is available for each private parameter $pp^{i,k}$ ε PP^i. In the following, $pp^{i,k}$ is generically noted pp for the sake of simplicity. Ageing models of a private parameter describe the evolution of the parameter value pp with environmental constraints (i.e., temperature, humidity, vibration, stress conditions). It follows that an ageing model $ag_x^{i,k}$ can be represented as an algebraic relation depending on operating conditions of the system mode m_x^{Σ}. According to the current system mode m_i^{Σ} that defines the range of the private parameters of the system, the well-suited ageing models are selected for each private parameter pp ε PP. The remaining time until the parameter pp becomes faulty is noted as remaining time to fault or RTF (pp). A fault probability for each private parameter in PP is established from the selected ageing models.

Basically, let f_{pp} denote a probability density function (pdf for short) representing the fault probability of a private parameter pp in the mode m_x^{Σ}, and P_{max} be the maximal fault probability value acceptable for the parameter pp; the remaining time to fault (RTF) of pp consists, then, of determining the time t_p for which the fault probability has reached the threshold P_{max} as shown by:

$$RTF(pp) = t_p \quad such\,that \int_0^{t_p} f_{pp}(t)dt = P_{\max} \tag{4.61}$$

It follows that $t_{j+1} = t_{prog} + \min(RTF(pp), pp\ \varepsilon\ PP)$ and the system mode m_{j+1}^{Σ} is such that pp is out of range [54].

Generic Modelling For Prognostics.

Private parameters usually represent physical attributes of the component and are totally heterogeneous. The difficulty is to find a common representation of the prognosis for any private parameter that also must be as flexible as possible to represent any type of probability density functions. For these reasons, the Weibull distribution is often used. In this case, the probability density function is:

$$W(t,\beta,\eta,\theta) = \frac{\beta}{\eta}\left(\frac{t-\theta}{\eta}\right)^{(\beta-1)} e^{-\left(\frac{t-\theta}{\eta}\right)^{\beta}} \tag{4.62}$$

where β characterises the shape of the distribution, η characterises the scale and θ characterises the location of the distribution. For a given private parameter $pp\ \varepsilon\ PP^i$ and a given probability threshold P_{max}, we then get:

$$RTF(pp) = t_p \quad such\,that \int_0^{t_p} W(t,\beta_{pp},\eta_{pp},\theta_{pp})dt = P_{\max} \tag{4.63}$$

Because β_{pp}, η_{pp}, θ_{pp} fully characterise the failure probability distribution, they model how the parameter pp is ageing. The description of the characteristics β_{pp}, η_{pp}, θ_{pp} relies on the ageing models associated with pp. Whatever the technique used to obtain such models, the available knowledge is then characterised by algebraic relations that define β_{pp}, η_{pp}, θ_{pp}: $\beta_{pp} = ar_\beta(ip^{i,1},.....,ip^{i,n}), \theta_{pp} = ar_\theta(ip^{i,1},...., ip^{i,n})$, $n_{pp} = ar_\eta(ip^{i,1},...., ip^{i,n})$.

- *Functional Prognosis.*

The RUL of the system is defined as the remaining time until the system cannot perform successfully its full set of complex functions that rely on the basic functions FU^i of any component C^i ε Comps. In systems like aircraft, for instance, the system functions usually rely on redundant implementations of the basic functions (the functions of one component can be performed by another). That is why prognostics relies on the functional view of the system. In order to get a prognosis of the availability of any system function in the future, the prognosis of the availability of every basic function F ε FU^i implemented on every component C^i of the system must be acquired. The ageing of F naturally depends on the ageing of the private parameters of the component implementing F. The remaining time until a basic function F becomes failed is noted as *ettf (F)* which stands for estimated time to failure [55]:

$$ettf(F) = t_f \quad such\,that \int_0^{t_p} W(t, \beta_F, \eta_F, \theta_F)dt = P_F \tag{4.64}$$

where P_F is a probability threshold and βF , ηF , θF are fully determined by a combination of the characteristics $\{ \beta_{pp}, \eta_{pp}, \theta_{pp} \}_{pp2PP(F).}$ By extension, the ettf (Fsys) of any system function F_{sys} is defined as *ettf (F)* where *F* is the basic function whose availability is necessary for Fsys to be available and whose ettf (F) is minimal. The RUL of a component is then $RUL(C^i) = min(ettf\,(Fu^{i,j}), Fu^{i,j}\ \varepsilon\ FU^i)$ and the RUL of the system is estimated to the minimal ettf (Fsys). Finally, maintenance can be scheduled based on the ettf (Fsys) and consists at least of replacing a component Ci implementing the basic function $F\ \varepsilon\ FU^i$ by a new one so that *ettf (Fsys)* is increased.

