DEVELOPMENT OF A STATE PREDICTION MODEL TO AID DECISION MAKING IN CONDITION BASED MAINTENANCE

B. HUSSIN

Ph.D. Thesis 2007
DEVELOPMENT OF A STATE PREDICTION MODEL TO AID DECISION MAKING IN CONDITION BASED MAINTENANCE

Burairah HUSSIN

Centre for Operational Research and Applied Statistics (CORAS)
University Of Salford, Salford UK

Submitted in Partial Fulfilment of the Requirements of the Degree of Doctor of Philosophy, March 2007
# TABLES OF CONTENTS

<table>
<thead>
<tr>
<th>TABLES OF CONTENTS</th>
<th>..................................................</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>..................................................</td>
<td>IV</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>..................................................</td>
<td>VI</td>
</tr>
<tr>
<td>NOTATIONS</td>
<td>..................................................</td>
<td>VIII</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>..................................................</td>
<td>IX</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>..................................................</td>
<td>X</td>
</tr>
</tbody>
</table>

## 1 CHAPTER 1: INTRODUCTION

1.1 BACKGROUND AND MOTIVATION .................................................. 1  
1.2 ORGANIZATION OF THE THESIS .............................................. 4  

## 2 CHAPTER 2: LITERATURE REVIEWS

2.1 INTRODUCTION ........................................................................ 7  
2.2 AN INTRODUCTION TO MAINTENANCE ....................................... 7  
2.3 CONDITION-BASED MAINTENANCE .......................................... 11  
2.4 COMPONENTS OF CONDITION-BASED MAINTENANCE ....................... 14  
2.5 CONDITION-MONITORING TECHNIQUES .................................... 17  
2.5.1 Vibration Monitoring ..................................................... 17  
2.5.2 Oil Analysis ..................................................................... 18  
2.5.3 Temperature Monitoring .................................................. 19  
2.6 DATA INTERPRETATION ....................................................... 19  
2.7 CBM DECISION MAKING ..................................................... 20  
2.8 MODELLING IN CONDITION-BASED MAINTENANCE ..................... 24  
2.8.1 Proportional Hazards Model (PHM) .................................. 24  
2.8.2 Proportional Intensities Model (PIM) ............................... 27  
2.8.3 Markov Models ............................................................. 28  
2.8.4 Stochastic Filtering ....................................................... 30  
2.9 COMPUTERIZED MAINTENANCE MANAGEMENT SYSTEMS .......... 34  
2.10 SUMMARY ........................................................................... 35  

## 3 CHAPTER 3: FAULT PREDICTION USING CONDITION MONITORING INFORMATION

3.1 INTRODUCTION ................................................................. 36  
3.2 BACKGROUND OF HIDDEN MARKOV MODELS .......................... 37  
3.3 MODELLING METHODOLOGY ............................................... 39  
3.4 FORMULATION OF THE TRANSITION PROBABILITIES .............. 42  
3.5 FORMULATION OF THE RELATIONSHIP BETWEEN THE OBSERVED DATA AND THE HIDDEN STATE .................................................. 44  
3.6 MODELLING DEVELOPMENT ............................................... 44  
3.7 NUMERICAL EXAMPLES .................................................... 48  
3.8 PARAMETER ESTIMATION .................................................. 50  
3.8.1 Maximum Likelihood Estimator ...................................... 51  
3.8.2 Expectation-Maximization (EM) Algorithm ....................... 55  
3.9 GOODNESS-OF-FIT TEST .................................................. 58  
3.10 SUMMARY ........................................................................ 63
CHAPTER 8: A WEAR PREDICTION MODEL BASED ON SPECTROMETRIC OIL ANALYSIS PROGRAMME USED IN DIESEL ENGINES ............................................................................................................148

8.1 INTRODUCTION ..............................................................................................................148
8.2 MODELLING DEVELOPMENT .........................................................................................149
  8.2.1 Notation ...................................................................................................................149
  8.2.2 Assumptions .............................................................................................................150
8.3 MODEL FORMULATION .................................................................................................150
8.4 MODEL APPROXIMATIONS ..........................................................................................151
  8.4.1 Approximated Grid Method ..................................................................................152
  8.4.2 Particle Filtering ......................................................................................................153
8.5 PARAMETER ESTIMATION ...........................................................................................154
8.6 NUMERICAL RESULTS .................................................................................................156
8.7 SUMMARY .....................................................................................................................156

CHAPTER 9: CONCLUSION, DISCUSSION AND FUTURE RESEARCH
.................................................................................................................................167

9.1 CONCLUSION OF THE RESEARCH ...............................................................................167
9.2 CONTRIBUTIONS OF THE RESEARCH .........................................................................168
9.3 FUTURE RESEARCH AND OTHER ISSUES ..................................................................170

REFERENCES ..................................................................................................................173

APPENDIX

Paper A ..........................................................................................................................A1-A9
LIST OF FIGURES

Figure 2-1: P-F curve ................................................................. 12
Figure 2-2: Detection of potential failures ........................................ 13
Figure 2-3: Delay time concept .................................................. 13
Figure 2-4: Hard and soft failures .............................................. 16
Figure 2-5: Illustration of pattern of the hazard function and the baseline hazard function. .......................................................... 25
Figure 2-6: Illustration of 3 states Markov model .............................. 29
Figure 3-1: Graphical representation of the structure of a hidden Markov model .............. 37
Figure 3-2: Processes of observation and hidden states .......................... 38
Figure 3-3: Two-stage failure process .......................................... 40
Figure 3-4: Vibration data for six bearings ..................................... 44
Figure 3-5: Set-up algorithm to generate simulated pattern data .................. 46
Figure 3-6: Relationship between \( y \) and \( x \) ........................................ 48
Figure 3-7: Case 1: Observed monitoring information and the probabilities of system state given \( \mathcal{Z}_t \) ................................................................. 49
Figure 3-8: Case 2: Observed monitoring information and the probabilities of system state given \( \mathcal{Z}_t \) ................................................................. 49
Figure 3-9: Simulation of five life cycles of data imitating the bearing case .......... 52
Figure 3-10: The likelihood function at each monitoring point .................... 53
Figure 3-11: Comparison between simulated, \( \tilde{y} \), and observed values, \( y \), for a particular case of simulated data ................................................................. 60
Figure 4-1: Case 1: Simulated and observed vibration levels for Gu-b3 ..................... 67
Figure 4-2: Case 2: Simulated and observed vibration levels for Gu-b4 ............... 68
Figure 4-3: Partition the pdf of \( p(y_t | \mathcal{Z}_{t-1}) \) into equal probability and transform it to uniform distribution ................................................................. 70
Figure 5-1: Trellis diagram for grid-based filter .................................. 80
Figure 5-2: Identification of a random defect of bearing Gu-b3 using SIS algorithm (optimal density) .................................................. 88
Figure 5-3: Identification of a random defect of bearing Gu-b3 using SIS algorithm (sub-optimal density) .................................................. 89
Figure 6-1: Row format for monitoring data ..................................... 94
Figure 6-2: Monitoring data after column manipulation ............................ 94
Figure 6-3: A sample of total metal concentration after the transformation ............ 96
Figure 6-4: Case 1 – Choosing the dimensions of principal component analysis for engine 830001/19 ................................................................. 98
Figure 6-5: Case 2 – Choosing the dimensions of principal component analysis for engine 830001/26 ................................................................. 98
Figure 6-6: Re-organizing condition-monitoring data from original reading .......... 100
Figure 6-7: The difference before and after re-organising .......................... 100
Figure 6-8: Examples of regular total metal concentration (1st PCA) used in diesel engine since new ................................................................. 101
Figure 6-9: Regression of total metal concentration and failure time .................. 114
Figure 6-10: 95% prediction interval from regression of failure data ................. 115
Figure 6-11: Typical monitoring observation ...................................... 120
Figure 6-12: Case 1 – pdf and actual residual time with monitoring information for engine 830001/28 ................................................................. 121
Figure 6-13: Case 2 – pdf and actual residual time with monitoring information for engine 830001/30 ................................................................. 122
Figure 6-14: pdf and actual residual time without monitoring information for engine 830001/28 ................................................................. 123
Figure 6-15: Case 1 – pdf of residual time of 830001/28 at last observation point ..... 123
Figure 6-16: Case 2 – pdf of residual time of 830001/30 at last observation point ..... 124
Figure 6-17: Partition the pdf of $p(x_i \mid \mathcal{I}_i)$ into equal probability and transform to uniform distribution ................................................................. 127
Figure 6-18: Case 1 – Expected cost per day in terms of planned replacement at time $T$ given that the current monitoring check is time $t_i$ .................................................. 130
Figure 6-19: Case 2 – Expected cost per day in terms of planned replacement at time $T$ given that the current monitoring check is time $t_i$ .................................................. 130
Figure 7-1: Case 1 – pdf and actual residual time with mixed monitoring information for engine 830001/28 ................................................................. 145
Figure 7-2: Case 2 – pdf and actual residual time with mixed monitoring information for engine 830001/30 ................................................................. 145
Figure 7-3: Case 1 – pdf residual time of engine 830001/28 at the last observation point ................................................................. 146
Figure 7-4: Case 2 – pdf residual time of engine 830001/30 at the last observation point ................................................................. 146
Figure 8-1: Algorithm for simulating a general beta distribution ......................... 160
Figure 8-2: Simulated and actual paths of monitoring information $y_i$ for 901001/1 .. 161
Figure 8-3: Simulated and actual paths of monitoring information $y_i$ for 830000/5 . 161
Figure 8-4: $p(w_i)$ based on failure data only ................................................................. 163
Figure 8-5: $p(w_i \mid \mathcal{I}_i)$ based on failure data and condition monitoring information of engine 830001/28 ................................................................. 163
LIST OF TABLES

Table 3-1: Simulation values .............................................................................................................. 48
Table 3-2: The estimated parameters and their true values ................................................................. 53
Table 3-3: The estimated parameters and true values using the likelihood function with failure information ............................................................................................................................. 54
Table 3-4: The estimated parameters and true values \( Q(\theta, \theta^j) \) ................................................. 57
Table 3-5: The estimated parameters and true values with modified \( Q(\theta, \theta^j) \) ................... 58
Table 3-6: Variances and covariances of estimated parameters ......................................................... 59
Table 3-7: \( R^2 \) for comparison between actual \( y_i \) and simulated \( \tilde{y}_j \) .......................... 61
Table 3-8: The observed and simulated values for the initial time of a random defect.. 61
Table 4-1: Estimated parameters from vibration data ................................................................. 64
Table 4-2: Variance and covariance matrix ....................................................................................... 65
Table 4-3: Case 1: \( P(x_i | \mathcal{F}_i) \) and the starting point of the abnormal stage for Gu-b3 ... 65
Table 4-4: Case 2: \( P(x_i | \mathcal{F}_i) \) and the starting point of the abnormal stage for Gu-b6.. 66
Table 4-5: Values for Pearson product moment correlation coefficient ........................................ 68
Table 4-6: Calculating the goodness-of-fit test .................................................................................. 71
Table 4-7: \( E(L_1 - t_i, x_i | L_1 > t_i, \mathcal{F}_i) \) with exponential distribution for Gu-b3 ............. 71
Table 4-8: \( E(L_1 - t_i, x_i | L_1 > t_i, \mathcal{F}_i) \) with Weibull distribution for Gu-b3 ............... 72
Table 4-9: \( E(L_2 + L_1 - t_i, x_i | L_2 + L_1 > t_i, \mathcal{F}_i) \) with exponential distribution for Gu-b3 72
Table 4-10: \( E(L_2 + L_1 - t_i, x_i | L_2 + L_1 > t_i, \mathcal{F}_i) \) with Weibull distribution for Gu-b3... 73
Table 4-11: Starting point of the abnormal stage using SPC techniques and state prediction model .......................................................................................................................................... 74
Table 5-1: Estimated parameters using the grid-based approach ................................................... 87
Table 5-2: Estimated parameters using a particle filter with prior importance density.. 88
Table 5-3: Estimated parameters using a particle filter with optimal importance density .......... 88
Table 6-1: Estimated parameters ....................................................................................................... 112
Table 6-2: Variances of estimated parameters .................................................................................. 112
Table 6-3: Estimated parameters for failure and interval-censored information ......................... 113
Table 6-4: Variances of estimated parameters for failure and interval-censored information .......... 113
Table 6-5: Estimated parameters for 95 % prediction interval from regression of failure data ............................................................................................................................................ 116
Table 6-6: Variances of estimated parameters for 95 % prediction interval from regression of failure data ............................................................................................................................................ 116
Table 6-7: Estimated parameters using interval-censored data...................................................... 119
Table 6-8: Estimated parameters using anticipated data ................................................................. 119
Table 6-9: MSE of anticipated and interval data .............................................................................. 119
Table 6-10: Variance and covariance results .................................................................................... 120
Table 6-11: Calculating the confidence level ..................................................................................... 126
Table 6-12: Calculating the goodness-of-fit ...................................................................................... 128
Table 7-1: Parameter estimation for responsive and reflective variables ......................................... 144
Table 7-2: Variance and covariance of estimated parameters ....................................................... 144
Table 8-1: Estimated parameters for beta distribution ..................................................................... 156
Table 8-2: Variances and covariance for the estimated parameters ................................................. 157
Table 8-3: Estimated parameter values from data ............................................................................. 159
Table 8-4: Null hypothesis for Pearson product-moment correlation coefficient........ 162
Table 8-5: Conditional failure probability from a beta wear model and the residual time model ................................................................. 165
# NOTATIONS

<table>
<thead>
<tr>
<th>Symbol(s)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i, j, k$</td>
<td>Integers counter</td>
</tr>
<tr>
<td>$t$</td>
<td>Monitoring Time</td>
</tr>
<tr>
<td>$X_i, Y_i$</td>
<td>Random variable at time $t_i$</td>
</tr>
<tr>
<td>$x_i$</td>
<td>A realization of $X_i$, which is an underlying state of the equipment.</td>
</tr>
<tr>
<td></td>
<td>1. Defined as a residual time for a continuous case</td>
</tr>
<tr>
<td></td>
<td>2. Defined as true states for a discrete case, use $X_i$ as the notation.</td>
</tr>
<tr>
<td>$y_i$</td>
<td>A realization of $Y_i$, the measurement obtained at time $t_i$ by condition monitoring tools.</td>
</tr>
<tr>
<td>$\mathcal{S}_i$</td>
<td>The condition monitoring history obtained up to time $t_i$ where $\mathcal{S}_i = {y_1, y_2, \ldots, y_i}$</td>
</tr>
<tr>
<td>$P(x_i</td>
<td>\mathcal{S}_i)$</td>
</tr>
<tr>
<td>$p(x_i</td>
<td>\mathcal{S}_i)$</td>
</tr>
<tr>
<td>$p(x_i</td>
<td>\mathcal{S}_i, \mathcal{N}_i)$</td>
</tr>
<tr>
<td>$\mathcal{N}_i$</td>
<td>The condition monitoring history (covariates) obtained up to time $t_i$ where $\mathcal{N}_i = {z_1, z_2, \ldots, z_i}$</td>
</tr>
<tr>
<td>$P(x_i</td>
<td>x_{i-1})$</td>
</tr>
<tr>
<td>$L_i$</td>
<td>Random variable for stage $i$ duration with $l_i$ is their realization.</td>
</tr>
<tr>
<td>$P(\bullet)$</td>
<td>Probability mass function</td>
</tr>
<tr>
<td>$p(\bullet)$</td>
<td>Probability density function</td>
</tr>
<tr>
<td>$h(t)$</td>
<td>Hazard function</td>
</tr>
<tr>
<td>$h_0(t)$</td>
<td>Baseline hazard function</td>
</tr>
<tr>
<td>$h(t, z(t))$</td>
<td>Hazard function in the presence of a covariate $z(t)$</td>
</tr>
<tr>
<td>$z(t)$</td>
<td>Vector of covariates at time $t$</td>
</tr>
<tr>
<td>$n(t)$</td>
<td>Intensity function</td>
</tr>
<tr>
<td>$n_0(t)$</td>
<td>Baseline Intensity function</td>
</tr>
<tr>
<td>$Z(t)$</td>
<td>Cumulative intensity function</td>
</tr>
<tr>
<td>Greek letters $\alpha, \beta, \lambda_1, \lambda_2, \alpha_1, \theta, \eta, \Gamma$</td>
<td>Parameters coefficient form chosen distribution (e.g. Weibull and exponent) or parameters coefficient in a model.</td>
</tr>
<tr>
<td>$C_p$</td>
<td>Cost of preventive replacement</td>
</tr>
<tr>
<td>$C_f$</td>
<td>Cost of failure replacement</td>
</tr>
</tbody>
</table>
Acknowledgments

I would like to express my gratitude to all those who have made it possible for me to complete this thesis.

My foremost thanks go to my supervisor Professor Wenbin Wang, for the patience and encouragement that carried me through difficult times, and for the detailed and constructive comments that helped to shape my research skills. Without him, this thesis would not have been possible.

I am deeply indebted to the late Professor A.H. Christer whose stimulating suggestions and encouragement helped me throughout this research.

I warmly thank other members from Centre for Operational Research and Applied Statistics (CORAS), Dr. D.F Percy, Prof. H. Kobbacy, Dr. P. Scarf, Dr. D. Jackson, Prof. R. Baker, Dr. I. Mchale and Ms. S. Sharples, for their kind supports and guidances, which have been wonderful in this study.

During this work I was able to exchange ideas with many colleagues for whom I have great regard, and I wish to extend my warmest thanks to all those who helped me with my work.

I owe my loving thanks to my wife Rozilah Kamarudin, my daughters Nurul Syafiqah and Nurul Nabihah and my son Imran Tahir. Without their encouragements and understanding it would have been impossible for me to finish this work. My special gratitude is due to my family for their devoted and loving support.

Lastly, the financial support from Universiti Teknikal Malaysia, Melaka and the Public Service Department of Malaysia is gratefully acknowledged.
Abstract

Condition monitoring and fault diagnosis for operational equipment are developing and showing their potential for enhancing the effectiveness and efficiency of maintenance management, including maintenance decision-making. In this thesis, our aim is to model the condition of equipment items subject to condition-monitoring in order to provide a quantitative measure to aid maintenance decision-making. A key ingredient towards dealing with the modelling work is to define the state or condition of the equipment with an appropriate measure and the observed condition monitoring may be a function of the state or condition of the operational equipment concerned. This leads to the two elements that are important in our modelling development; the need to develop a model that describes the system condition subject to its monitoring data and a decision model that is based upon the predicted system condition.