In summary, prognosis to implement complex systems should take into account that on many occasions, different prognoses must be applied separately, as a complex system consists of several subsystems, and these subsystems are, in turn, composed of components. As given above, the most common way to implement prognosis is using the Wiebull distribution, but other methods can also be applied, such as fuzzy neural networks. It is important to know that each complex system has a different mathematical model; thus, when we want to estimate the remaining useful life, we must first make a mathematical model of the system.

Part IV
RESULTS AND CONCLUSIONS

Results 5

In this chapter a comparison chart where you may view different prognosis techniques used for the estimation of the remaining useful life of the assets listed above will be made.

Table 1. Approaches to Prognosis

Approaches Prognosis in Different Industrial Assets			
Assets			Approaches
A	B	C	
Model Based			
✶	✶		Particle Filtering
✶	★		Physics- based fatigue models
			ARMA, ARMAX, and ARIMA Methods
Data Driven			
✶			Linear Regression
✶		✶	Artificial Neural Networks
✶		✶	Fuzzy Logic Systems
			Gaussian Process Regression
			Relevance vector machines (RVM)
Experienced Based			
			Bayesian Probability Theory
✶	✶	✶	The Weibull Model: Analysis of Time to Failure
			HMM
Models			
		✶	Mathematical

A: rotating machines: Pumps, Turbines, Compressors, Motors.

B: Structures: Bridges and Buildings

C: Complex Systems: Planes, Cars, etc.

As Table 1 shows, prognosis is a technique that can be applied in different industrial assets. Rotating machines only require particle filters and physics-based fatigue; other techniques are rarely used for this type of asset. In terms of data-based models, ANNs and fuzzy logic may be used; on occasion, linear regression is used. For experience-based models, the most common is Weibull because it is the most conventional.

Models for structures such as bridges and buildings use fatigue because as stated in the previous chapter, the monitoring applied to these assets is used to calculate the remaining useful life of fatigue as this is usually the main problem with such assets. But depending on the type of structure, Weibull can be used.

Finally, the calculation of the remaining useful life of more complex systems uses mathematical models; as stated earlier, most complex systems require the development of a model, because these systems are not alike. Similar systems often use techniques such as neural networks, fuzzy networks, and occasionally Weibull.

Conclusion 6

This work summarises how a CBM program is applied to estimate remaining useful life. Various techniques, models and algorithms have been reviewed following the three main steps of a CBM program: data acquisition, data processing and maintenance decision-making. Although advanced maintenance techniques are described in the literature, there are two common extremes in industry. One extreme is to always adopt a run-to-failure (breakdown) policy. The other is to always apply an as-frequent-as-possible maintenance policy. The two can be applied to some special cases with satisfactory results, but in many situations, especially when both maintenance and failure are very costly, CBM is a better choice. Expert knowledge in both application and theory is required for choosing the best maintenance policies.

Advanced maintenance technologies have not been well implemented in industry for the following possible reasons: (1) lack of data due to incorrect data collecting approach, or even no data collection and/or data storage; (2) lack of efficient communication between theory developers and practitioners in the area of reliability and maintenance; (3) lack of efficient validation approaches; (4) difficulty of implementation due to frequent change of design, technologies, business policies and management executives. Next generation prognostic systems will likely focus more on various aspects of continuous monitoring and automate the prognostics.