A quantification of the system condition in this thesis is modelled using either discrete or continuous measures. In the case of a discrete state space, this thesis presents details of how the initiation of a random defect can be identified. In the case of a continuous state space, two approaches, which were used to identify the system condition, are discussed. The first is adopted from the concept of the conditional residual time and secondly, a wear process determined from a beta distribution. In developing these models, we used vibration and oil analysis data. Note that understanding, manipulating and analysing of the data played an important role in this thesis. This is needed not only for model development, but also for validating the model. Methods for estimating model parameters are discussed in detail. In addition, since the models presented are generally beyond the scope for analytical solutions, two numerical approximation methods are proposed. Simple decision models, which minimize the expected cost per unit time over a time interval between the current monitoring time and the next monitoring time, are shown. Numerical examples to demonstrate the modelling ideas are also illustrated throughout the thesis.
1 CHAPTER 1: INTRODUCTION

1.1 Background and Motivation

Operational equipment such as pumps, conveyors, motors and others generate a large number of signals that can be monitored. As the parts of the equipment move and rotate, they produce vibration, sound and may change temperatures and pressures. In addition, the condition of the oil used as a lubricant also has a significant effect on the working condition of the equipment. These signals can act as maintenance indicators, which could be used to describe the key relationship between equipment condition and a maintenance decision. Using these condition-monitoring signals, we could assess equipment condition in its present operating environment and maintenance actions are carried out only when necessary. This could result in a safe, effective and economical maintenance operation. Furthermore, the importance of condition monitoring in maintenance has been increasingly recognized, due to the availability of modern condition monitoring technologies. With these technologies, continuous conditional indicators are provided, which can help maintenance departments to develop, measure and improve maintenance actions in the organization.

However, most of these developments focused on the technical aspects of condition monitoring (Rao, 1995; 2001; 2002; 2003), such as advanced tools and techniques in monitoring technologies, signal processing, data acquisition and interpretation. These are mainly for diagnosis on issues of what to do, but the issue of when to do it received less attention. Yet, deciding when to do it (preventive repair) also requires some justifications.

It is a common practice that a certain threshold value has to be set for a chosen condition monitoring parameter of the equipment. The threshold level may be set up based upon manufacture recommendations, personal experiences or other subjective criteria to provide a warning that a significant change has occurred and immediate action needs to be taken. From the condition-based maintenance perspective, this threshold level may not be optimal. This is due to the fact that each machine is an individual, which may behave differently even though they are supposed to be identical,
so such a common threshold level is not appropriate in most cases. Furthermore, setting up a common threshold level may pose a difficulty for maintenance actions, and would incur extra costs. For example, if we set up a lower threshold value, it may result in needing more early replacements, and could waste much of the useful remaining lifetime of the equipment. In contrast, if we set up a higher threshold value, it will result in an increase of machine failures.

Thus, from an economical or safety point of view, the basic idea is to use all the monitored data (current and past) of a particular system, and make corresponding decisions to maintain the production equipment based upon cost, safety or other criteria. There is obviously a need for an appropriate model to aid such a decision support for plant maintenance managers. It is noted however, only a very few tools are available in the market and only a small amount of research has been devoted to this area. Thus, developing a suitable model for effective decision-making in condition-based maintenance is regarded as an important addition to the subject.

To achieve our primary goal as described above, this research aims to predict the underlying state of a piece of production equipment, given its observed condition monitoring measurements at each monitoring point to date. How to define the underlying state of the equipment is a difficult issue and in this thesis, we used residual time and wear as examples. The residual time is chosen due to the fact that it represents an important characteristic in deciding an appropriate maintenance decision such as when to replace the equipment (Reinertsen, 1996). Similarly, cumulative wear could also be used, as it is a direct indication of the deterioration process. If these underlying states can be predicted, maintenance actions including manpower, equipment and tools, and spare parts can be planned and scheduled (Al-Sultan and Duffuaa, 1995).

In this study, by using available monitored-condition information, we believe that prediction of the underlying state could be much better than the conventional method, which uses only the current age to predict the remaining life or wear of a machine. It is noted that the result of the underlying state prediction can only be described by a probability distribution due to the fact that it is random and unknown. This poses the question of how we can obtain such a probability distribution of the underlying state. This is the main question addressed in this thesis. It is also highlighted that this
distribution is a key element in the subsequent maintenance decision model that we aimed to achieve.

However, there are several critical challenges, which make the establishment of the underlying state distribution a difficult problem. The first challenge is how we can define the failure, based upon the chosen underlying state and its relation to the observed monitoring data. That is, how well the condition monitoring data reflects the deterioration or failure process. Hence, understanding and modelling the processes of deterioration and failures themselves become essential in this research.

The second challenge is that we may have rich sources of condition-monitoring data but very little failure information. Also as reported by Ascher et al. (1995) the data captured from the field suffers from many problems and it makes the data manipulation task both important and challenging. Hence, most of the effort has been placed on understanding, manipulating and preparing the data for development of the model.

The third challenge is that we may have a good theoretical model, but can it be implemented in real applications? Scarf (1997) surveyed the available papers on modelling condition-based maintenance and appealed for the applicability of the models in practice. Applicability in this case implies how the complexity and computation time can be reduced. Therefore, in developing the model, a few assumptions have been made, not only to simplify our modelling but also to ensure practical advantages.

To overcome all the challenges stated above, the objectives of this research are as follows:

1. To investigate the appropriateness of the defined state used in the model to quantify the equipment condition.
2. To identify ways in which established model can be improved.
3. To explore approximation solution for the analytical model.
4. To investigate some application of the model subjected to the chosen condition monitoring data and their relations toward maintenance planning and scheduling.
The following section contains an outline of the thesis, and gives an overview of the work.

1.2 Organization of the Thesis

This thesis is organized in nine chapters. In Chapter 1, we first introduce the common problems arising in condition-based maintenance and their challenges, which motivate us to carry out this research.

In Chapter 2, a brief introduction to maintenance and management strategy in undertaking maintenance actions is presented. We introduce the concept of condition monitoring and condition-based maintenance in detail. We investigate the literature and current monitoring techniques used within industry. It is noted that the scope of the literature review is mainly concerned with the modelling aspect of decision making in condition-based maintenance, thus several approaches and concepts related to this issue are discussed.

Chapter 3 presents a new development of a conditional residual time model that is different from the literature as reviewed in Chapter 2. A discrete state space is used to define the condition of the operational equipment. The methodology and formulation used in this development are discussed, and simulation studies with numerical results are presented. In this chapter, we show how the development of this model can be used to predict the initial point of a random fault in a system. The process of model fitting and testing using an actual dataset is shown in Chapter 4. Also in Chapter 4, we had an attempt to compare our results with a statistical process control based method, which had been developed in a previous study conducted by Zhang (2004).

In Chapter 5, we investigate numerical approaches as an alternative solution to the model developed in Chapter 3. Approximate approaches, such as grid based and particle filtering, are discussed. We demonstrate these approaches using simulated and actual datasets from Chapter 4.

Chapter 6 presents further developments for predicting the conditional residual time using actual oil monitoring data. The data collected is not organised, with missing
values and many unexplained, so we have to re-organise the data into a format that is suitable for our model. One of the processes of re-organising the data is by transforming the collected data into a measure known as the total metal concentration. In general, three components of data are available to us such as the metal elements, lubricant performances and contaminant indicator while conducting the oil analysis programme. In this chapter, we only used the metal elements component indicator as it provides important information about the wear of the internal engine parts. Although there are many metal concentrations in the oil sample, not all of them are useful. Hence, a technique to reduce the dimension or size of metal elements is discussed. In addition, several procedures dealing with incomplete data are also presented. The model developed is fitted to the data and the numerical results are given. We carried out several tests to show the robustness of the model developed, and produced significant results.

In Chapter 7, the model developed in Chapter 6 is enhanced with more monitoring information. Using all three components presented in the oil-monitoring data, we grouped them into two groups. Here, we consider that these two types of condition-monitoring information are not correlated with each other but have different relationships with the residual life. The assumption and formulation for the new model are discussed, with a focus on interpreting and preparing the data required in the new model. As the lubricant performance and contaminant indicators are significantly correlated, we carry out an independent component analysis to separate each variable. The model is supplied with the actual data and the numerical results are given. Several other tests are also conducted.

As seen from Chapters 3, 4, 6, and 7, the residual time is used to characterize the failure or deterioration process of the production equipment. But, in Chapter 8, we attempt to model the deterioration process using a measure called “wear”, without using the residual time concept. To do this, we introduce a model that uses a continuous random variable to represent the process of deterioration of the system at any monitoring point. This is done using a beta distribution and allows us to have a more generic wear model that can be applied to many situations. This model is tested with the same oil data used in Chapter 6, and the numerical result is given. Furthermore, an analysis between this model and the residual time model in Chapter 6 is made.
Finally, in Chapter 9, the findings of the thesis are summarised and recommendations for future work are also discussed.
CHAPTER 2: LITERATURE REVIEWS

2.1 Introduction

This review of literature is divided into three sections, according to the nature of the problems discussed. The first section introduces the area of maintenance in general, the second section covers the concept of condition-based maintenance and its components, and the last section is concerned with the modelling aspects of condition-based maintenance, designed to aid maintenance decision-making. A fundamental understanding of the above topics is essential to the conduct of this research.

2.2 An Introduction to Maintenance

The purpose of system maintenance is to ensure the viability of the operation of equipment, as most equipment will deteriorate while in operation and with the lapse of time. A consequence of the deterioration process may be a failure of the system. Some failures are minor and result in inconvenience and small economic loss, while other failures are catastrophic, lead to uncountable cost and may be dangerous to personnel. There are many contributory factors to deterioration such as degradation, corrosion, wear, erosion, aging and production process are known as some of the root causes; and maintenance is necessary to prevent equipment from continuously deteriorating and suffering breakdowns (Reinertsen, 1996). Therefore, a well-planned maintenance scheme is important in reducing costly breakdowns (Dohi et al., 2001) while at the same time maintaining a high level of quality (Paz and Leigh, 1994) or improving the viability of the system (Murthy and Hwang, 1996).

In the past, maintenance has been labelled by managers as a ‘necessary evil’, as reported by Tsang (1995) during his study of a tool for decision-making in condition-based maintenance. Only in recent years has maintenance gradually emerged as a strategic issue in business processes (Murthy et al, 2002) and it is now seen as important in that it contributes to reducing costs, minimizing equipment downtime, improving quality, increasing productivity and providing safety and reliability (Mechefske and Wang, 2003).
Formally, maintenance has been defined by BS EN 13306:2001 as

*Combination of all technical, administrative, and managerial actions during the life cycle of an item intended to retain it in, or restore it to, a state in which it can perform the required function.*

A more general understanding of maintenance is given by Pintelon and Gelders (1992) as

*all activities necessary to restore equipment to or keep it in a specified operating condition.*

This would imply that maintenance consists of actions taken to make sure that items of equipment are fit to fulfil their required functions. It should be noted here that we should have a clear notion of the meaning of ‘required function’ or ‘specified operating condition’ as stated in the above definitions. It should be understood that equipment is usually designed with some pre-defined life expectancy or operational life. For example, equipment may be designed to operate at a full design load for such amount of age or usage. In the case of an engine, belts and hoses need adjustment, alignment must be maintained, and proper lubrication on rotating equipment is required. Such a specification of the design life assumes necessary maintenance and any failure to perform maintenance activities intended by the equipment’s designer would shorten its operating life.

Performing maintenance activities requires justification, as it contributes to cost, safety and other criteria. Jabar (2003) divides maintenance costs into two main groups namely direct and hidden costs. The former consist of items such as labour, materials, services and overheads, while hidden or indirect costs are harder to measure and are classified into six main areas of losses:

1. Breakdowns and unplanned plant shutdown losses.
2. Excessive set-up, changeovers and adjustments losses.
3. Idling and minor stoppages.
4. Running at reduced speed.
5. Start-up losses.
6. Quality defects.

An attempt should be made to ascertain how maintenance can be performed to ensure equipment reaches or exceeds its design life, taking into account economic considerations, safety or other criteria. Therefore, it is very important for companies to have a good maintenance strategy for managing the effectiveness of maintenance and maximizing equipment uptime in their organizations. A review of the maintenance objectives is summarized by Wang et al. (2004) as follows:

1. Ensuring system function (availability, efficiency and product quality).
2. Ensuring system life (asset management) and safety.

To accomplish the maintenance objectives above, three types of maintenance policies have been used such as corrective, preventive and condition-based maintenance. Corrective or breakdown maintenance has been used for many decades and is still in practice today (Luce, 1999). The basic principle of, ‘fix it when it fails’ is an approach where no preventive maintenance activities are carried out until failure occurs, when maintenance is resorted to in order to restore the required function. As a result,

1. Equipment or machines may be exposed to catastrophic failure.
2. Excessive secondary damage may occur.
3. Production downtime during excessive maintenance repair time is very costly.
4. Parts are not always readily accessible and may be expensive.
5. There is danger to the operating personnel and environment.

To minimize the impact of unexpected equipment breakdown or failure, maintenance practitioners started to adopt what was known as a planned preventive or time-based maintenance policy (Silver and Fiechter, 1992). This strategy aims at preventing unexpected failures by carrying out planned inspections or maintenance at regular intervals and undertaking replacements or repairs when necessary. The operational advantage of this pre-set schedule of inspections is that it is deterministic and relatively
simple to manage. However, the difficulty in determining the appropriate time to 
perform maintenance and the need to interrupt production at scheduled time points 
reduce the impact on saving maintenance costs, as claimed by Swanson (2001).

The requirement for increased plant productivity associated with plant automation and 
the need for improved safety and reduction in maintenance costs have led to the growth 
in popularity of condition-based maintenance (CBM), which uses various condition 
monitoring techniques to aid the planning of plant preventive maintenance and 
operational policies (Christer et al., 1997). With condition monitoring, the signals of 
system deterioration may be easy to observe, hence equipment maintenance can be 
scheduled at the appropriate time and not on an emergency basis. Therefore, the plant 
availability can be increased and the cost of unscheduled production shutdowns can be 
reduced.

In practice, however, the choice of the optimum maintenance strategy is not as simple as 
suggested above, as industry has become aware of the fact that a single maintenance 
policy, however efficient it may be, cannot eliminate all breakdowns or restore the plant 
to its full potential (Saranga, 2002). Nowlan and Heap (1978) coined the term 
Reliability Centered Maintenance (RCM) in 1978 while writing a report on aircraft 
maintenance policy for United Airlines. Since then, the term was popularly adapted to 
be used in other maintenance industrial application. Formal definition of RCM is given 
by Moubray (1997), which stated RCM as

\[
a \text{process used to determine the maintenance requirements of any physical asset in its operating context.}
\]

In short, RCM is a structured procedure for analyzing the functions and potential 
failures of physical assets in order to determine and apply the most appropriate 
maintenance policy for corresponding items and modules in a system. However, RCM 
itself cannot be used as a tool for deciding the optimal interval between two 
maintenance inspections, such as in the case of time-based maintenance, which in turn 
may improve the system availability (Saranga, 2002).
This research is limited to investigating CBM as an option in maintenance decision-making, as it shows dominance over other two strategies. Kimura (1997) offers four reasons in support of this argument:

1. When the reliability of the item does not follow the ‘bathtub’ curve, time-based maintenance loses its significance.
2. If preventive maintenance is perfectly performed, resulting in no failure, then we cannot have failure statistics.
3. Time-based maintenance is inapplicable to a new item for which failure statistics do not exist.
4. The development of a condition-monitoring tool and techniques has facilitated the implementation of CBM.

Hence, the next section will further discuss the CBM strategy and its components.

2.3 Condition-Based Maintenance

The purpose of CBM is to allow maintenance to be done only when necessary, with the help of condition monitoring data. As Kelly and Harris (1978) stated, the only aspect of CBM that distinguishes it from both run-to-failure and timed-based preventive maintenance is the sense that it requires the monitoring of some condition indicating parameters of the unit being maintained. Run-to-failure maintenance requires no monitoring activities, while time-based preventive maintenance is based on statistical failure data and some sort of simple manual inspection. Referring to BS EN 13306:2001 CBM is defined as

Preventive maintenance based on performance and/or parameter monitoring and the subsequent actions.

During operation, plant and equipment are under a variety of stresses such as load and speed. With all these stresses and the lapse of time, the condition of the item may have deteriorated so that it moves towards a failure. This is the time when the idea of condition monitoring is appropriate. According to Moubray (1997) most failures give a warning before they occur and if such a warning is identified, possible maintenance
actions can be carried out to prevent the equipment from failing completely or causing more damage to the plant. This statement illustrates the scope of condition monitoring as the techniques that can be used to identify warning signs by measuring any significant deterioration indicated by changes in the values of monitored variables. Moubray (1997) introduced the concept of P-F (Potential failure-Functional failure) curves; these are intended to explain his statement, as shown in Figure 2-1 below.

\[ \text{P - F curve} \]

P refers to the earliest time at which a potential failure can be detected and F refers to the time at which the functional failure occurs. If a potential failure or fault can be detected between points P and F or earlier, a possible maintenance action can be performed to prevent the functional failure from occurring. Moubray (1997) also mentions that there are many ways of finding out that failure is in the process of occurring. As an example, consider the P-F curve in Figure 2-2 below, which illustrates how the single failure mode could be preceded by a variety of potential failures, each of which could be detected by a different condition-monitoring technique defined by P_1, P_2, P_3, and P_4. This means that the initial defect time for any failure is not an absolute point in time, but is dependent on the method of monitoring the plant condition.
An alternative way to understand condition monitoring is to look at the delay time concept, introduced twenty years ago by Christer and Waller (1984), which is a model for industrial inspection maintenance problems. The concept has a similar definition to the P-F interval, but allows far more insight into the failure process. It defines failure as a two-stage stochastic process, where the first stage is the initiating phase of a defect, which was not defined in Moubray (1997). The second stage is where the defect leads to a failure in the absence of maintenance actions. The time lapse from the time a defect can be first identified at an inspection point, $u$, to the time that the defect causes a failure, is called the delay time, $h$. If an inspection is carried out during this (variable) time period, the defect may be identified and removed. **Figure 2-3** below illustrates the delay time concept.

**Figure 2-3: Delay time concept**

These concepts are sufficient to provide a practical view of condition monitoring and its consequences. Formally, BS EN 13306:2001 defines condition monitoring as

*Activity performed either manually or automatically, intended to observe the actual state of an item.*
Note that, if the measured condition parameters have not changed, it does not mean that
the monitoring is a waste of time. It provides a peace of mind that the equipment may
be in a satisfactory condition. Also, if the measured parameters did not show any trend
or change until failure, it is likely that a wrong type of information was collected. Thus,
the brief discussion of condition monitoring above enables us to interpret condition-
based maintenance as an outcome of the measurement of the system condition, based on
information called condition data with the aim of determining required maintenance
actions. Therefore, CBM is shown to be a method which attempts to provide a diagnosis
and prognosis approach towards maintenance problems. These analyses describe the
processes of the assessment of equipment health for present and future, based upon
observed data and available knowledge of the system (Mathur et al., 2001).

Here, diagnosis is concerned with identifying the causes of failures or anomalous
conditions in a system or its subsystems and determining the severity of given faults
once detected. Prognosis, on the other hand, is a very challenging task, which aims at
predicting failures, based on observed data and available knowledge of the system, and
may lead to recommending preventive maintenance prior to the onset of catastrophic
failures.

2.4  Components of Condition-Based Maintenance

Wang (2000) while reporting a review of the modelling of CBM decision support
claims that there are two stages of condition-based maintenance. The first relates to
condition-monitoring data acquisition and its interpretation, followed by the second
stage of making a decision based on the monitored information. Similar arguments also
can be found in Jardine et al. (2005) suggesting that a CBM programme should consist
of three key steps:

1. Data acquisition, to obtain data relevant to system health.
2. Signal processing, to handle the data or signals collected in step 1 for better
   understanding and interpretation of the data.
3. Maintenance decision-making, to recommend efficient maintenance policies.
Hence, we may conclude that the theory and implementation of condition-based maintenance must have these components. The first component, data acquisition, plays an important role in this approach, in which the condition of the equipment needs to be known. In general, condition monitoring can be divided into two categories (Kelly and Harris, 1978), namely monitoring which can be carried out without interruption of production, and monitoring which requires the shutdown of the unit. This categorization is important in order to select the appropriate techniques and tools for data acquisition (Shearman, 2001). Attempting to fulfil this requirement, various data acquisition techniques and tools can be utilized within condition monitoring (Williams et al., 1995). Vingerhoeds et al. (1995) discussed additional kinds of condition monitoring, i.e. off-line condition monitoring and on-line condition monitoring. The diagnostic system for on-line monitoring enables quick and reliable fault diagnosis for the warning that occurs. However, many challenges still exist in practical applications of on-line monitoring, such as rapid data processing, diagnosis procedure and the high operating cost. To overcome such problems, off-line monitoring approaches have been used, in which the data is measured on-line but the analysis is carried out on a regular basis (off-line).

However, a key difficulty in condition monitoring is to detect changes that are not necessarily directly observed and that are indirectly measured together with other types of noise. In a study of the condition monitoring of a component which has an observable measure of condition called ‘wear’, Christer and Wang (1995) classify the information collected during the monitoring exercise into two categories, namely direct and indirect information. They define the direct information as the measurement of a variable that can directly determine the state of the system, for example the thickness of a brake pad or the depth of a crack. Normally, these direct methods may be based on visual inspections or other types of monitoring such as non-destructive sensing which may not be economical to use for some equipment (Jantunen, 2006). Indirect information is defined as the associated information that is influenced by the component condition, which cannot be directly observed; for example, the information gathered in oil analysis or vibration frequency analysis.
Another categorization is found in the work of Martin (1994) who distinguishes between two different types of fault, namely soft and hard faults (see Figure 2-4). The difference between these types is important, as soft faults lead to predictable situations, hence being amenable to condition monitoring, while hard faults are basically unpredictable; but there is a view that even hard faults must exhibit some changes before the occurrence of the failure (Martin, 1994).

![Figure 2-4: Hard and soft failures](image)

One example of such a categorization is that a large amount of recent industrial research has put more effort into systems consisting of mechanical plant and equipment such as power turbines, diesel engines and other rotating machinery, rather than electronic or electrical systems. The reason for this is that failure in mechanical systems tends to occur slowly, so that if condition monitoring is performed, it will provide an opportunity to assess the deterioration and to compute the expected remaining life of a system or machine, while in electronic systems, failures tend to occur without any warning or delay time.

Thus, the categorization of condition monitoring helps us to understand more about the process of deterioration, as developing prognosis models can be varied according to the type of observed condition-monitoring data in conjunction with the defect time for the potential failure, which is not a deterministic point (Moubray, 1997; Christer and Wang, 1995). Given that the initial defect time for any potential failure is not certain, it provides us with a broad view of the state of the art in fault-diagnosis techniques. Moubray (1997) has chosen the techniques to detect these potential failures by monitoring measurable parameters or faults as ‘equipment condition monitoring’.
Others (Andersen and Rasmussen, 1999) have referred to it as ‘information about technical health’. It has been reported that major improvements have occurred in the technology, practice and use of equipment condition monitoring over the past sixty years (Mitchell, 1999). An example is the development from the mechanical instruments that were used 20 years ago to capture a simple low frequency dynamic waveform to today’s high-performance digital instrumentation. Methods of equipment condition monitoring can be classified according to the monitored parameters that were influenced by the potential failure (Moubray, 1997). To support his argument, Moubray divides condition-monitoring techniques into six categories:

1. Dynamic effects, such as vibration and noise levels.
2. Particles released into the environment.
3. Chemicals released into the environment.
4. Physical effects, such as cracks, fractures, wear and deformation.
5. Temperature rise in the equipment.
6. Electrical effects, such as resistance, conductivity, dielectric strength, etc.

However, irrespective of the condition monitoring techniques used, the key elements of condition monitoring are the same: the condition data that becomes available needs to be converted into a meaningful form and appropriate actions must be taken accordingly. As examples in this discussion, a few condition monitoring techniques that are popular in industry have been selected from the survey conducted by Higgs et al. (2004). For other references to such methods, see (Moubray, 1997) and (Williams et al., 1995).

2.5 Condition-monitoring techniques

2.5.1 Vibration Monitoring

The vibration-monitoring technique is perhaps the most popular and widely used of all monitoring techniques (Higgs et al., 2004). It can be used to detect several system conditions, such as fatigue, imbalance, misalignment, loosened assemblies and turbulence, which can occur in rotational or reciprocating parts such as bearings, gearboxes, shafts, pumps, motors, engines and turbines. The operating processes of these components will release energy in the form of vibration, whose amplitude will remain in a steady state unless there is a change in the operating dynamic of the system.
The changes at this stage may signal a warning of the impending failures that may occur. Reeves (1998) explains that vibration monitoring consists of identifying two quantities: the magnitude of the vibrations and their frequency. The former is used to establish the severity of the vibration, while the latter indicates the origin of the defect. It should also be noted that the severity of the vibration in any particular case would depend on these factors:

1. Type of machine
2. Flexibility of the mounting/foundation
3. Position or direction of measurement
4. Operating conditions during measurement

A summary of techniques used in vibration monitoring is given by Reeves (1998).

2.5.2 Oil Analysis

Condition monitoring through oil analysis provides a means of analysing oil at regular intervals to determine if it still meets the lubrication requirements of the equipment. The particles contained in a lubricating fluid carry detailed and important information about the condition of the machine components. There are many methods, which can be utilised to perform an analysis to obtain such information. A comparison between different methods, which were available, is given by Roylance (2005). The features of the analysis can be deduced from particle shape, size, composition, size distribution and concentration, which can be classified into three categories namely quantitative, qualitative and material properties, see Khan and Starr (2006) for details of the classifications.

A change in the rate of particles collected indicates a change in the condition of the machine. When this condition reaches an unacceptable state, the machine must be replaced to maintain satisfactory system operations. The two most commonly used methods in oil analysis are spectrometric and ferrographic analysis. Spectrometric analysis is the method commonly used to detect small wear particles. It is also used to identify the possible introduction of contaminants. The results are typically reported in parts per million (PPM). According to Edwards et al. (1998), it is important to note that
this method monitors only the smaller particles present in the oil (<10 microns). This disadvantage of spectrometric analysis is due to the fact that large and medium particles (> 10 microns) are likely to exit the oil flow via some filtration. This leaves the small particles which passed through the filter to remain suspended within the engine and their oil measurements to provide an indication of machine condition (Edwards et al., 1998).

Ferrographic analysis produces similar results to spectrometric analysis, but with two main exceptions. First, ferrographic analysis separates wear particles by using a magnetic field, rather than burning a sample as in spectrographic analysis. Secondly, wear particles that are larger than 10 microns can be separated and analysed, which provides a better representation of the wear particles in used oil analysis. The only criticism of this technique is that the analysis of the wear particles is very skill-dependent, subjective and time-consuming as well (Whitlock, 1997) and (Roylance, 2005).

2.5.3 Temperature Monitoring

Thermography has become a common technique for non-destructive inspections in various engineering fields which is mention by Lo and Choi (2004). It allows the monitoring of temperatures and thermal patterns to be conducted while the equipment is in operation. Generally, all mechanical systems generate thermal energy during normal operation, which permits infrared thermograph instruments such as infrared detectors to evaluate their operating conditions. Thermal anomalies, where components are colder or hotter than they should be, are taken as alarm signals of potential problems within the system. However, Lo and Choi (2004) conclude that these results can also be affected by the experience of the assessor, equipment capabilities, construction details and environmental factors.

2.6 Data Interpretation

The second component of CBM is how the collected data is interpreted; in other words, how the analysis and diagnosis are carried out to reveal operational patterns, any faults and subsequent consequences of fault conditions. This relates to the concept of signal
processing. The purposes of signal processing in diagnostic and prognostic applications of CBM are given by Bengtsson et al. (2004) as:

1. To remove distortions and restore the signal to its original shape.
2. To remove sensor data that is not relevant for diagnostics or predictions.
3. To transform the signal to make more explicit any relevant features, which may be hidden in the signal.

This also correlates with the work of Mathur et al. (2001) who noted that the real issues in signal processing are features of selection/reduction and how to treat any missing data. In addition, they noted that two characteristics that are common to all applications are incomplete or imprecise knowledge about the system of interest, especially in the failure space, and the uncertainty of the observed data. In certain cases, signal processing may also manipulate the signal so that some characteristics become more visible, enabling better prognosis. An example of the data cleaning procedure is introduced by Jardine et al. (2001) in monitoring the condition of mine haul truck wheel motors. Another example is given by Christer and Wang (1992), who assumed that condition monitoring is capable of indicating wear trends, which can then be used to determine the critical warning level and frequency of monitoring inspections.

2.7 CBM Decision Making

The final component of CBM is decision-making, as it is responsible for providing a justification for maintenance actions rather than arbitrary actions. A study performed by Tsang (1995) has shown three general types of condition monitoring decision that need to be made; nevertheless there can be many decisions related to CBM. The general types are selecting the parameters to be monitored, establishing the warning limit and determining the monitoring interval. The first and second decisions are problems that relate to a wide range of engineering issues that are taken into consideration before the maintenance decision policy can be established and which can be found in the series of proceedings of COMADEM (Rao, 1995; Rao, 2001; Rao, 2002; Rao, 2003). Other engineering issues relate to diagnosis and investigating faults (Pennacchi and Varia, 2003; Jennings and Drake, 1997), designing and implementing monitoring systems (Leem and Dornfield, 1996; Pritchard et al., 2003), condition-monitoring data diagnosis
(Prabhakaran and Jagga, 1999) and implementation of computerized condition monitoring (O'Sullivan, 1991; Bekiaris and Amditis, 2003). Recent developments in artificial intelligence such as neural networks (Timusk and Mechefsk, 2001), expert systems, fuzzy logic (Liu et al., 1996) and genetic algorithms also contribute to the techniques of fault diagnosis.

The third problem mentioned above (Tsang, 1995) relates to decisions on how and when to schedule the condition inspections in an effective way. Decision-making in this perspective is concerned with the selection of an economic inspection schedule to balance the cost and downtime due to breakdowns and inspections. Grall et al. (2002) reporting on the CBM policy for a stochastic deterioration system, state that the choices made in determining the monitoring intervals and establishing the critical threshold values will influence the economic performance of the maintenance policy, so that, in practice, a conservative approach could lead to the setting of a threshold and inspections being conducted more often than necessary, leading to non-optimal maintenance policies. As an example, Wang (2000) presented a model dealing with the selection of the best condition monitoring interval and the optimal critical level in terms of a criterion of interest, which can be cost, downtime or reliability.

Now, let us consider the action to be taken at the time of each inspection or monitoring; at this point, two decisions will be made: first, what maintenance action to take, either to replace or repair the system to a specific state or to leave as it is; and secondly, when the next inspection will take place (Lin and Wiseman, 2005). Again, the decision at this stage can be complicated and entails considerations of cost, downtime, production demand, preventive maintenance shutdown windows and most importantly, the likely survival time of the item monitored. Implementation of condition monitoring costs money, and therefore the purpose of condition-monitoring strategy development is to ensure the optimum return by maximizing the effectiveness of the monitoring data.

Jardine et al. (2005) state that most companies are very enthusiastic in their general commitment to CBM, but often confined themselves to the data acquisition step, leaving the maintenance decision to follow only certain warning limits from their experience-based knowledge. As a result, end-users tend to schedule machinery maintenance based on subjective alarm criteria, rather than on objective estimates of the
remaining life. Hence, an unknown amount of remaining life is wasted and this does not afford the end-user any means of economically evaluating condition-monitoring data. A good example of this is a study of the oil-based condition monitoring of locomotive gearboxes used by the Canadian Pacific Railway (Aghjagan, 1989), which indicated that the failure rate of gearboxes while in use fell by 90% after condition monitoring was commissioned; however, further details of reconditioning/overhaul showed that there was nothing evidently wrong in 50% of the cases. Obviously, this is not an efficient way to perform CBM. An effective and efficient CBM programme must include not only advanced condition-monitoring equipment for data acquisition but also advanced technologies for signal processing and maintenance decision making.

This can lead to a challenge regarding how to build up an appropriate formulation for learning and inference concerning problems of maintenance decision making using condition-monitoring data. Lofsten (2000) defines maintenance modelling as a mathematical model used to represent the behaviour of the plant under different maintenance actions and thereby to identify the ‘best’ maintenance decisions in terms of cost, downtime, output, availability or any criterion of interest. Further details of model development are discussed by Dekker (1995), according to whom maintenance optimisation models should address four aspects:

1. Description of a technical system, its function and its important features.
2. Modelling of the deterioration of the system over time, and possible consequences for the system.
3. Description of the available information about the system and the actions open to management.
4. An objective function and an optimisation technique, which help in finding the best balance.

However, because at the early stages decision making was not emphasized, or justified via quantitative rather than qualitative measurement, little attention has been paid to modelling the appropriate decision making in CBM (Wang, 2000). A related argument has also been found in Mathur et al. (2001), who observed that such efforts are still in their infancy and therefore the results are not easily available in the public domain. The
literature shows that modelling CBM is under-explored and the few related studies available are small in number, having been conducted by the same groups of authors.

Only a few of the papers, which have been assessed, are based on the types of monitoring data. Christer and Wang (1995) developed a simple monitoring model based on direct condition information. Hontelez et al. (1996) developed an optimum CBM policy for deteriorating systems based on a Markov decision process. Aven (1996) developed a condition-based replacement policy using a counting process approach. A model based on indirect monitoring information has been presented by Wang et al. (1997) and Christer et al. (1997). Wang and Christer (2000) used the filtering approach in modelling the residual time distribution subject to condition monitoring. Love and Guo (1991), Kumar (1996) and Jardine et al. (1998) used a ‘proportional hazards’ model as an attempt at modelling CBM.

A study by Wang et al. (1996) on stochastic decision modelling of CBM shows how a relationships between the condition of the system and monitored information can be established. Wang et al. (1996) using a regression model to predict the future condition information and an accelerated life model to predict the future development of the true condition of the system. Based upon the accelerated life model, a decision model can be established to determine the decision variable of interest, such as the optimal replacement time. Later, a study done by the same author Wang (2000) used a random coefficient growth model to describe the deterioration process of an item monitored to explore the relationship between the critical level, monitoring intervals and the objective function of interest. Scarf (1997) surveys the available papers on modelling CBM and appeals for the applicability of the models in practice. He finds that too much attention was being paid to the invention of new models without practical implementation. In particular, this criticism recognizes that since CBM is based on monitored data, the development of a modelling approach should consider the relationship between the system condition and the monitored data. The next section will discuss this matter in detail.
2.8 Modelling in Condition-Based Maintenance

The objective in developing a CBM model is to acquire knowledge about the equipment condition based upon its monitored parameters, so that any necessary maintenance decisions can be made accordingly. Generally, models can be divided into two parts: the deterioration model and a decision model (Frangopol et al., 2004). The deterioration model is used to approximate and predict the actual deterioration process, taking into consideration age and condition-monitoring information. The decision model uses the deterioration model to determine the optimal decision that we would like to decide in order to minimise a criterion of interest, such as cost, safety and others. Since the current and future states of a system are unknown unless it is directly observed, a probabilistic approach is suitable for the model.

In this section, a general idea is given on how this uncertainty is modelled as a key element of CBM optimisation. Several approaches are being researched and applied in the development of CBM models. One of the methods that have been widely used is Proportional Hazards Modelling (PHM).

2.8.1 Proportional Hazards Model (PHM)

Introduced by Cox (1972), the Proportional Hazards (PH) model is used to identify significant covariates and to quantify their effects on survival (using hazard function) as a function of covariate values and the working age (time). The model has been widely used in the biomedical field, for which medical research and drug trials are examples. Recently there has been an increasing interest in its application for reliability engineering. The use of PH models in reliability assumes that the hazard function of a machine is the product of both a baseline hazard and a term containing explanatory variables or covariates at time \( t \) as illustrated by Kumar (1996) in Figure 2-5.
Figure 2-5: Illustration of pattern of the hazard function and the baseline hazard function.

Kumar (1996) stated that effect of a covariate is to increase or to decrease the hazard function. Mathematically, the hazard function can be written as

\[ h(t, z(t)) = h_0(t)\phi(\gamma, z(t)) \]  \hspace{1cm} (2-1)

where \( h_0(t) \) is the baseline hazard function, \( \phi(\gamma, z(t)) \) is the functional covariate term, \( z(t) \) is a vector of covariates and \( \gamma \) is a vector of covariate coefficients. These covariates may be measurements of machine condition, such as the levels of metallic components in oil analysis or vibration amplitude. \( \gamma \) may be estimated from the data to provide a quantitative measure of the importance of each covariate and their impact on the hazard.

To use PHM, a stopping rule (the interval for maintenance) needs to be defined. In Jardine et al. (1997) and Jardine et al. (1998), the interval is defined as \( T_d \), where \( d > 0 \). To find the optimum value of \( d \), the cost model (Jardine et al., 1998; Kobbacy et al., 1997) is used, and shown below

\[ C(d) = \frac{C_p(1 - Q(d)) + C_fQ(d)}{W(d)} \]  \hspace{1cm} (2-2)
where
\[ C_p \] is the mean cost of a preventive replacement.

\[ C_f \] is the mean cost of a failure replacement.

\[ Q(d) \] is the probability that an item will fail before a preventive replacement.

\[ W(d) \] is the expected time between two consecutive replacements (regardless of reason for replacement – preventive or at failure).

The details of this formulation can be found in Jardine et al. (1998). The two most common techniques for estimating the PHM coefficients are (i) Cox’s partial likelihood method, which estimates the coefficients without making any assumptions about the form of the base hazard, and (ii) the maximum likelihood estimation (MLE) method, in which an explicit form of base hazard (e.g. Weibull hazard) is assumed (Gurvitz, 2005).

Higher values of \[ \phi \] will reduce the estimated survival time for a machine, so it can be viewed as a factor accelerating the failure rate of the machine (Mann et al., 1995).