This work has also discussed many concepts associated with prognostics in differente industrial assets. It has reviewed prognostic approaches and implementation issues including current CBM developments, and has provided several examples. It has discussed the variations in data, modelling and reasoning for the different prognostic approaches and illustrated these in different assets: rotating machines (pumps, turbines, compressors, motors), structures (bridges and buildings) and complex systems (planes, cars etc.). Data availability, dominant failure or degradation mode of interest, modelling and system knowledge, accuracies required and criticality of the application are some of the variables that determine the choice of prognostic approach. The ability to predict the time to conditional or mechanical failure (on a real-time basis) is of enormous benefit; health management systems that can effectively implement the capabilities presented herein offer a great opportunity in terms of reducing the overall Life Cycle Costs (LCC) of operating systems and decreasing maintenance tasks.

Further Research 7

In future work, it would be interesting to study how to apply prognosis in other types of industrial assets. In addition, it would be useful to conduct a experimental study using a real case with the approaches outlined here, in order to compare the results. Studies could also compare two approaches to the same data. Finally, the hybrid approaches could be explored in greater detail, especially their applicability in industry.

For CBM, it would be interesting to study the progress and trends of this type of program at present (data cleaning, data quality, issues of context awareness), and to explore the use of data when estimating the remaining useful life. Another future work could study the advances in sensors and data acquisition systems, along with the computer programs used for this purpose.

References 8

[1]. R. C. Mishra, K. Pathak, "*Maintenance Engineering and management*", PHI Learning Pvt. Ltd., 1/8/2004 - 228 pages.

[2]. Otilia E, DRAGOMIR. Rafael GOURIVEAU. Florin DRAGOMIR. Eugenia MINCA. Noureddine ZERHOUNI, "*Review of prognostic problem in condition– based maintenance*", hal-00418761, version 1-21, Sep 2009.

[3]. A.K.S. Jardine, D. Lin and D. Banjevic, " *A review on machinery diagnostics and prognostics implementing condition – based maintenance*", *Mech. Sys. & Sig. Pro.,* Vol. 20, pp. 1483- 1510, 2006,

[4]. Jezdimir Knezevic, "*Maintenence*". Ed. Isdefe, 1996.

[5]. J. Lee, R. Abujamra, A.K.S. Jardine, D. Banjevic, " *An integrated platform for diagnostics, prognostics and maintenance optimization*", in: The IMS, 2004 International conference on advances in maintenance and in modelling, simulation and intelligent monitoring of degradations, Arles, France, 2004.

[6]. Gerardo Trujillo C. – Noria Latín América, "Condition Monitoring - A Strategy Integration Technologies", 1st Mexican Congress of reliability and maintenance, October, Mexico, 2003.

[7]. Albert, H.C. Tsang "*Condition – based maintenance: tools and decision making*" Vol,1 No, 3, 1995, pp 3-17

[8]. Mike Sondalini, Managing Director, Lifetime Reliability Solutions HQ, "*How to use Condition Based Maintenance Strategy for Equipment Failure Prevention*" http://www.lifetime-reliability.com

[9]. T. Brotherton, G. Jahns, J. Jacobs and D. Wroblewski, " *Prognosis of faults in gas turbine engines*", in Proc. IEEE International conference on aerospace, Vol. 6, pp. 163-171, 2000.

[10]. Pecht, Michael G. (2008). "*Prognostics and Health Management of Electronics".* Wiley. ISBN 978-0-470-27802-4..

[11]. O. Dragomir, R. Gouriveau, N. Zerhouni and F. Dragomir, "*Framework for a distributed and hybrid prognostic system*", in: 4th IFAC Conf. on Manag. and Control of Prod. and Logistics, 2007.

[12]. Yu, Wei Kufi; Harris (2001). "*A new model based on stress fatigue life of ball bearings*, " Tribology Transactions, 11 18 doi : 10.1080/10402000108982420.

[13]. J. Luo, M. Namburu, K. Pattipati, L. Qiao, M. Kawamoto and S. Chigusa, "*Model-based prognostic techniques*", in: Proc. of IEEE Autotestcon, pp. 330-340, 2003.

[14]. Carl S. Byington, Michael J. Roemer, Thomas G. " *Prognostics enhancements to diagnostic systems for improved condition based maintenance*" IEEE 2002.

[15]. Shane Butler "*Prognostic Algorithms for Condition Monitoring and Remaining Useful Life Estimation*" A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy, September 2012.