Jardine et al. (1998) consider the baseline hazard \[ h_o(t) \] as a Weibull hazard function, which has the following form

\[
h_o(t) = \frac{\beta}{\eta} \left( \frac{t}{\eta} \right)^{\beta-1}
\] (2-3)

where \( \eta \) and \( \beta \) are the scale and shape parameters. Once the optimal threshold level \( d^* \) is calculated, it is easy to utilize the optimal replacement rule where the replacement is taking place at the first time \( t \) for which

\[
\gamma_1 z_1(t) + \gamma_2 z_2(t) + \cdots \geq \ln \frac{\eta^\beta d^*}{\beta} - (\beta - 1) \ln t
\] (2-4)

This implies that PHM combines all the significant measurements into one single value with appropriate weights. The decision rule then suggests preventive replacement at the time \( t \) when the combined covariate values reach the warning level, which depends on \( t \). Several authors have considered proportional hazards models for modelling the
condition monitoring process (Jardine et al., 2001; Love and Guo, 1991; Vlok et al., 2002). A list of applications of PHM in reliability studies can be found in Kumar (1996). The advantage of this method is that it includes both the age and the condition of the equipment in the calculation of the hazard at time \( t \) (Jardine et al., 1998). Needless to say, PHM assumes that the covariates will change the hazards that may be true in some cases, but in reality it may be the state of the system that causes the change in the observed parameters and not vice versa, such as in the case of vibration monitoring. This would be a significant problem if PHM were used in CBM, where the relationship between the monitored parameters and the underlying state of the system does not follow the assumptions made in PHM.

### 2.8.2 Proportional Intensities Model (PIM)

Another approach used in modelling CBM is the proportional intensities model (PIM), in which the proneness of the unit to failure can be characterized by the failure intensity. A typical example of this type of problem is a repairable system, in which, when the failure occurs, a small part of the system can be repaired without replacing the entire system. The system may be repaired several times before being finally replaced; therefore a system that is considered repairable cannot be modelled by the conventional hazard function, as successive failures are not identically distributed and not independent. Instead, it can be modelled using a counting process in which the hazard function with time-dependent covariates can be replaced with the intensity function, which may be defined in terms of the number of failures, \( N(t) \) over time \( (0,t) \). In other words, this describes the failure intensity. Wang (2000) defines the intensity function with a baseline intensity multiplied by a multiplicative factor involving covariates, which can be written as

\[
n(t) = n_0(t) \exp(y(t))
\]  

(2-5)

where \( n_0(t) \) is the baseline intensity and \( y(t) \) is a vector of covariates. A non-homogeneous Poisson process (NHPP) forms the basis of this model. Common models used to describe the baseline intensity of an NHPP include the power law model, \( n_0(t) = \alpha t^{\beta-1} \) and the exponential model, \( n_0(t) = \alpha e^{\beta t} \). If \( \{N(t), t > 0\} \) is an NHPP
having intensity function \( n(t) \), then \( N(t) \) is a Poisson random variable having mean \( Z(t) \), where \( Z(t) \) is the cumulative intensity function.

\[
Z(t) = \int_0^t n(u)du
\]  

(2-6)

Consider that the system has an increasing intensity function, which tells us that failures will be likely to occur more frequently and at some time \( t \) it will become more economical to replace the system. To find the optimum value of \( t \), \( (t^*) \), we minimize the cost model,

\[
E[C(t)] = \frac{1}{t} \left[ C_p + C_f Z(t) \right]
\]  

(2-7)

where

- \( C_p \) is the cost of replacing/preventive the system.
- \( C_f \) is the cost of a failure.

Such applications can be found in Aven (1996), Ascher et al. (1995), Percy et al. (1998) and Watson et al. (2002). PIM is just an extension of PHM (Kumar, 1996), which also assumes that the covariates will influence the intensity, hence it experiences the same shortcomings as those of PHM discussed above.

### 2.8.3 Markov Models

Another approach that is widely used for modelling CBM uses Markov models. In all cases of condition monitoring, the condition of a technical unit is changing continuously. To use a Markov model, the deterioration process is normally assumed to be described in terms of a limited number of condition states, such as ‘normal’, ‘defective’ and ‘failure’. Figure 2-6 illustrated the concept of Markov model with a transition probabilities link the current state with a maintenance action to a future state.
A transition probability \( \{P_{12}, P_{21}, \ldots, P_{13}\} \) is defined as the probability that a component will move from one state to another, depending on the action taken. The Markov property assumes that the probability of deteriorating to another state does not depend on the history of the process, which implies that the future lifetime of a component/system does not depend on how long it has already operated. Applications of this approach in CBM can be found in Coolen and Dekker (1995) and Hontelez et al. (1996). However, the ‘memoryless’ property of this approach, which considers the future condition as dependent solely on the current information, without using the whole history of observation, limits the capability of this technique.

As an extension of the Markov model, a general model of the deterioration process pioneered by Thorstensen and Rasmussen (1999), can also be used to estimate the deterioration function of an asset and then incorporate a stochastic component to represent the failure of the system. Thorstensen and Rasmussen (1999) present the deterioration process at time \( t \), \( \xi^t \) as

\[
\xi^t = g(t) + b\sqrt{t}U \quad U \sim N(0,1)
\]

where \( g(t) \) represents the deterministic deterioration function of an asset and \( b\sqrt{t}U \) represents the stochastic nature of the deterioration. The deterioration function is then divided into \( M \) condition levels in which condition level 1 represents the new condition of the component, while condition level \( M \) represents a total breakdown. Here, the concept of the Markov model is used and the boundary of each condition level is introduced. This model looks interesting, but as future deterioration is independent of the deterioration in current time units, it has similar drawbacks to the original Markov model. It is noted that the Markov model assumes that the observed information

![Figure 2-6: Illustration of 3 states Markov model](image)
completely describes the system state while in most CBM cases this assumption does not hold.

### 2.8.4 Stochastic Filtering

Another convenient way of modelling a deterioration process in CBM is by regarding it as a stochastic process with hidden states. Consider the condition of a monitored system deteriorating according to a stochastic process $x_i$, which is not directly observed, where $x_i$ may be the true condition of the system at time $t_i$ since new. At each monitoring point, there exists observable $y_i$, the monitored information that is assumed to be correlated with $x_i$. Once the relationship between the true condition of the system and the monitored information is established, a decision model can be developed to recommend the appropriate maintenance action. One serious challenge for the modeller is how to explain this relationship. Christer et al. (1997) use a state space model for the unobserved $x_i$, the slot size in a furnace given the observed condition information history, $\mathcal{X}_i = \{y_1, y_2, ..., y_t\}$ which is the history of the conductance ratio at a particular time $t_i$, $i = 1, 2, ...$. They established a relationship that can be used to predict the probable slot size, $x_i$, in the furnace from the information given by the conductance ratio, $y_i$. This relationship is written as

$$
x_i = \beta x_{i-1} + \eta_i
$$

$$
y_i = \alpha x_i + \varepsilon_i
$$

(2.9)

where $\varepsilon_i$ and $\eta_i$ are Gaussian noise (disturbance factors) and $\alpha$ and $\beta$ are parameters of the state space model ($\alpha$ and $\beta$ can be vectors). Thus, they seek to model the distribution of the state, conditional on $\mathcal{X}_i$, which they denote by $(x_i \mid \mathcal{X}_i)$. The two-step Kalman filtering approach was used to solve the equation. In their study, $(x_i \mid \mathcal{X}_i)$ is assumed to follow a normal distribution. It is noted, however, that the linear and Gaussian assumptions restrict the application of the Kalman filtering approach.

Irrespective of the method used to monitor system behaviour, one important variable of operational interest is how long it can survive given the condition information history.
(Wang and Christer, 2000). Hence, the concept of conditional residual time (CRT) is introduced to represent the system state based on the condition-monitoring information. CRT is defined by Wang and Christer (2000) as

\[
\text{the time lapse from the time point a potential failure is identified, based upon measured condition information, to the time that it would fail given no other maintenance action.}
\]

Once the relationship between the CRT and the measured condition is established, a decision model can be developed to recommend the appropriate maintenance actions. As an example, Wang and Christer (2000) develop a residual time-predicting model using non-Gaussian and non-linear filtering. The relationships between the CRT and the measured condition are described in general by the following recursive formula:

\[
x_i = \begin{cases} 
      x_{i-1} - (t_i - t_{i-1}) & \text{if } x_{i-1} > t_i - t_{i-1} \\
      \text{not defined} & \text{otherwise}
   \end{cases} \quad (2-10)
\]

and

\[
y_i = g(x_i, \eta_i) \quad (2-11)
\]

where \( x_i \) is defined as the CRT at time \( t_i \), \( g(x_i, \eta_i) \) is a function to be determined, \( t_i \) for \( i = 1,2,\ldots \) are the condition monitoring check time points and \( \mathcal{I}_i = \{y_1, y_2, \ldots, y_i\} \). Applying the multiplication law, \( p(x_i \mid \mathcal{I}_i) \) can be written as

\[
p(x_i \mid \mathcal{I}_i) = \frac{p(x_i, y_i \mid \mathcal{I}_{i-1})}{p(y_i \mid \mathcal{I}_{i-1})} \quad (2-12)
\]

Hence, it is shown that a relationship between \( x_i \) and \( y_i \) can be established by recursive filtering. The dataset from the furnace case described in Christer et al. (1997) and referred to above, is again used to provide the numerical example. It is also shown by Wang and Christer (2000) that the effectiveness of maintenance action can be modelled, given that an additional function is required to describe the relationship between the residual time before and after the maintenance actions. Wang (2002) applied the above
technique to a dataset from six rolling element bearings as an application of the model developed to determine the bearing residual time, given vibration monitoring information, and produced acceptable results. Soon after, Wang and Zhang (2002) applied the same technique, but with different condition information, namely oil analysis data, to demonstrate a model to predict the residual life of equipment. Wang (2003) uses the filtering approach to model optimal condition monitoring intervals by analysing five different criterion functions and comparing those using simulations.

However, the method of finding the distribution of conditional residual time as a state of the plant could have drawbacks in practice (Wang and Zhang, 2002). These drawbacks occur as a result of how we defined the state of the plant in a meaningful and usable way. First, it is shown from previous work (Wang and Christer, 2000; Wang, 2002) that conditional residual time is used to measure the system state, which implies how long the equipment could survive before failure. We notice that the residual life is only a measure used to describe the system deterioration but other measures could also be used. As an example, people prefer the plant state to be characterized in a simple and meaningful way, such as ‘good’, ‘bad’ or ‘about to fail’, and the residual time cannot provide a direct answer to this. Secondly, the residual time is assumed to be fixed at birth (installation), see equation (2-10), but is unknown. The latter problem is solved by Wang and Christer (2000) allowing $x_i$ to change, but the former still remains.

This motivates us to carry out an attempt to use a discrete space to describe the state of the system, which is described in detail in Chapter 3. In order to use the concept of conditional residual life to explain the relationship between $x_i$ and $y_i$, e.g. equation (2-11), an initial set-up must be identified. However, as in each case different condition-monitoring data is used, a different initial set-up will also apply. We will return to this point in subsequent chapters.

Last but not least, the deterioration of the system can also be modelled by age and condition monitoring variables using a gamma stochastic process. It is assumed that when the component/system is new, the deterioration is equal to 0 and a component/system is said to fail when the deterioration exceeds a certain threshold value. Examples of this application, to model the deterioration processes and use in
maintenance decision making are Noortwijk and Klatter (1999) who model the scour-hole development in block mats, and Grall et al. (2002) who consider a single-unit system or structure subject to a continuous accumulation of wear in time to obtain optimal thresholds and inspection scheduling. An advantage of modelling deterioration processes through a gamma process is that the required mathematical calculations are relatively straightforward (Pandey et al., 2005). However, the set-up for an appropriate benchmark failure level is a non-trivial course of action and it needs to be done carefully. A recent attempt conducted by Wang (2006) using a beta model to describe the deterioration process seems partially to resolve this issue.

Detecting the initiation point of the second stage, \( u \), while referring to a two-stage failure process (Christer and Waller 1984) is also an important issue in CBM modelling. In most cases, the detection of an initial defect is made either statistically or by expert judgment (Wang, 2002; Zhang, 2004). The basic principle of both techniques uses the captured information or experienced observations to create a reference for a normal condition and then measures the variation of the present monitoring data from a normal condition to see whether the present signal is within the control limit or not (Lin, 2006). Predicting the initiation of the second stage, \( u \), is one of the proactive approaches on handling faults. In the case of condition monitoring using vibration data, it is preferable to develop a combined model to predict when the fault will occur and then to link it to a subsequent CRT model to determine the residual life.

It is clear here that fault prediction is very different from fault detection. The latter deals with faults that have already happened, while the former deals with faults that may happen in the future. By considering the condition-monitoring information, Chapter 3 will report a modelling development for fault identification, where \( x_i \) is not defined as the residual time but as a discrete set of finite numbers representing ‘normal’, ‘defective’ or ‘failed’. With this concept, we show how the initial defect can be found using the filtering approach itself and a residual time can still be formulated indirectly from there.
2.9 Computerized Maintenance Management Systems

The complexity of manufacturing systems and products, as well as the growth in computer technology, has opened new possibilities in the area of computerized maintenance management systems (CMMS). According to Cooke (2003) the main motivation behind designing a CMMS is the need to facilitate the management of maintenance resources, to monitor maintenance efficiency and to provide appropriately analysed management information for further consideration. Indeed, (Labib, 2004) has suggested some benefits that CMMS can offer, as follows:

1. It can support condition-based monitoring (CBM) of machines and assets, to offer insight into wear and imminent failure.
2. It can track the movement of spare parts and requisition replacements when necessary.
3. It allows operators to report faults faster, thus enabling maintenance staff to respond to problems more quickly.
4. It can facilitate improvements in the communication between operation and maintenance personnel, and is influential in ameliorating the consistency of information passed between these two departments.
5. It provides maintenance planners with the historical information necessary for developing PM schedules.

In spite of providing significant characteristics of CMMS that may fit with the needs of industry, Labib (2004) also finds that the majority of commercially available CMMS software applications lack decision supports for management. One of the reasons highlighted by Labib (2004) is that managers are unaware of the various types of maintenance optimisation models. The reason for this lack of awareness is simply that none or little assessment of the successful applications of maintenance optimisation models has been reported (Tsang et al., 2006). This is because optimisation models have

1. computational difficulties.
2. difficulties of collection of data and modelling of failure distribution.
3. the gap between theory and practice.
As far as this research is concerned, few packages have been identified that involve optimisation models in condition-based maintenance decision support. One good example, is EXAKT (The CBM Optimizer) (Tsang et al., 2006). This CBM optimisation model was embedded in software developed by the CBM laboratory at the University of Toronto. It is designed to use condition-monitoring data, along with age-based records and financial estimates, to indicate maintenance actions based upon proportional hazards and Markov approaches.

As mentioned in the discussion above, maintenance modelling, especially using condition-monitoring data, is under-explored. Few approaches have been applied and could be applied to obtain an optimised CBM. This is the motivation for this research, which aims to give more choices to the system developer for plugging into their CMMS software.

2.10 Summary

In this chapter, an introduction to the general maintenance concept is presented, while the discussion of condition-based maintenance is highlighted. During the rationalization of CBM, its application can be categorized according to different types of monitoring data, which share the same principle, namely undertaking maintenance actions based on observed anomalies in monitoring data. However, little effort has been paid to model maintenance decision-making based upon its condition data, compared with the extensive research on the engineering side of CBM. Here we reviewed and discussed recent developments in condition-based maintenance modelling, such as the proportional hazards model, proportional intensities model, Markov models, stochastic filtering model and conditional residual time model to aid maintenance decision making. Lastly, a review of current CMMS packages has revealed a lack of optimisation maintenance decision models.
3 CHAPTER 3: FAULT PREDICTION USING CONDITION MONITORING INFORMATION

3.1 Introduction

One of the key advantages of condition monitoring in a maintenance strategy is its ability to identify faults, and this has attracted substantial attention. However, users would not only like to identify faults in processes, but also to predict when the fault will occur, as far in advance as possible. One of the few previous attempts to solve these problems is the condition residual time (CRT) model (see Chapter 2), which aims to obtain the residual time (an indirect measurement to measure the state condition) of a system before it fails, given the observed condition monitoring data.

Once the state or condition of the equipment is known, maintenance actions, which involve manpower, equipment and tools, and spare parts, can be planned and scheduled (Al-Sultan and Duffua, 1995). However, predicting the state is often difficult because of these factors:

1. The state cannot be measured directly.
2. The information is sometimes corrupted due to measurement or other errors.
3. The transient behaviour of the failure process is unknown.

However, as we discussed in the literature review, it is helpful to have a simple description of the plant condition. Here, we propose a new model that meets the requirements and can be used to predict the state/condition of a system given monitored information, while at the same time allowing the conditional residual time prediction to be made. This new model defines the state/condition of a system in a standard way, using a discrete set of integers to represent the state of the system. It is assumed that a state can change from one to another over time, and therefore the Markov chain concept is chosen to develop the model. Furthermore, a filtering approach is adopted in order to predict the future state recursively. We also assume that the observed \( y_i \) is a function of unobserved \( x_i \) and not vice versa, which is different from the PHM. The combination of these two main approaches enables us to use the Hidden Markov Model (HMM) concept, which is in parallel with the filtering approach but mostly on discrete state spaces.
HMMs have become very popular in modern statistics, as they have been used widely in the field of speech recognition (Rabiner, 1989). The usage has been recently spread more to other areas such as bioinformatics and engineering, including machine tool monitoring (Ertunc et al., 2001), image segmentation (Li ad Gray, 2000) and fault detection and diagnosis (Bunks et al., 2000).

3.2 Background of Hidden Markov Models

An HMM is a Markov chain observed in noise (Cappe et al., 2005). This Markov chain is assumed to have a finite number of states, $x_i$ but these are not directly observable (Rabiner, 1989). Instead, we can observe another random measurement $y_i$, which is related to each state by a probability distribution. The term ‘hidden’ comes from the notion of making observations and drawing conclusions without knowing the exact states of the system, which are hidden. The structure of an HMM may be represented graphically (Cappe et al., 2005), as in Figure 3-1 below.

![Figure 3-1: Graphical representation of the structure of a hidden Markov model](image)

Figure 3-1 implies that the hidden states $x_{i+1}$ at time $t_{i+1}$ are independent of the history of the process, $x_0, \ldots, x_{i-1}$ but conditional on the previous state $x_i$ at time $t_i$. This is called the first order Markov assumption and the resulting model becomes a first-order HMM. Generally, $x_{i+1}$ may depend on part of past states until $x_j, j < i$; it is possible to obtain such a model, called an $(i + 1 - j)^{th}$ order HMM, but a higher-order HMM will have greater complexity. Similarly, the observed values of $y_{i+1}$ are independent of the
past observations $y_0, \ldots, y_i$ but conditional on the value of $x_{i+1}$. Hence, in order to define the HMM, the following elements are needed:

1. A set of hidden states, which are the actual states of the process, but cannot be observed directly.
2. Transition probabilities $P(x_i \mid x_{i-1})$, which describes the probability of moving from state $x_{i-1}$ at time $t_{i-1}$ to $x_i$ at time $t_i$.
3. A set of values that can be observed or measured, which is a random function of the current state.
4. Observation probability density function (pdf.) $p(y_i \mid x_i)$, which describes the pdf. of observing $y_i$ given that the actual state is $x_i$ at time $t_i$.

The relationship between the system and the data observed is best explained by way of an example. Consider some observed data represented by vibration monitoring readings, as shown in Figure 3-2 below. $y_i$ is the observed data and $x_i$ shows the hidden states of the system.

![Figure 3-2: Processes of observation and hidden states](image)

At each particular time $t_i$, condition-monitoring data is collected through an observation, $y_i$. We assume that condition-monitoring data is stochastically correlated with the true state $x_i$, which is hidden. The true state may be assumed to be describable by a set of terms such as ‘good’, ‘minor defect’, ‘major defect’ or ‘failed’. We are
interested in what the system state is at any particular moment and beyond. Surely the exact states cannot be known but their probabilities can be calculated. Hence, at each time $t_i$, we seek to find the probability of $x_i$, the true state of the system, given the monitored condition information history $\mathcal{Y}_i = \{y_1, y_2, \ldots, y_i\}$, namely $P(x_i | \mathcal{Y}_i)$.

### 3.3 Modelling Methodology

In this section, we use the concept of HMM to establish a model to predict the probability of the true states $x_i$, given $\mathcal{Y}_i$, i.e. $P(x_i | \mathcal{Y}_i)$. For the purpose of model building, we adopt the following assumptions and notation:

1. The state of a monitored system deteriorates according to a discrete Markov chain $X_i$, where $X_i$ is the random variable for the true state at time $t_i$ with $x_i$ of its realization.
2. Once the state is $x_i$, it cannot go back to state $x_{i-1}$.
3. Condition monitoring takes place at time $t_i$, $i = 1, \ldots, n$, where the intervals between consecutive monitoring checks may be equal or unequal.
4. The measurement obtained at time $t_i$ by condition monitoring is denoted by a random variable $Y_i$, with $y_i$ of its realization, which is assumed to be correlated with the true state $x_i$. Note that $Y_i$ could be a vector if several types of collected information are used.
5. The condition monitoring history obtained up to time $t_j$ is denoted by $\mathcal{Y}_j$, where $\mathcal{Y}_j = \{y_1, y_2, \ldots, y_j\}$.
6. $L_j$ is a random variable for the stage $j = \{1, 2\}$ duration with $l_j$ of its realization.
7. The equipment considered here is a single-component system subject to one dominant failure mode.

In order to demonstrate the basic model, we apply the concept of a two-stage failure process, introduced by Christer and Waller (1984) with the first stage being the length of time from new to the first point when a defect is identified, $u$ is $L_1$; the second stage running from this point to a failure is $L_2$, if no maintenance intervention is involved. At each stage, the state and the duration of each state in the system are defined. Figure 3-3
shows an illustration of the two-stage failure process with three system states with their durations before failure.

![Diagram of two-stage failure process]

Figure 3-3: Two-stage failure process

We are interested in identifying:

2 at time $t_i$, the probability of a particular state given past observed monitoring information;
3 at time $t_i$, the expected time to the initial point of the second stage if $t_i < l_1$; and
4 at time $t_i$, the expected time to failure if $t_i < l_1 + l_2$.

In order to solve the first problem, let us assume that there are three states, $X_i \in \{1, 2, 3\}$ in the system, where 1 means a normal state, 2 is a defective and 3 is a failed state. The failed state is always identifiable, but the other two stages are random in their durations and not directly observable. To model this transitional state process as a Markov chain, we need to formulate the transitional matrix. We define the transitional probability that, given a system in state $k$ at time $t_{i-1}$, the system will be in state $m$ at time $t_i$ as $P(X_i = m \mid X_{i-1} = k)$. Next, the relationship between $y_i$ and $x_i$ is yet to be identified. It is assumed that a probability density function can be used to describe the relationship, namely $p(y_i \mid x_i)$, as suggested by Wang and Christer (2000).

To find the probability of $x_i$ given $\mathcal{Z}_i = \{y_i, y_{2i}, ..., y_{ni}\}$ at every monitoring point time $t_i$, namely $P(x_i \mid \mathcal{Z}_i)$, we adopt a filtering technique shown in Wang and Christer (2000), that gives $P(x_i \mid \mathcal{Z}_i)$ as

$$P(x_i \mid \mathcal{Z}_i) = P(x_i \mid y_i, \mathcal{Z}_{i-1}) = \frac{P(x_i, y_i \mid \mathcal{Z}_{i-1})}{p(y_i \mid \mathcal{Z}_{i-1})}$$

(3-1)
Using another chain rule, the joint distribution for \( P(x_i, y_i \mid \mathcal{Z}_{i-1}) \) and \( p(y_i \mid \mathcal{Z}_{i-1}) \) is given by Wang (2002) as

\[
P(x_i, y_i \mid \mathcal{Z}_{i-1}) = p(y_i \mid x_i, \mathcal{Z}_{i-1})P(x_i \mid \mathcal{Z}_{i-1}) = p(y_i \mid x_i)P(x_i \mid \mathcal{Z}_{i-1}) \tag{3-2}
\]

where we use \( p(y_i \mid x_i, \mathcal{Z}_{i-1}) = p(y_i \mid x_i) \) from the Markovian assumption stated earlier.

\( P(x_i \mid \mathcal{Z}_{i-1}) \) has the following form

Then, \( p(y_i \mid \mathcal{Z}_{i-1}) \) would have the following form if \( x_i \) were continuous

\[
p(y_i \mid \mathcal{Z}_{i-1}) = \int p(y_i \mid x_i, \mathcal{Z}_{i-1})P(x_i \mid \mathcal{Z}_{i-1})dx_i = \int p(y_i \mid x_i)P(x_i \mid \mathcal{Z}_{i-1})dx_i \tag{3-3}
\]

As the true state \( x_i \) is a discrete variable, where \( N \) is the number of states excluding the failed state, the use of another chain rule enables \( P(x_i \mid \mathcal{Z}_{i-1}) \) to be written as

\[
P(x_i \mid \mathcal{Z}_{i-1}) = \sum_{x_{i-1}=1}^{N} P(x_i \mid x_{i-1})P(x_{i-1} \mid \mathcal{Z}_{i-1}) \tag{3-4}
\]

and

\[
p(y_i \mid \mathcal{Z}_{i-1}) = \sum_{x_i=1}^{N} p(y_i \mid x_i)P(x_i \mid \mathcal{Z}_{i-1}) \tag{3-5}
\]

Hence, it can be shown that \( P(x_i \mid \mathcal{Z}_{j}) \) can be written as

\[
P(x_i \mid \mathcal{Z}_{j}) = \frac{p(y_i \mid x_i) \sum_{x_{i-1}=1}^{N} P(x_i \mid x_{i-1})P(x_{i-1} \mid \mathcal{Z}_{i-1})}{\sum_{x_i=1}^{N} p(y_i \mid x_i) \sum_{x_{i-1}=1}^{N} P(x_i \mid x_{i-1})P(x_{i-1} \mid \mathcal{Z}_{i-1})} \tag{3-6}
\]

The same formulation was proposed by Jazwinski (1970) as a method for estimating the state of a stochastic dynamic system from a noisy observation. It can be seen from equations (3-4) and (3-5) that if \( P(x_0 \mid \mathcal{Z}_0) \), \( P(x_i \mid x_{i-1}) \) and \( p(y_i \mid x_i) \) \( i=1,2,... \) are known, then equation (3-6) may be solved. Since \( \mathcal{Z}_0 \) is not available in most cases, we
can set \( P(x_0 \mid \mathcal{Z}_0) = P(x_0) \). In our case it is assumed that \( P(x_0) = 1 \) as we start our system from new. Since we assume that the system starts from new, at \( i = 0 \), the true state \( x_0 \) can only be a normal state that is 1, hence equation (3-6) for \( i = 1 \) can be written as

\[
P(x_1 \mid \mathcal{Z}_1) = \frac{p(y_1 \mid x_1)P(x_1 \mid X_0 = 1)P(x_0 \mid \mathcal{Z}_0)}{\sum_{x_i=1} p(y_1 \mid x_i)P(x_1 \mid X_0 = 1)P(x_0 \mid \mathcal{Z}_0)}
\]  

(3-7)

Hence, for every \( i = 2,3,.... \) \( P(x_i \mid \mathcal{Z}_i) \) can be calculated from equation (3-6) recursively.

### 3.4 Formulation of the Transition Probabilities

Now, we go into the details of the formulation of \( P(X_i = m \mid X_{i-1} = k) \) and \( p(y_i \mid x_i) \).

First, we adopt the concept of delay-time modelling (Christer and Waller, 1984) to model the transitional probability. Using a Markov model, it is assumed that \( x_i \) must always be in one of a finite number of discrete states, \( X_i \in \{1,2,3\} \). We define \( F_1(l_1) \) and \( F_2(l_2) \) as the cumulative probability of random variables \( L_1 \) and \( L_2 \). The transition probabilities that state \( m \) will occur at time \( t_i \), given that the current state is \( k \) at time \( t_{i-1} \), are as follows, see Christer et al. (2001).

\[
P(X_i = 1 \mid X_{i-1} = 1) = P(L_1 > t_i \mid L_1 > t_{i-1}) = \frac{\int_{l_{i-1}}^{l_i} f_1(l_1) dl_1}{\int_{l_{i-1}}^{\infty} f_1(l_1) dl_1} = \frac{1 - F_1(t_i)}{1 - F_1(t_{i-1})}
\]  

(3-8)

\[
P(X_i = 2 \mid X_{i-1} = 1) = P(L_2 > t_i - l_1, L_1 < t_i \mid L_1 > t_{i-1})
\]

\[
= \frac{\int_{l_{i-1}}^{l_i} f_1(l_1) \int_{l_i}^{\infty} f_2(l_2) dl_2 dl_1}{\int_{l_{i-1}}^{\infty} f_1(l_1) dl_1} = \frac{\int_{l_{i-1}}^{l_i} f_1(l_1)(1 - F_2(t_i - l_1)) dl_1}{1 - F_1(t_{i-1})}
\]  

(3-9)
Equation (3-10) assumed that if the item failed before \( t_i \), it will stay failed until \( t_i \).

\[
P(X_i = 1 \mid X_{i-1} = 2) = 0 \tag{3-11}
\]

\[
P(X_i = 2 \mid X_{i-1} = 2) = P(L_2 > t_i - l_i, L_4 < t_i \mid L_2 > t_{i-1} - l_i, L_4 \in (0, t_{i-1}))
\]

\[
= \frac{\int_{t_{i-1}}^{t_i} f_1(l_i) \int_{l_i}^{t_i} f_2(l_2) dl_2 dl_1}{1 - F_1(t_i)} = \frac{\int_{t_{i-1}}^{t_i} f_1(l_i)(1 - F_2(t_i - l_i)) dl_1}{1 - F_1(t_i)} \tag{3-12}
\]

\[
P(X_i = 3 \mid X_{i-1} = 2) = P(L_2 < t_i - l_i, L_4 < t_i \mid L_2 > t_{i-1} - l_i, L_4 \in (0, t_{i-1}))
\]

\[
= \frac{\int_{t_{i-1}}^{t_i} f_1(l_i) \int_{l_i}^{t_i} f_2(l_2) dl_2 dl_1}{1 - F_1(t_i)} = \frac{\int_{t_{i-1}}^{t_i} f_1(l_i)(F_2(t_i - l_i) - F_2(t_{i-1} - l_i)) dl_1}{1 - F_1(t_i)} \tag{3-13}
\]

\[
P(X_i = 1 \mid X_{i-1} = 3) = 0 \tag{3-14}
\]

\[
P(X_i = 2 \mid X_{i-1} = 3) = 0 \tag{3-15}
\]

\[
P(X_i = 3 \mid X_{i-1} = 3) = 1 \tag{3-16}
\]

Note that the stationary assumption is no longer valid, as at every time \( t_i \), the transition probability is clearly time-dependent. This is the property we needed, since most existing applications of Markov models are time-independent.
3.5 Formulation of the Relationship between the Observed Data and the Hidden State

In order to specify the relationship between the observed data $y_i$ and the hidden state $x_i$, we propose to use the method of floating scale parameters. This can be achieved by assuming that $y_i$ will follow a two-parameter family of distributions, such as gamma or Weibull, denoted by a probability density function (pdf.), $f(y_i; \alpha, \beta)$, where $\alpha$ is the scale and $\beta$ is the shape parameter respectively. Then, by letting

$$ p(y_i | x_i) = \begin{cases} f(y_i; \alpha_1, \beta) & \text{if } X_1 = 1 \\ f(y_i; \alpha_2, \beta) & \text{if } X_2 = 2 \\ \end{cases} $$

(3-17)

the relationship between $y_i$ and $x_i$ can be established. The details of this establishment of the relationship will be shown when presenting the development of the model using simulated and real data.

3.6 Modelling Development

We started the development of the model by observing the data in Wang (2002), which shows a typical stochastic nature of a bearing life distribution with vibration information. See Figure 3-4 below.

![Figure 3-4: Vibration data for six bearings](image)
Figure 3-4 shows the data of the overall vibration level obtained from six rolling element bearings, which is conducted in a laboratory fatigue experiment several years ago. Generally, the overall vibration level is a measure of the total vibration amplitude over a wide range of frequencies. The velocity amplitude in this case is expressed in term of root mean square (rms) value, which tells us the energy of the vibration. It is noted, this is the only condition monitoring data that are available to us, since other associated information about the experiment such as size and type of the bearings, running speed and loading have been lost. Figure 3-4 shows that for every bearing, the overall vibration level was fairly flat until a time where it started to increase rapidly. This illustrates the idea of the two-stage failure process stated by Christer and Waller (1984) as discussed in Chapter 2.

Since this is an early developmental stage of model building, we are reluctant to use this dataset to develop our model. Hence, a simulation approach was purposely used to approximate the pattern of the data and test the feasibility of our model. By using simulation, we know the true parameters and the model used, so it helps us to check whether we can successfully recover the model parameters from simulated data. Also, using these simulated $\tilde{y}_i$, we tested our model to see whether it was able to predict the underlying state accurately. Generally, by simulation it is easy to validate a model under development, by comparing the results produced by the model, since we know all the $l_1$’s and $l_2$’s. Figure 3-5 below shows the set-up algorithm for the simulation.
We now demonstrate the modelling development through a numerical example based on our simulated data. For this example, to calculate $P(x_1 | \mathcal{F}_i)$ we assume:

$$f_1(l_1) = \lambda_1 e^{-\lambda_1 l_1} \quad (3-18)$$
$$f_2(l_2) = \lambda_2 e^{-\lambda_2 l_2} \quad (3-19)$$
$$F_1(l_1) = P(L_1 \leq l_1) = 1 - e^{-\lambda_1 l_1} \quad (3-20)$$
$$F_2(l_2) = P(L_2 \leq l_2) = 1 - e^{-\lambda_2 l_2} \quad (3-21)$$
with \( \lambda_2 \), \( \lambda_2 = 0.025 \) and \( \lambda_2 = 0.05 \) purposely selected in order to repeat the pattern in

**Figure 3-4.** As the relationship between \( y_i \) and \( x_i \) can be treated through a conditional distribution \( p(y_i \mid x_i) \), a Weibull distribution is chosen for the time being. The distribution of \( y_i \) given \( x_i \) is given below:

\[
p(y_i \mid x_i) = \begin{cases} 
\alpha_i \beta(\alpha_i y_i)^{\beta-1} e^{-(\alpha_i y_i)^\beta} & \text{if } x_i = 1 \\
\hat{\alpha}_2 \beta(\hat{\alpha}_2 y_i)^{\beta-1} e^{-(\hat{\alpha}_2 y_i)^\beta} & \text{if } x_i = 2 
\end{cases}
\]

where \( \hat{\alpha}_2 \approx \frac{0}{\int_0^t f_i(l_i)dl_i} \) equation (3-22)

\[
\int_0^t \frac{\alpha_i}{1 + \alpha_i b(t_i - l_i)} f_i(l_i)dl_i
\]

Equation (3-22) requires explanation, particularly when \( X_i = 2 \). It is known that during the normal working period of \( l_i \), the measured condition information data may display no particular trend and fluctuate around a constant mean; then by letting the scale parameter be \( \alpha_i \) while the shape parameter is fixed, we have \( E(Y_i \mid X_i = 1) \propto 1/\alpha_i \).

During the second stage, the readings may start to increase and display a trend, in which case we assume \( E(Y_i \mid X_i = 2) \propto 1/\alpha_2 = 1/\alpha_i + b \ast (t_i - l_i) \); that is, the expected value of the observed data when \( X_i = 2 \) is approximately a linear function of the time since the start of \( l_2 \), with the intercept approximately equal to the expected value of the observed data in stage 1. It is straightforward to show that

\[
\alpha_2 = \frac{\alpha_i}{1 + \alpha_i b(t_i - l_i)}
\]

Since we don’t know exactly when \( L_i \) finishes, we use an expected value for \( \alpha_2 \) conditional on \( L_i < t_i \), which is given by equation (3-22). **Figure 3-6** demonstrates the working principle of the above modelling idea.
The above is just one of many possible ideas to model the relationship between $y_i$ and $x_i$ motivated by the vibration-based monitoring cases and the real-life data we have. It is possible to have an exponential function instead of a linear relationship between $y_i$ and $x_i$, but let us use the linear case first.

### 3.7 Numerical Examples

Now, assuming all the parameter values are known, as in Table 3-1, we run the simulation. Using equation (3-6), we show two results to describe our initial findings, as in Figure 3-7 and Figure 3-8 below.

<table>
<thead>
<tr>
<th>$\alpha_1$</th>
<th>$\beta$</th>
<th>$b$</th>
<th>$\lambda_1$</th>
<th>$\lambda_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2176</td>
<td>4.0</td>
<td>0.8</td>
<td>0.05</td>
<td>0.025</td>
</tr>
</tbody>
</table>

Table 3-1: Simulation values
The simulation provides almost the same pattern data as in Figure 3-4. The data $y_i$ in both figures are generated from simulation with given $l_1$ and $l_2$, so we know the timing for the start of $l_2$ or the length of $l_1$. For case 1, $l_1 = 18.8$ and the model identifies its defective state with probability 1 at time = 22. In case 2, $l_1 = 116.3$ and the defective state was identified at time = 120. In both cases the model performed well, with a slight delay due to the usual filtering delay. It is interesting to note that in case 2 the prediction was not accurate at the earlier stages of the system’s life, due to a relatively large
fluctuation of the data observed. As an example, at epoch $t_{12}$ the readings show that $y_{12}$ is near the expected mean of $y$, but there is a sudden jump at $t_{18}$ followed by a rapid decrease to the expected level of $y$ at $t_{22}$. The large changes in data readings between the two checking points affected the probability; however, it became stable when more data became available. Here, we show the capability of the model to identify the fault, but it is only limited to a two-stage failure process. It can be extended to more than 2 stages failure process where 4 or more states may be required.