[16]. Z. Tian. *"An artificial neural network method for remaining useful life prediction of equipment subject to condition monitoring. Journal of Intelligent Manufacturing",* N/A:1{11, 2009.

[17]. W. Q. Wang, M. F. Golnaraghi, and F. Ismail. "*Prognosis of machine health condition using neuro-fuzzy systems. Mechanical Systems and Signal Processing",* 18(4):813{831, 2004.

[18]. Wikipedia the free encyclopedia. "*Prognostics, Hybrid approaches*" http://en.wikipedia.org/wiki/Prognostics#Hybrid_approaches

[19]. Rafael Gouriveau, "*An introduction to Prognostics*" FEMTO-ST Institute, UMR CNRS 6174 FCLAB Research Federation, FR CNRS 3539.

[20]. M. Orchard, G. Kacprzynski, K. Goebel, B. Saha, and G. Vachtsevanos. *"Advances in uncertainty representation and management for particle filtering applied to prognostics. In Prognostics and Health Management",* PHM 2008. International Conference on, pages 16, October 2008.

[21]. M. E. Orchard and G. J. Vachtsevanos. *"A particle filtering approach for on-line fault diagnosis and failure prognosis. Transactions of the Institute of Measurement and Control",* 31:221 246, 2009.

[22]. G. Vachtsevanos, F. Lewis, M. Roemer, A. Hess, B. Wu, "*Intelligent Fault Diagnosis and Prognosis for Engineering Systems*", John Wiley and Sons Inc., Hoboken, New Jersey, 2006.

[23]. J. Luo, M. Namburu, K. Pattipati, L. Qiao, M. Kawamoto, S. Chigusa, "*Model-Based Prognostic Techniques*" , Anaheim, CA, United States: 2003, Institute of Electrical and Electronics Engineers Inc., Piscataway, NJ, United States, 2003, pp. 330–340.

[24]. J.Z. Sikorska, M.Hodkiewicz, L.Ma, "*Prognostic modelling options for remaining useful life estimation by industry"* Mechanical Systems and Signal Processing, 2010.

[25]. W.R. Blischke, D.N. Murthy, Reliability:*"Modelling, Prediction and Optimization",* John Wiley & Sons Ltd., 2000.

[26]. M.Todinov, Reliabilityand Risk, "*Models Setting Reliability Requirements*" John Wiley & SonsLtd., Chichester England, 2005.

[27]. J. Neter, M.H. Kutner, C. J Nachtsheim, and W. Wasserman, "*Applied Linear Statistical Models"* Irwin, 1996.

[28]. S. Agatonovic-Kustrin and R. Beresford, "*Basic concepts of artificial neural network (ANN) modelling and its application in pharmaceutical research*". Journal of pharmaceutical and biomedical analysis, Vol. 22, 717-727, 2000.

[29]. Y. Tan and A. Van Cauwenberghe, *"Neural-network-based d-stepahead predictors for nonlinear systems with time delay*". Engineering applications of artificial intelligence, Vol. 12, 21-35, 1999.

[30]. Siti Azirah Asmai, Abd. Samad Hasan Basari, Abdul Samad Shibghatullah, Nuzulha Khilwani Ibrahim and

Burairah Hussin, “*Neural Network Prognostics Model for Industrial Equipment Maintenance”* Department of Industrial Computing, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia, 2011.

[31]. Y. Sun, V. Babovic, and E. Chan, "*Multi-step-ahead model error prediction using time-delay neural networks combined with chaos theory*". Journal of Hydrology, Vol., 2010.

[32]. J. Heaton, “*Introduction to Neural Networks for Java*”, Heaton Research Inc, 2008.

[33]. W. Wang, F. Golnaraghi, F. Ismail, “*Prognosis of machine health condition using neuro-fuzzy systems*,” Mech. Syst. Signal Process., vol. 18, pp. 813-831, 2004.

[34]. C. Lin and C. Lee, *Neural Fuzzy System: A Neuro-Fuzzy Synergism to Intelligent Systems.* Upper Saddle River, NJ: Prentice-Hall, 1996.