3.8 Parameter Estimation

In order to approximate the real situation where the model parameters are unknown, we seek to estimate the model parameters from our simulated data and compare them with the actual values supplied during the simulation. If the two sets of values are close, the method can be said to be feasible and robust. In this thesis, two methods of parameter estimation are reviewed, namely, least-squares estimation (LSE) and maximum likelihood estimation (MLE). MLE exhibits several criteria to be a good estimator (John, 1997). These criteria include:

1. MLE is unbiased, i.e. its bias tends to zero as the number of samples increases to infinity.

2. MLE is efficient, i.e. it has minimal variance when the number of samples tends to infinity.

As the number of samples increases, the distribution of the MLE tends to the Gaussian distribution with mean $\theta$, where $\theta$ are the parameters to be estimated. This enables us to calculate the variance and covariance of the estimated parameters by taking the inverse of the Fisher information matrix.

In contrast, no such statistical characteristic can be pointed out in LSE (Myung, 2003), accepting it as a method to find the parameters that minimize the sum of the squares of residuals. Since in this study we are dealing with a large sample size, we will use MLE as our approach of parameter estimation, until another method is specified.
### 3.8.1 Maximum Likelihood Estimator

The idea behind maximum likelihood is to obtain the most likely values of the parameters for a given likelihood function that best describes the data. In our case, \( y_i \) is the only data gathered from the observed information, which may depend on past observations, so the likelihood function for one component life cycle can be written as

\[
L(\theta \mid \mathcal{Y}_T) \propto \prod_{i=1}^{T} p(y_i \mid \mathcal{Y}_{i-1}, \theta)
\]

\[
= \prod_{i=1}^{T} \left\{ \sum_{y_i=x_{i-1}, x_{i-1}}^{N} \sum_{y_i=x_{i-1}, x_{i-1}}^{N} p(y_i \mid x_i, \theta_1) P(x_i \mid x_{i-1}, \theta_2) P(x_{i-1} \mid \mathcal{Y}_{i-1}, \theta) \right\}
\]

where \( \theta = \{\theta_1, \theta_2\} \) is a set of unknown parameters, \( T \) is the number of monitoring checks during the cycle, \( N \) is the number of states, \( y_i \) is the condition monitoring reading at time \( t_i \) and \( \mathcal{Y}_{i-1} = \{y_1, y_2, \ldots, y_{i-1}\} \). Parameter estimates that maximize the likelihood of equation (3-24) are the most likely values for the parameters. However, the products of the likelihood could turn out to be very small numbers and inconvenient for computations. For numerical convenience, the log of the likelihood is used. Therefore, equation (3-24) can be rewritten as

\[
\log L(\theta \mid \mathcal{Y}_T) = \sum_{i=1}^{T} \log p(y_i \mid \mathcal{Y}_{i-1}, \theta)
\]

\[
= \sum_{i=1}^{T} \log \left\{ \sum_{y_i=x_{i-1}, x_{i-1}}^{N} \sum_{y_i=x_{i-1}, x_{i-1}}^{N} p(y_i \mid x_i, \theta_1) P(x_i \mid x_{i-1}, \theta_2) P(x_{i-1} \mid \mathcal{Y}_{i-1}, \theta) \right\}
\]

In our case, we could have more than one component life cycle; hence we simulated five life cycles, which are shown in Figure 3-9.
Each set contains one complete life cycle with \( T \) (variable) monitoring checks in each cycle. Similarly, the likelihood for \( m \) life cycles is given by

\[
L(\theta \mid \mathcal{Z}_{mT}) = \prod_{j=1}^{m} \prod_{i=1}^{T_j} p(y_j \mid \mathcal{Z}_{j-1}, \theta)
\]

\[
= \prod_{j=1}^{m} \prod_{i=1}^{T_j} \left\{ \sum_{x_i=x_{i-1}, x_{i+1}=1}^{N} p(y_{ji} \mid x_{ji}, \theta_1) P(x_{ji} \mid x_{j-1}, \theta_2) P(x_{j-1} \mid \mathcal{Z}_{j-1}, \theta) \right\} \quad (3-26)
\]

where \( m \) is the number of complete life cycles, \( \theta = \{\theta_1, \theta_2\} \) is a set of unknown parameters, \( T_j \) is the number of monitoring checks of the \( jth \) cycle, \( y_{ji} \) is the condition monitoring reading at \( t_j \) for the \( jth \) cycle and the \( ith \) monitoring, \( x_{ji} \) and \( x_{ji-1} \) are the states of the system at \( t_i \) and \( t_{i-1} \) for the \( jth \) cycle and \( ith \) monitoring respectively, and \( \mathcal{Z}_{j-1} = \{y_{j1}, y_{j2}, \ldots, y_{ji-1}\} \).

Similarly, for numerical convenience, the log of the likelihood is used. Therefore, equation (3-26) can be rewritten as
\[
L(\theta \mid \mathcal{Z}_{mT}) \propto \sum_{j=1}^{m} \sum_{i=1}^{T_j} \log p(y_{ij} \mid \mathcal{Z}_{j-1}, \theta)
\]

\[
= \sum_{j=1}^{m} \sum_{i=1}^{T_j} \log \left\{ \sum_{x_{ji}}^{N} \sum_{x_{j+1}}^{N} p(y_{ji} \mid x_{ji}, \theta_1)P(x_{ji} \mid x_{j-1}, \theta_2)P(x_{j+1} \mid \mathcal{Z}_{j-1}, \theta) \right\}
\]

(3-27)

We coded equation (3-27) in FORTRAN95 and used NAG E04JYF optimisation routines to obtain the values of the estimated parameters \(\alpha_1, \beta, b, \lambda_1\) and \(\lambda_2\) as shown in Table 3-2.

<table>
<thead>
<tr>
<th>Estimated Value</th>
<th>(\hat{\alpha}_1 = 0.2062)</th>
<th>(\hat{\beta} = 3.1710)</th>
<th>(\hat{b} = 0.8109)</th>
<th>(\hat{\lambda}_1 = 0.0231)</th>
<th>(\hat{\lambda}_2 = \text{not available})</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Value</td>
<td>(\alpha_1 = 0.2176)</td>
<td>(\beta = 4.0)</td>
<td>(b = 0.8)</td>
<td>(\lambda_1 = 0.05)</td>
<td>(\lambda_2 = 0.025)</td>
</tr>
</tbody>
</table>

Table 3-2: The estimated parameters and their true values

We obtained good estimated values for \(\hat{\alpha}, \hat{\beta}, \hat{b}\) and \(\hat{\lambda}_1\), but had no success for \(\hat{\lambda}_2\), as it tended to the lowest setting boundary in the optimisation routine. In order to explain this, let us look at each monitoring point where we have \(p(y_i \mid \mathcal{Z}_i)\) until the last monitoring point, \(p(y_{T_j} \mid \mathcal{Z}_{T_j})\), before failure occurs, as shown in Figure 3-10. Here, the failure information is not used and because of that, the likelihood of equation (3-26) contains little information on the length \(l_2\), which is the reason \(\hat{\lambda}_2\) is “non-identifiable” from these data.

![Figure 3-10: The likelihood function at each monitoring point](image)

By incorporating the failure information, the probability of \(X_{T_{j+1}} = 3\) given \(\mathcal{Z}_{jT_j}\) is
\[
P(X_{T_j+1} = 3 \mid \mathcal{Z}_{jT_j}) = \]
\[
P(X_{jT_j+1} = 3 \mid x_{jT_j} = 1)P(X_{T_j} = 1 \mid \mathcal{Z}_{jT_j}) + P(X_{jT_j+1} = 3 \mid x_{jT_j} = 2)P(X_{jT_j} = 2 \mid \mathcal{Z}_{jT_j})
\]

(3-28)

Multiplying this term in equation (3-26), the likelihood function now becomes

\[
L(\theta \mid \mathcal{Z}_{mT}) \propto \prod_{j=1}^{m} \prod_{i=1}^{T_j} \left\{ \sum_{y_{ji}} \sum_{x_{ji}=0, \ldots, x_{ji}=N} p(y_{ji} \mid x_{ji}, \theta_{ji})P(x_{ji} \mid x_{ji-1}, \theta_{ji-1})P(x_{ji-1} \mid \mathcal{Z}_{ji-1}, \theta_{ji-1}) \right\} P(X_{jT_j+1} = 3 \mid \mathcal{Z}_{jT_j})
\]

(3-29)

By taking logs on both sides of equation (3-29) and maximizing the log likelihood, we coded it using the same E04JYF routines and obtained the results as shown in Table 3-3 below.

<table>
<thead>
<tr>
<th>Estimated Value</th>
<th>$\hat{\alpha}_1 = 0.2056$</th>
<th>$\hat{\beta} = 3.1702$</th>
<th>$\hat{b} = 0.8126$</th>
<th>$\hat{\lambda}_1 = 0.0232$</th>
<th>$\hat{\lambda}_2 = 0.0362$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Value</td>
<td>$\alpha_1 = 0.2176$</td>
<td>$\beta = 4.0$</td>
<td>$b = 0.8$</td>
<td>$\lambda_1 = 0.05$</td>
<td>$\lambda_2 = 0.025$</td>
</tr>
</tbody>
</table>

Table 3-3: The estimated parameters and true values using the likelihood function with failure information

The results show a significant improvement in the estimated value of $\lambda_2$ by incorporating the failure information, which is shown to be a vital piece of information needed for parameter estimation, although it may not be available in practice.

Although the likelihood function with failure information was successful for estimating parameters and some of the results were very good, it was felt that the results might be better if another technique was used; hence we turned our attention to the Expectation-Maximization (EM) algorithm, which is particularly useful in HMM (Rabiner, 1989).
3.8.2 Expectation-Maximization (EM) Algorithm

The Expectation-Maximization algorithm was originally developed by Dempster et al. (1977) for finding the maximum-likelihood estimate of the parameters of an underlying distribution from a given data set when the data is incomplete or has missing values (Bilmes, 1998). To demonstrate this algorithm using our data, we start with the general notation of the EM algorithm given by Bilmes (1998). Assume data \( y \) is observed and called as incomplete data and \( x \) denotes the hidden or missing value. A complete dataset exists, which is \( z = (y, x) \) and has a joint density function as follows:

\[
p(z \mid \theta) = p(y, x \mid \theta) = p(y \mid x, \theta)p(x \mid \theta)
\]

Here we assume both \( x \) and \( y \) are continuous.

This joint density function describes a joint relationship between the missing and observed values. With this density function, we can define a likelihood function of dataset \( z = (y, x) \) as

\[
L(\theta \mid z) \propto L(\theta \mid y, x) = p(y, x \mid \theta)
\]

which is called the complete-data likelihood. The EM algorithm consists of two steps, the first being to find the expected value of the complete data log likelihood \( \log p(y, x \mid \theta) \) with respect to the unknown state, \( x \), given the observed data \( y \) and the current parameter estimates, \( \theta^j \). We define this expected log likelihood as

\[
Q(\theta, \theta^j) = E[\log p((y, x \mid \theta) \mid x, \theta^j)]
\]

where

\[
E[\log p(y, x \mid \theta) \mid x, \theta^j] = \int_{x \in X} \log p(y, x \mid \theta)p(x \mid \theta^j)dx
\]

and where \( X \) is the complete set of \( x \). The second step is to find the maximum of the expectation that we have computed in the first step, that is \( \max Q(\theta, \theta^j) \). To obtain
$Q^{j+1}$, these two steps are repeated until $|Q^{j+1} - Q^j| < tol$, where tol is a pre-set tolerance level.

Applying this to our case study, given the observed value, the complete data likelihood has the following form, in which $p(y_i \mid \mathcal{S}_{i-1}, \theta)$ describes the observation values and $P(x_i \mid \mathcal{S}_{i-1}, \theta)$ describes the hidden states. Note $X_i$ is a discrete random variable in our case. To calculate the estimate parameter using one life cycle of data using the EM approach, the algorithm is as follows:

**Start:** an initial estimate of $\theta^{(0)}$

**E-step:** calculate the conditional expectation

$$Q(\theta, \theta^j) = \sum_{i=1}^{T} \log \left\{ \sum_{x_i} p(y_i \mid x_i, \theta_j) P(x_i \mid \mathcal{S}_{i-1}, \theta^j) \right\}$$

$$= \sum_{j=1}^{T} \log \left\{ \sum_{x_i} \sum_{x_{i-1}} p(y_i \mid x_i, \theta_j) P(x_i \mid x_{i-1}, \theta_j) P(x_{i-1} \mid \mathcal{S}_{i-1}, \theta) \right\} P(x_i \mid \mathcal{S}_{i-1}, \theta^j)$$

where $\theta = \{\theta_1, \theta_2, \ldots\}$ is a set of unknown parameters, $T$ is the number of monitoring checks during the cycle, $N$ is the number of states, $y_i$ is the condition monitoring reading at time $t_i$ and $\mathcal{S}_{i-1} = \{y_1, y_2, \ldots, y_{i-1}\}$.

**M-step:** maximize $Q(\theta, \theta^j)$ to obtain the next estimate, max $Q(\theta, \theta^j)$ to obtain $Q^{j+1}$.

Similarly, if we have $m$ life cycles, the new algorithm is as follows

**Start:** an initial estimate of $\theta^{(0)}$

**E-step:** calculate the conditional expectation

$$Q(\theta, \theta^j) = \sum_{j=1}^{m} \sum_{i=1}^{T_j} \log \left\{ \sum_{x_{i-j}} p(y_{i-j} \mid x_{i-j}, \theta_j) P(x_{i-j} \mid \mathcal{S}_{i-j-1}, \theta) \right\}$$

$$= \sum_{j=1}^{m} \sum_{i=1}^{T_j} \log \left\{ \sum_{x_{i-j}} \sum_{x_{i-j-1}} p(y_{i-j} \mid x_{i-j}, \theta_j) P(x_{i-j} \mid x_{i-j-1}, \theta_j) P(x_{i-j-1} \mid \mathcal{S}_{i-j-1}, \theta) \right\} P(x_{i-j} \mid \mathcal{S}_{i-j-1}, \theta^j)$$
where \( m \) is the number of complete life cycles, \( \theta = \{\theta_1, \theta_2\} \) is a set of unknown parameters, \( T_j \) is the number of monitoring checks of the \( j \text{th} \) cycle, \( y_j \) is the condition monitoring reading at \( t_j \) for the \( j \text{th} \) cycle and the \( i \text{th} \) monitoring, \( x_{ji} \) and \( x_{ji-1} \) are the states of the system at \( t_i \) and \( t_{i-1} \) for the \( j \text{th} \) cycle and \( i \text{th} \) monitoring respectively, and 
\[
\mathcal{I}_{ji} = \{y_{ji}, y_{j2}, \ldots, y_{ji-1}\}.
\]

**M-step:** maximize \( Q(\theta, \theta^i) \) to obtain the next estimate, \( \max Q(\theta, \theta^i) \) to obtain \( Q^{i+1} \).

Applying this to our simulated data without using failure information, we obtained the results in Table 3-4 below.

<table>
<thead>
<tr>
<th>Estimated Value</th>
<th>( \hat{\alpha}_1 = 0.2033 )</th>
<th>( \hat{\beta} = 3.1886 )</th>
<th>( \hat{b} = 0.8013 )</th>
<th>( \hat{\lambda}_1 = 0.0239 )</th>
<th>( \hat{\lambda}_2 ) = not available</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Value</td>
<td>( \alpha_1 = 0.2176 )</td>
<td>( \beta = 4.0 )</td>
<td>( b = 0.8 )</td>
<td>( \lambda_1 = 0.05 )</td>
<td>( \lambda_2 = 0.025 )</td>
</tr>
</tbody>
</table>

Table 3-4: The estimated parameters and true values \( Q(\theta, \theta^i) \)

In our expectations, the EM should be better since, in theory, we cannot observe the hidden state and the algorithm is guaranteed to converge to a local maximum of the likelihood function (Dempster et al., 1977). Surprisingly, the EM procedure explained above produces almost the same result as the ordinary likelihood function without the failure information; see (Table 3-2). It is believed that the complete likelihood function in the E-step is not enough to describe the data. Therefore, the function for the E-step was added with failure information, as follows:

\[
Q(\theta, \theta^i) = \sum_{i=1}^{N} \left\{ \log \left( p(y_i \mid \mathcal{I}_{i-1}, \theta) \right) P(x_i \mid \mathcal{I}_{i-1}, \theta^i) \right\} + \log \left( P(X_{N+1} = 3 \mid \mathcal{I}_N) \right) \quad (3-34)
\]

where \( t_{N+1} \) is the failure point. When the same procedure was reapplied for \( m \) sets of data, it generated a set of new estimates, shown in Table 3-5 below.
Again, the EM procedure produces almost the same results as the ordinary likelihood function with failure information. This is because the approach taken to formulate the EM algorithm is almost the same as for the ordinary likelihood function, except that we multiplied it with $p(x_i | \mathcal{X}_{i-1})$ in equations (3-34) to boost the optimisation calculation.

We were able to estimate $\hat{\lambda}_1$ simply because we had the information about when the time $l_1$ began and ended. In contrast, we were unable to estimate $\hat{\lambda}_2$ because we did not have enough information concerning when the time $l_2$ ended. However, the failure information provides new data on the ending of the time $l_2$, allowing the parameter $\hat{\lambda}_2$ to be estimated. In practice, failure information is difficult to obtain due to preventive replacements, but expert judgments (Zhang, 2004) can be used to overcome this problem.

### 3.9 Goodness-of-Fit Test

The point of interest now is how good our estimated parameters are. Based upon the assumption of asymptotically Gaussian estimator, we could compute the variance of estimated parameters, which is useful for constructing confidence regions for the estimated parameters. We used a numerical approach developed by Baker and Wang (1992) to calculate the covariance matrix of fitted parameters. The resulting for variances and covariances for parameter estimates $\hat{\alpha}_1$, $\hat{\beta}$, $\hat{b}$, $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are given in Table 3-6:

<table>
<thead>
<tr>
<th>Estimated Value</th>
<th>$\hat{\alpha}_1 = 0.2032$</th>
<th>$\hat{\beta} = 3.1886$</th>
<th>$\hat{b} = 0.8013$</th>
<th>$\hat{\lambda}_1 = 0.0239$</th>
<th>$\hat{\lambda}_2 = 0.016$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Value</td>
<td>$\alpha_1 = 0.2176$</td>
<td>$\beta = 4.0$</td>
<td>$b = 0.8$</td>
<td>$\lambda_1 = 0.05$</td>
<td>$\lambda_2 = 0.025$</td>
</tr>
</tbody>
</table>

Table 3-5: The estimated parameters and true values with modified $Q(\theta, \theta')$
Table 3-6: Variances and covariances of estimated parameters

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\alpha}_1$</th>
<th>$\hat{\beta}$</th>
<th>$\hat{b}$</th>
<th>$\hat{\lambda}_1$</th>
<th>$\hat{\lambda}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\alpha}_1$</td>
<td>5.509E-5</td>
<td>-2.375E-12</td>
<td>7.117E-13</td>
<td>4.432E-14</td>
<td>-9.678E-15</td>
</tr>
<tr>
<td>$\hat{\beta}$</td>
<td>-2.375E-12</td>
<td>0.037</td>
<td>1.974E-11</td>
<td>1.487E-12</td>
<td>-6.778E-14</td>
</tr>
<tr>
<td>$\hat{b}$</td>
<td>7.117E-13</td>
<td>1.974E-11</td>
<td>1.814E-3</td>
<td>-1.174E-12</td>
<td>1.578E-14</td>
</tr>
<tr>
<td>$\hat{\lambda}_1$</td>
<td>4.432E-14</td>
<td>1.487E-12</td>
<td>-1.174E-12</td>
<td>9.259E-6</td>
<td>-2.445E-15</td>
</tr>
</tbody>
</table>

The variances (diagonal) of the parameter estimates are small, which shows that our estimated parameters are good. An apparent exception is the case of $\hat{\beta}$, where the variance is relatively large; but comparing with the actual value of $\hat{\beta}$ shows it is comparatively small. The covariances are also small, which shows that there are no correlations between the parameter estimates.

An important issue to discuss here is how well the model output fits the actual data, or in other words, how the model can be validated. There are many statistical tools that can be used to measure the goodness-of-fit of a model, which can be grouped into two categories: graphical and numerical. Graphical methods illustrate a broad range of complex aspects of the relationship between the model output and the actual data, whereas numerical methods tend to be narrowly focused on a particular aspect of the relationship between the model output and the actual data, and often try to compress that information into a single number or test result.

In practice, we only managed to observe the vibration signal $y_i$ at each monitoring point from the normal state to the failure state, and we do not know what happened to the actual state during the life cycle. By producing some calculated values of $\tilde{y}_i$ via simulation at every monitoring point using the estimated parameters, we were able to compare the observed $y_i$ with the simulated $\tilde{y}_i$. Figure 3-11 below shows an example of the simulation.
From Figure 3-11 it can be seen that there is not much difference between the observed and the simulated data, which gives us more confidence in the model we have developed. To determine how well the simulated model represents the observed data, we use the Pearson product moment correlation coefficient, $R^2$, which is normally used to assess the relationship of two variables (Albright et al., 1999). One of its formulations is:

$$r = \frac{\sum (\tilde{y} - \bar{y})(y - \bar{y})}{\sqrt{\sum (\tilde{y} - \bar{y})^2 \sum (y - \bar{y})^2}}$$

(3-35)

where

$\tilde{y}$ refers to the output from the simulated data and $y$ refers to the observed data.

Normally, $r$ could take any value from $+1$ to $-1$, where a positive sign means a positive correlation and a negative sign means an inverse correlation, while if $r = 0$, there is no correlation between the two variables. However, we are interested in the coefficient value; thus, taking the square of $r$ into $R^2$ would not change its relationship. This statistic $R^2$ measures the correlation between the observed values and the simulated values (Albright et al., 1999). Once we have learned the correlation coefficient $R^2$, we need to test the significance of $R^2$. To do this, we need to set the critical alpha level that defines the error rate we are willing to accept (normally 0.05 or 0.01) and we will use critical value tables for Pearson’s correlation coefficient. If our
$R^2$ statistic is higher than the critical value from the table, we could say our relationship is significant.

We calculated the $R^2$ for comparison between observed $y_i$ and simulated $\tilde{y}_i$ and calculated the significance at 5% and 1% level.

The results given in Table 3-7 show that for every test, $R^2$ is always greater than the critical value at the 5% and 1% significance levels, which shows the relationship between the simulated and observed data obtained from Figure 3-9, and confirms the use of the likelihood function to obtain parameter estimates.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$R^2$</th>
<th>Alpha=0.05</th>
<th>Alpha=0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1</td>
<td>0.7699</td>
<td>0.297</td>
<td>0.384</td>
</tr>
<tr>
<td>Test 2</td>
<td>0.5303</td>
<td>0.374</td>
<td>0.478</td>
</tr>
<tr>
<td>Test 3</td>
<td>0.6932</td>
<td>0.433</td>
<td>0.549</td>
</tr>
<tr>
<td>Test 4</td>
<td>0.6849</td>
<td>0.2690</td>
<td>0.3486</td>
</tr>
<tr>
<td>Test 5</td>
<td>0.6294</td>
<td>0.3310</td>
<td>0.4870</td>
</tr>
</tbody>
</table>

**Table 3-7: $R^2$ for comparison between actual $y_i$ and simulated $\tilde{y}_i$**

Another test that we carried out to validate the model is looking at the initiation time of a random fault that occurred. Using the simulated data shows in Figure 3-9, we know when the time $l_i$ starts and these values can be compared with those from the model, which locates its defective state with probability 1. The results for this assignment are shown in Table 3-8.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Observed</th>
<th>Simulated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1</td>
<td>77.3</td>
<td>80</td>
</tr>
<tr>
<td>Test 2</td>
<td>18.4</td>
<td>22</td>
</tr>
<tr>
<td>Test 3</td>
<td>30.1</td>
<td>34</td>
</tr>
<tr>
<td>Test 4</td>
<td>49.6</td>
<td>52</td>
</tr>
<tr>
<td>Test 5</td>
<td>33.3</td>
<td>36</td>
</tr>
</tbody>
</table>

**Table 3-8: The observed and simulated values for the initial time of a random defect**

After doing some calculations, $R^2$ was found to be 0.9998, which means that our model explains 99% of the variation in the data. Again, this provides a good measure of confidence in the likelihood function.
The next values that interest us at this stage were the expected time to the initial point of
the second stage and the expected time to failure at every time \( t_i \). To model the first
situation, suppose at time \( t_i \) where the system has not yet failed, it must be either in
state 1 or 2. The expected time to the initial point of the second stage is simply the
product of the conditional expected value of \( L_1 \) given \( L_1 > t_i \) and the probability of the
system being in state 1, which is:

\[
E(L_1 - t_i, x_i \mid L_1 > t_i, \mathcal{F}_i) = P(X_i = 1 \mid Y_i) \frac{\int_{L_1}^{\infty} (l_1 - t_i) f_1(l_i) \, dl_1}{\int_{L_1}^{\infty} f_1(l_i) \, dl_1}
\]  

(3.36)

Using a similar argument, we calculated the expected time to failure as

\[
E(L_2 + L_1 - t_i, x_i \mid L_2 + L_1 > t_i, \mathcal{F}_i)
\]

\[
P(X_i = 1 \mid \mathcal{F}_i) \frac{\int_{L_1}^{\infty} f_1(l_i) \int_0^{l_1 + t_i - t_i} f_2(l_2) \, dl_2 \, dl_1}{\int_{L_1}^{\infty} f_1(l_i) \, dl_1}
\]

\[
+ P(X_i = 2 \mid \mathcal{F}_i) \frac{\int_{L_1}^{\infty} f_1(l_i) \int_{l_1 - t_i}^{\infty} f_2(l_2) \, dl_2 \, dl_1}{\int_{L_1}^{\infty} f_1(l_i) \, dl_1}
\]

(3.37)

We also formulated the distribution of the times to the initial point of the second stage
and to failure, which are given below:

\[
\text{Pdf}(L_1 - t_i = t_k, x_i \mid L_1 > t_i, \mathcal{F}_i) = P(X_i = 1 \mid \mathcal{F}_i) \frac{f_1(t_i - t_i)}{\int_{L_1}^{\infty} f_1(l_i) \, dl_1}, \quad k = i + 1, i + 2, \ldots
\]

(3.38)

and

\[
\text{pdf}(L_2 + L_1 - t_i = t_k, x_i \mid L_2 + L_1 > t_i, \mathcal{F}_i) =
\]

\[
P(X_i = 1 \mid \mathcal{F}_i) \frac{\int_{L_1}^{\infty} f_1(l_i) f_2(t_k - t_i) \, dl_1}{\int_{L_1}^{\infty} f_1(l_i) \, dl_1}
\]

\[
+ P(X_i = 2 \mid \mathcal{F}_i) \frac{\int_{L_1}^{\infty} f_1(l_i) f_2(t_k - t_i) \, dl_1}{\int_{L_1}^{\infty} f_1(l_i) \, dl_1}
\]

(3.39)

\[k = i + 1, i + 2, \ldots\]
Equation (3-39) shows how the residual time of the system could be calculated from our model. In Chapter 4, we fit the model we had developed with the observed data. We then show all the results and attempt to compare those results with a previous study conducted by Zhang (2004).

3.10 Summary

This chapter proposed a model to calculate the residual time of the monitored items by defining a $N$ state space of the system. The concept of a Hidden Markov Model is applied to a two-stage failure process. We explained this process and formulated the transition probabilities between stages and their relationship between the monitoring data and the state of the system. Using the filtering technique, the probability of the state given its monitoring history is calculated. Two parameter estimation techniques are used, namely the maximum likelihood estimator and EM algorithms. Both methods produced the desired results given that the failure information was used.
4 CHAPTER 4: EARLY FAULT IDENTIFICATION – A CASE STUDY USING VIBRATION DATA

4.1 Introduction

In this chapter, the model introduced in Chapter 3 is fitted to a set of vibration data obtained from a laboratory experiment (Wang, 2002); see Figure 3-4 in Chapter 3. Also, an attempt has been made to compare the results from this model with those of a past study using a Statistical Process Control (SPC) chart conducted by Zhang (2004).

4.2 Numerical Results

Here, we seek to identify the initial point of a random defect or an abnormal stage, as explained in Chapter 3, from observed vibration data. To start with, we need to estimate the parameters used in our model from the dataset. To do that, we randomly selected three sets of observed vibration data (Gu-b2, Gu-b5 and Gu-b1) for parameter estimations and reserved the other three (Gu-b3, Gu-b4 and Gu-b6) for model testing. Using this approach, we could verify our model later. Given all the formulations for estimating model parameters (see Chapter 3), and based on observed vibration data, we carried out the calculations and produced the results depicted in Table 4-1 below.

<table>
<thead>
<tr>
<th>$\hat{\alpha}$</th>
<th>$\hat{\beta}$</th>
<th>$\hat{\theta}$</th>
<th>$\hat{\lambda}_1$</th>
<th>$\hat{\lambda}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4314</td>
<td>3.635</td>
<td>0.0983</td>
<td>0.0089</td>
<td>0.0093</td>
</tr>
</tbody>
</table>

Table 4-1: Estimated parameters from vibration data

The estimated parameters are only point estimates; hence, to show the amount of uncertainty of these estimates, a variance-covariance matrix was produced based on the log-likelihood, equation (3-28). The variance-covariance matrix of the estimated parameters is shown in Table 4-2 below.

<table>
<thead>
<tr>
<th>$\hat{\alpha}$</th>
<th>$\hat{\beta}$</th>
<th>$\hat{\theta}$</th>
<th>$\hat{\lambda}_1$</th>
<th>$\hat{\lambda}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.593E-5</td>
<td>-3.032E-4</td>
<td>-4.926E-6</td>
<td>-2.149E-6</td>
<td>1.179E-8</td>
</tr>
<tr>
<td>-3.032E-4</td>
<td>8.481E-4</td>
<td>4.942E-5</td>
<td>1.492E-5</td>
<td>-2.316E-7</td>
</tr>
<tr>
<td>-4.926E-6</td>
<td>4.942E-5</td>
<td>2.813E-6</td>
<td>-1.058E-6</td>
<td>1.11E-8</td>
</tr>
<tr>
<td>-2.149E-6</td>
<td>1.492E-5</td>
<td>-1.058E-6</td>
<td>1.542E-6</td>
<td>-1.787E-8</td>
</tr>
</tbody>
</table>
The variance of each parameter is relatively small, which indicates that our estimation is relatively good. The covariance also shows that there is little correlation between parameters. These results allow the model to be used with confidence.

Using these estimated parameters, the $P(x_i \mid \mathcal{Y}_i)$ model was calculated. From the model, the starting point of defect (stage 2) is easily identified when $P(X_i = 2 \mid \mathcal{Y}_i)$ has a probability 1. Two of the results are shown in Table 4-3 and Table 4-4 below.

<table>
<thead>
<tr>
<th>time</th>
<th>vib. in rms</th>
<th>$P(X_i = 1 \mid \mathcal{Y}_i)$</th>
<th>$P(X_i = 2 \mid \mathcal{Y}_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>1.6012</td>
<td>0.9535</td>
<td>4.65E-02</td>
</tr>
<tr>
<td>25.5</td>
<td>1.7306</td>
<td>0.9448</td>
<td>5.52E-02</td>
</tr>
<tr>
<td>35</td>
<td>1.8917</td>
<td>0.9680</td>
<td>3.19E-02</td>
</tr>
<tr>
<td>50</td>
<td>2.4231</td>
<td>0.9587</td>
<td>4.13E-02</td>
</tr>
<tr>
<td>57</td>
<td>2.5086</td>
<td>0.9749</td>
<td>2.50E-02</td>
</tr>
<tr>
<td>74</td>
<td>3.4853</td>
<td>0.6641</td>
<td>0.335847</td>
</tr>
<tr>
<td><strong>85</strong></td>
<td><strong>5.3865</strong></td>
<td><strong>0.0000</strong></td>
<td><strong>1</strong></td>
</tr>
<tr>
<td>102</td>
<td>13.0421</td>
<td>0.0000</td>
<td>1</td>
</tr>
<tr>
<td>105</td>
<td>23.7722</td>
<td>0.0000</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4-3: Case 1: $P(x_i \mid \mathcal{Y}_i)$ and the starting point of the abnormal stage for Gu-b3
<table>
<thead>
<tr>
<th>time</th>
<th>vib. in rms</th>
<th>$P(X_i = 1 \mid \mathcal{Z}_i)$</th>
<th>$P(X_i = 2 \mid \mathcal{Z}_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.5</td>
<td>2.5478</td>
<td>0.9422</td>
<td>0.0577</td>
</tr>
<tr>
<td>17.5</td>
<td>2.7402</td>
<td>0.8472</td>
<td>0.1527</td>
</tr>
<tr>
<td>30</td>
<td>2.857</td>
<td>0.7633</td>
<td>0.2366</td>
</tr>
<tr>
<td>40</td>
<td>3.0721</td>
<td>0.6356</td>
<td>0.3643</td>
</tr>
<tr>
<td>50</td>
<td>3.2599</td>
<td>0.4231</td>
<td>0.5768</td>
</tr>
<tr>
<td>61</td>
<td>3.342</td>
<td>0.2420</td>
<td>0.7579</td>
</tr>
<tr>
<td>74.5</td>
<td>3.7751</td>
<td>0.0257</td>
<td>0.9742</td>
</tr>
<tr>
<td>86</td>
<td>3.9638</td>
<td>0.0009</td>
<td>0.9990</td>
</tr>
<tr>
<td>94</td>
<td>4.0591</td>
<td>0.0000</td>
<td>0.9999</td>
</tr>
<tr>
<td><strong>106</strong></td>
<td><strong>4.3209</strong></td>
<td>0.0000</td>
<td>1</td>
</tr>
<tr>
<td>. . . . .</td>
<td>. . . . .</td>
<td>. . . .</td>
<td>. . . .</td>
</tr>
<tr>
<td>. . . . .</td>
<td>. . . . .</td>
<td>. . . .</td>
<td>. . . .</td>
</tr>
<tr>
<td>. . . . .</td>
<td>. . . . .</td>
<td>. . . .</td>
<td>. . . .</td>
</tr>
<tr>
<td>269.5</td>
<td>13.9153</td>
<td>0.0000</td>
<td>1</td>
</tr>
<tr>
<td>280.5</td>
<td>19.6412</td>
<td>0.0000</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4-4: Case 2: $P(x_i \mid \mathcal{Z}_i)$ and the starting point of the abnormal stage for Gu-b6

The figures show different paths of bearing failures but share the same properties in which they stay flat at an early stage and then increase rapidly before a failure. This justifies the two-stage failure process defined by (Christer et al., 1984). Determining the changing point is very important for maintenance personnel to enable suitable maintenance decisions to be made. For case 1, the defective state is picked up with probability 1 at time = 85 with vibrations reading 5.3865 in rms; in case 2 the defective state is detected at time = 106 with vibrations reading 4.3209 in rms. According to the study using SPC with the same dataset, Zhang (2004) recommended that a vibration level around 5 in rms seems to be a reasonable point at which to say that the bearings
are in a defective state. This suggests that our model performed well, reaching nearly the same conclusion. The full comparison of results using the SPC chart method and the model developed is shown in Section 4.4.

4.3 Testing the Model

One way to check the model’s fit is by simulating and generating a new dataset \( \tilde{y} \) and comparing it with the observed data \( y \), as a graphical comparison, which is usually a good initial indicator. The simulation of \( \tilde{y} \) is carried out by substituting the estimated parameters from the observed data into the model \( P(x_i | \mathcal{Z}_j) \), which is used to identify the initialisation of stage 2 and the duration of time \( l_1 \). Since the failure is also observed from the data, we can calculate the time \( l_2 \). Knowing \( l_1, l_2 \) and estimated parameters in the model, we could run the simulation that generates the simulated \( \tilde{y} \). See Figure 3.5 for the algorithm. If the matching between the observed \( y \) and simulated \( \tilde{y} \) is good, then we may say that our model is reasonable to explain the hidden process. Thus, the Pearson product moment correlation coefficient was calculated, to assess whether there was a relationship between these two variables and to see how well the model explained the observed data. Here, we show two plots from our test datasets, which attempt to compare the actual vibration from observation and simulated vibration based on estimated parameters.

![Figure 4-1: Case 1: Simulated and observed vibration levels for Gu-b3](image)
From these graphical representations, (Figure 4-1 and Figure 4-2) it is sufficient to say that the model fits well with the observed data. To see the significance level of this relationship, we set a value for the confidence level, normally 95% or 99% (Albright et al., 1999). The result of this comparison is shown in Table 4-5 below.