[35]. Chaochao Chen, Bin Zhang, George Vachtsevanos, *“Prediction of Machine Health Condition Using Neuro-Fuzzy and Bayesian Algorithms”,* School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332 USA, Impact Technologies LLC, Rochester, NY, 14623.

[36]. C-H. Lee and C-C. Teng, *“Identification and control of dynamic systems using recurrent fuzzy neural networks,” IEEE Trans. Fuzzy Syst.,* vol. 8, no. 4, August 2000.

[37]. C. E. Rasmussen and C. K. I. Williams, *“Gaussian Processes for Machine Learning”*, The MIT Press, 2006.

[38]. C. K. I. Williams and C. E. Rasmussen, *"Gaussian Processes for Regression",* D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (eds.), Advances in Neural Information Processing Systems, vol. 8, pp. 514-520, The MIT Press, Cambridge, MA, 1996.

[39]. K. V. Mardia and R. J. Marshall, "*Maximum Likelihood Estimation for Models of Residual Covariance in Spatial Regression*", Biometrika, vol. 71, no. 1, pp. 135-146, 1984.

[40]. M. E. Tipping, "*The Relevance Vector Machine*", Advances in Neural Information Processing Systems, vol. 12, pp. 652-658, Cambridge MIT Press, 2000.

[41]. Lian-Yin Zhai, Wen-Feng Lu, Ying Liu, Xiang Li, and George Vachtsevanos,"*Analysis of Time-to-Failure Data with Weibull Model in Product Life Cycle Management*" *20th CIRP International Conference on Life Cycle Engineering, Singapore, 2013.*

[42]. Pasha, G.R.; M. Shuaib Khan; Ahmed Hesham Pasha,"*Empirical Analysis of The Weibull Distribution for Failure Data, Journal of Statistics",* Vol. 13, No.1, pp.33-45.(2006).

[43]. Abernethy, R. B. "*The New Weibull Handbook*", 5th edition,(2006).

[44]. D. A. Tobon-Mejia, K. Medjaher, N. Zerhouni, "*Hidden Markov Models for Failure Diagnostic and Prognostic" ,* version 1 - 21 Jun 2011.

[45]. S. M. Ross, "*Probability Models for Computer Science*". Academic Press, 2001.

[46]. M. Dong and D. He, "*A segmental hidden semi-markov model (hsmm)-based diagnostics and prognostics framework and methodology*," Mechanical Systems and Signal Processing, vol. 21, pp. 2248–2266, 2007.

[47]. A. Dempster, N. Laird, and D. Rubin, "*Maximum likelihood from incomplete data via the EM algorithm*," Journal of the Royal Statistical Society, vol. 39, pp. 1 – 38, 1977.

[48]. J. Zarei and J. Poshtan, "*Bearing fault detection using wavelet packet transform of induction motor stator current*," Tribology International, vol. 40, no. 5, pp. 763 – 769, 2007.

[49]. Natke, H.G. and Cempel, C."*Model-Aided Diagnosis of Mechanical Systems Fundamentals, Detection, Localization, and Assessment"*, Springer-Verlag. , 1996

[50]. Natke, H.G. and Cempel, C. "*Proceedings of the Society for Machinery Failure Prevention Technology" MFPT, formerly MFPG* 1968-1997

[51]. Y.Ni, K.Y. Wong. *"Integrating Bridge Structural Health Monitoring and Condition-Based Maintenance Management"* Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong.

[52]. Kohout, J. and Vechet, S."*A new Function for Fatigue Curves Characterization and Its Multiple Merits*", Int. Journal of Fatigue, 23 (2): 175-183. 2001.

[53]. S. Engel, B. Gilmartin, K. Bongort, and A. Hess, “Prognostics, the real issues involved with predicting life remaining,” in IEEE Aerospace Conference, vol. 6, USA, 2000, pp. 457–469.

[54]. P. Ribot, Y. Pencol´e, and M. Combacau, “*Functional prognostic architecture for maintenance of distributed systems*,” in In the 7th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes, 2009, to appear.

[55]. A. Voisin, E. Levrat, P. Cocheteux, and B. Iung, “*Generic prognosis model for proactive maintenance decision support: application to preindustrial e-maintenance test bed*,” Journal of Intelligent Manufacturing, p. Accepted article, 2009.