![Graph showing simulated and observed vibration levels for Gu-b4](image)

**Figure 4-2: Case 2: Simulated and observed vibration levels for Gu-b4**

<table>
<thead>
<tr>
<th>df (n-2)</th>
<th>Gu-b3</th>
<th>Gu-b4</th>
</tr>
</thead>
<tbody>
<tr>
<td>R²</td>
<td>0.958601</td>
<td>0.823308</td>
</tr>
</tbody>
</table>

| 95% Critical Values – Pearson Product Moment Correlation Coefficient | 0.666 | 0.423 |

**Table 4-5: Values for Pearson product moment correlation coefficient**

The null hypothesis for the Pearson product moment correlation coefficient is \( r = 0 \), indicating no correlation within variables. In the above cases, the critical value at 0.05 confidence level and \( n - 2 \) degrees of freedom is greater than those from the table; therefore the null hypothesis is rejected, which implies that there is a correlation between the observed and fitted data. Furthermore, we could see from Table 4-5 that the observed data for Gu-b3 and Gu-b4 represents 95% and 82% close to the model developed. With this high percentage of correlations, we could say that the model fits the data well.
The other test to determine how well the model actually reflects the data is by running a goodness-of-fit test. One statistical test that addresses this issue is the chi-square, $\chi^2$, goodness-of-fit test, which determines whether there is a significant difference between the sample data and the population distribution. We first formulated the conservative hypothesis, called the null hypothesis ($H_0$), which states that there is no difference between the sample and the population distribution. The test procedure begins by arranging the $n$ observations into a set of $k$ class intervals or cells. The test statistic is given by

$$
\chi^2 = \sum_{i=1}^{k} \left( \frac{(N_i - np_i)^2}{np_i} \right)
$$

(4-1)

where $N_i$ is the number of observations included in category $i$, $p_i$ is the probability that the random variable $X$ of the population falls into category $i$ and $np_i$ is the expected value of category $i$. The $\chi^2$ statistic is then compared with the critical value of $\chi^2$ distribution with the degree of freedom $k-1$. If the critical value of $\chi^2$ distribution is greater than the $\chi^2$ statistic, we can say that the observed data follows the theoretical distribution.

However, we cannot do this with this type of data as at each monitoring point we had only one observation. As an alternative, to divide the categories of data we could choose intervals with equal probability (Law and Kelton, 2000). To carry out the chi-square test, we could transform $p(y_i | \mathcal{X}_{i-1})$ into an equal probability, $\frac{1}{M}$, which will follow a uniform distribution; see Figure 4-3 below.
Thus, we partition \( p(y_i | \mathcal{Z}_{i-1}) \) into \( M \) cells so that each cell has an equal probability, and then we examine which cell each observation fell in and count the number of observations in each cell. Hence, we can test whether the newly transformed data follows the uniform distribution. As an example, Zhang (2004) specified that the intervals should have equal probability measures, when testing the conditional residual time (CRT) model developed for oil analysis. The difficulty with carrying out a chi-square test is that there is no ideal recommendation for choosing the number and size of the intervals, \( k \), which is guaranteed to produce good result in terms of validity (Law and Kelton, 2000). However, a few guidelines can be followed to ensure the validity of the test, such as the one given by Law and Kelton (2000) where for equal probability intervals, the validity condition can only be satisfied if \( k \geq 3 \) and \( N_i \geq 5 \) for all \( i \).

Applying all the considerations to satisfy the validity of the test, we choose the number of intervals \( = 4 \) with the size of equal probability \( 0.25 \). The computations of this test are given in Table 4-6 below, using two datasets (Gu-b3 and Gub4), which have 31 observation points.
Table 4-6: Calculating the goodness-of-fit test

<table>
<thead>
<tr>
<th>Number of Interval</th>
<th>( N_i )</th>
<th>( p_i )</th>
<th>( Np_i )</th>
<th>( (N_i - Np_i)^2 )</th>
<th>( \frac{(N_i - Np_i)^2}{Np_i} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13</td>
<td>0.25</td>
<td>7.75</td>
<td>5.25</td>
<td>3.556452</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.25</td>
<td>7.75</td>
<td>-2.75</td>
<td>0.975806</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.25</td>
<td>7.75</td>
<td>-2.75</td>
<td>0.975806</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0.25</td>
<td>7.75</td>
<td>0.25</td>
<td>0.008065</td>
</tr>
<tr>
<td>Total</td>
<td>31</td>
<td>1</td>
<td></td>
<td></td>
<td><strong>5.516129</strong></td>
</tr>
</tbody>
</table>

The value of the test statistic is \( \chi^2 = 5.5161 \). Referring to the chi-square table, we see that \( \chi^2_{3,0.05} = 7.815 \), which is not exceeded by the critical value of \( \chi^2 \), so we would not reject the null hypothesis at \( \alpha = 0.05 \). Thus, this test gives us a reason to conclude that the sample is from the population distribution.

The next stage was to use the data to calculate the expected time to stage 2 and calculate the expected time to failure, based upon the model developed in Chapter 3. We first model the problem with the expected time to the initial point of the second stage that is equation (3-36). The numerical results are shown in Table 4-7 below.

Table 4-7: \( E(L_t - t_i, x_i \mid L_t > t_i, \mathcal{F}_t) \) with exponential distribution for Gu-b3

<table>
<thead>
<tr>
<th>time</th>
<th>vibr</th>
<th>( P(X_i = 1 \mid \mathcal{F}_t) )</th>
<th>( \int_{t_i}^{\infty} (l_i - t_i) f_i(l_i) dl_i )</th>
<th>( \int_{t_i}^{\infty} f_i(l_i) dl_i )</th>
<th>( E(L_t - t_i, x_i \mid L_t &gt; t_i, \mathcal{F}_t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>1.6012</td>
<td>0.953525</td>
<td>112.359</td>
<td>107.14</td>
<td></td>
</tr>
<tr>
<td>25.5</td>
<td>1.7306</td>
<td>0.944808</td>
<td>112.359</td>
<td>106.16</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>1.8917</td>
<td>0.968074</td>
<td>112.359</td>
<td>108.77</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>2.4231</td>
<td>0.958727</td>
<td>112.360</td>
<td>107.72</td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>2.5086</td>
<td>0.974952</td>
<td>112.360</td>
<td>109.55</td>
<td></td>
</tr>
<tr>
<td>74</td>
<td>3.4853</td>
<td>0.664153</td>
<td>112.360</td>
<td>74.62</td>
<td></td>
</tr>
<tr>
<td><strong>85</strong></td>
<td><strong>5.3865</strong></td>
<td><strong>5.14E-08</strong></td>
<td><strong>112.360</strong></td>
<td><strong>0.00</strong></td>
<td></td>
</tr>
<tr>
<td>102</td>
<td>13.0421</td>
<td>9.04E-19</td>
<td>112.360</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>105</td>
<td>23.7722</td>
<td>9.07E-19</td>
<td>112.360</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

From the table, the conditional expected value of \( L_t \) given \( L_t > t_i \) is constant at every monitoring check. This is because \( f_i(l_i) \) is assumed to be an exponential distribution, which has the memoryless property indicates that the remaining lifetime of a component is independent of its current age. This situation can be resolved if a different distribution
is used to replace the exponential distribution $f_i(l_i)$. As an example, we chose the Weibull distribution with known $\alpha$ and $\beta$, and re-calculated the conditional expected value of $l_i$ given $l_i > t_i$. This result is given in Table 4-8 below:

<table>
<thead>
<tr>
<th>time</th>
<th>vibr</th>
<th>$P(X_i = 1 \mid \overline{X_i})$</th>
<th>$\int_{t_i}^{\infty} (l_i - t_i) f_i(l_i) dl_i$</th>
<th>$\int_{t_i}^{\infty} f_i(l_i) dl_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>1.6012</td>
<td>0.953525</td>
<td>72.665</td>
<td>69.29</td>
</tr>
<tr>
<td>25.5</td>
<td>1.7306</td>
<td>0.944808</td>
<td>56.552</td>
<td>53.43</td>
</tr>
<tr>
<td>35</td>
<td>1.8917</td>
<td>0.968074</td>
<td>47.895</td>
<td>46.37</td>
</tr>
<tr>
<td>50</td>
<td>2.4231</td>
<td>0.958727</td>
<td>35.825</td>
<td>34.35</td>
</tr>
<tr>
<td>57</td>
<td>2.5086</td>
<td>0.974952</td>
<td>30.996</td>
<td>30.22</td>
</tr>
<tr>
<td>74</td>
<td>3.4853</td>
<td>0.664153</td>
<td>21.481</td>
<td>14.27</td>
</tr>
<tr>
<td>85</td>
<td>5.3865</td>
<td>5.14E-08</td>
<td>16.861</td>
<td>0.00</td>
</tr>
<tr>
<td>102</td>
<td>13.0421</td>
<td>9.04E-19</td>
<td>11.638</td>
<td>0.00</td>
</tr>
<tr>
<td>105</td>
<td>23.7722</td>
<td>9.07E-19</td>
<td>10.915</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 4-8: $E(L_4 - t_i, x_i \mid L_4 > t_i, \overline{X_i})$ with Weibull distribution for Gu-b3

Next, we calculated the expected time to failure as stated in equation (3-37). A similar problem also occurs if we use an exponential distribution to calculate the conditional expected value for time $L_4$ and $L_2$, given $L_4, L_2 > t_i$ as shown in Table 4-9:

<table>
<thead>
<tr>
<th>$\int_{l_i}^{\infty} f_i(l_i) \int_{l_2 + t_i}^{\infty} (l_2 + l_i - t_i) f_2(l_2) dl_2 dl_i$</th>
<th>$\int_{l_i}^{\infty} f_i(l_i) \int_{l_1}^{\infty} (l_2 + l_i - t_i) f_2(l_2) dl_2 dl_i$</th>
<th>$\int_{l_i}^{\infty} f_i(l_i) \int_{l_1}^{\infty} f_2(l_2) dl_2 dl_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>219.891</td>
<td>107.527</td>
<td>214.67</td>
</tr>
<tr>
<td>219.891</td>
<td>107.527</td>
<td>213.69</td>
</tr>
<tr>
<td>219.891</td>
<td>107.527</td>
<td>216.30</td>
</tr>
<tr>
<td>219.893</td>
<td>107.527</td>
<td>215.26</td>
</tr>
<tr>
<td>219.893</td>
<td>107.527</td>
<td>217.08</td>
</tr>
<tr>
<td>219.893</td>
<td>107.527</td>
<td>182.16</td>
</tr>
<tr>
<td>219.893</td>
<td>107.527</td>
<td>107.53</td>
</tr>
<tr>
<td>219.893</td>
<td>107.527</td>
<td>107.53</td>
</tr>
<tr>
<td>219.893</td>
<td>107.527</td>
<td>107.53</td>
</tr>
</tbody>
</table>

Table 4-9: $E(L_2 + L_4 - t_i, x_i \mid L_2 + L_4 > t_i, \overline{X_i})$ with exponential distribution for Gu-b3

Thus, if we choose others distributions, e.g. Weibull, with known $\alpha$ and $\beta$, we could obtain different results such as shown in Table 4-10 below:
To summarize, we have shown how the expected time to first reach stage 2 or fault and failure can be calculated. Using exponential distributions for the two-stage failure process would yield a constant rate of the condition expected, which is inaccurate. Hence, to show an example to work this out, we choose a Weibull distribution with given parameters suitable for the two-stage failure process. For the case of the initialisation of random defects, Table 4-8 shows at first checking time = 9, the expected time to first reach stage 2 is 69.29; and Table 4-10 shows the expected time to first reach failure stage is 147.34. This process is continued until time = 85, where the expected time to first reach stage 2 is 0 (the second stage is reached), and the expected time to first reach failure stage is 59.28. Using this analysis it could aid maintenance managers to plan and schedule their maintenance actions.

### 4.4 Comparison with a Statistical Process Control (SPC) Chart Approach

The Statistical Process Control (SPC) has been formally defined as a methodology for monitoring a process to identify special causes of variations that signal the need to take corrective actions when it is appropriate (Evans and Lindsay, 2005). In the case of condition-based maintenance, the SPC model was developed to monitor the process of observing variables before any maintenance action can be taken. This section aims to summarize a study by Zhang (2004), which applied the SPC chart approach (Wetherill and Brown, 1991) to the same dataset we used here. In her work, Zhang proposed three
SPC chart models, which are useful for identifying the initial point of a random defect. The models are:

3. Shewhart average level
4. Adaptive moving average and moving range charts
5. Adaptive Shewhart average level

Without any discussion of those methods cited above, a comparison will be made here of the Zhang’s results with those obtained from our state prediction model, as the same data is used to achieve the same objectives. The results are shown in Table 4-11 below.

<table>
<thead>
<tr>
<th>Techniques used</th>
<th>Gu-b1</th>
<th>Gu-b2</th>
<th>Gu-b3</th>
<th>Gu-b4</th>
<th>Gu-b5</th>
<th>Gu-b6</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPC – Shewhart average level</td>
<td>6.6828</td>
<td>5.8982</td>
<td>5.3865</td>
<td>5.3062</td>
<td>5.1302</td>
<td>8.3994</td>
</tr>
<tr>
<td>SPC – Adaptive moving average</td>
<td>6.6828</td>
<td>5.4781</td>
<td>5.3865</td>
<td>3.7751</td>
<td>5.1302</td>
<td>2.6598</td>
</tr>
<tr>
<td>SPC – Adaptive moving range</td>
<td>6.6828</td>
<td>8.2242</td>
<td>5.3865</td>
<td>6.3926</td>
<td>5.1302</td>
<td>3.8376</td>
</tr>
<tr>
<td>SPC – Adaptive Shewhart average level</td>
<td>4.0744</td>
<td>3.8286</td>
<td>5.3865</td>
<td>3.9638</td>
<td>5.1302</td>
<td>8.3994</td>
</tr>
<tr>
<td>State prediction model</td>
<td>6.6828</td>
<td>4.6055</td>
<td>5.3865</td>
<td>4.0591</td>
<td>6.4194</td>
<td>8.3994</td>
</tr>
</tbody>
</table>

Table 4-11: Starting point of the abnormal stage using SPC techniques and state prediction model

The results in Table 4-11 shows the starting point abnormal of vibration readings, which indicate that, the system may be in an abnormal state. It is clearly shown that the state prediction model produce almost the same result as the SPC models discussed by Zhang (2004). Briefly, using the Shewhart average level to detect a fault, we need to have a threshold level indicating a warning limit, which can be established from experience by an expert, or from a manufacturer. The adaptive moving average and adaptive moving range models also need a threshold value and averaging the process will increase the delays in responding to sudden jumps. The adaptive Shewhart average level model performs well without a threshold value but is not suitable with one-at-a-time data unless several such data can be grouped together using a predetermined interval size. Again, by averaging the process, the delays in responding to a sudden jump will depend on the average difference of interval. See Zhang (2004) for details of
the techniques used. In contrast to the SPC models, the model that we developed, which we called a state prediction model, does not set any threshold values and does not use the averaging concept. In fact, it uses the weight of each state to predict the future state, which is more appropriate.

From this comparison of the techniques used, the advantages of state prediction can be summarized as:

1. Useful as a complete decision model for CRT, to determine the initial point of default and predict the residual time of the item monitored. A previous study had used SPC as a technique to identify random faults before it was possible to use the CRT model.
2. Can be extended to predict any state defined in a system, while SPC is limited to two states.
3. Given the structure of the state prediction model, maintenance actions could be carried out directly according to the state.
4. SPC cannot provide the probability prediction of state one, which is needed for other modelling, as explained in Section 4.3.

To summarize, we could say that the state prediction model enriches the range of existing techniques for fault identification.

4.5 Summary

This chapter has given an account of the fitting of observed vibration data obtained from a laboratory experiment to our model proposed in Chapter 3. The model was formulated to identify the initiation of a random defect from vibration data before it causes a failure. The rationale behind this idea is that once the defect is identified, necessary maintenance could be carried out or other modelling tasks could be performed. Numerical results were shown regarding the parameter estimation calculated from the observed data and the capability of our model to identify random faults. Two statistical tests were carried out to test the goodness-of-fit of the model to data, both giving convincing results. Finally, a comparison of the results with those of a past study was presented and the benefits of the model are pointed out.
5 CHAPTER 5: NUMERICAL APPROXIMATION TO FAULT PREDICTION TECHNIQUES

5.1 Introduction

Determining the underlying state of the system for maintenance management is crucial, as discussed in Chapter 3. It has been demonstrated how this problem can be solved using the concept of a Hidden Markov Model. As an extension, this chapter will explore approximation techniques that can be used to solve the computing problems associated with previous work, that calculating the $P(x_t | \mathcal{Z}_t)$ is complex and time-consuming particularly that $x_t$ is continuous. Thus, it is desirable to find an approximation model that will enable us to calculate a recursive filtering approach $P(x_t | \mathcal{Z}_t)$ more efficiently. The literature, (Arulampalam et al., 2001) shows that predicting the hidden state of a system from noisy and partial observations, usually in a Markov context, is known as Bayesian filtering. It has been suggested that a number of different types of approximation might be used to address this problem (Arulampalam et al., 2001; Doucet et al., 2000). To approximate $P(x_t | \mathcal{Z}_t)$, we investigate two of the techniques discussed, namely:

1. a grid-based method.
2. particle filtering.

In this section, we present a brief review of the theory for each approach and then show how it can be used for solving the random fault identification problem. We demonstrate its application with the vibration monitoring data used in Chapter 3.

5.2 Grid-based Method

The purpose of the classic filtering approach is to compute the joint posterior distribution of all states given in all the monitoring observations. In details, Sarkka (2006) noted that the filtering model consists of
1. an initial distribution that specifies the prior distribution of \( p(x_0) \), that is the hidden state, \( x_0 \) at initial time \( t_i = 0 \).

2. a dynamic model that models the dynamics of the model and its uncertainties as a Markov sequence. Here the model is described by transition probabilities \( p(x_i | x_{i-1}) \).

3. a measurement model that models the relationship of random observation \( y_i \) where \( y_i \) is its realization, is dependent on the random state, \( x_i \) where \( x_i \) is its realization. This relationship is modelled by specifying the distribution of the observation given state \( x_i \), \( p(y_i | x_i) \).

Hence, the joint posterior distribution of \( x_0, x_1, \ldots, x_t \) given the condition monitoring history has the following form:

\[
p(x_0, x_1, \ldots, x_t | y_0, y_1, \ldots, y_t) = p(x_0 | y_0) \prod_{j=1}^{t} p(x_j | y_j) p(x_j | x_{j-1})
\]

(5-1)

However, computing the joint posterior distribution in such a way is computationally inefficient and unnecessary in practice. An optimal filtering marginalized the distribution of the current \( x_t \) given the history of monitoring information which implies an estimation of state at time \( t_i \) from all monitoring information, \( \mathcal{Z}_t = \{y_i, y_{i-1}, \ldots, y_1\} \), up to time \( t_i \). This allows the filtering distributions to be computed recursively by the Bayesian filtering equations, which have the following form:

1. Predicting step which uses the Chapman-Kolmogorov equation for computing the step-ahead prediction distribution of the state \( x_i \) given \( \mathcal{Z}_{i-1} \).

2. Updating step, which uses the Bayes’ rule for computing the posterior distribution of state \( x_i \) given the current measurement \( y_i \).
Therefore, similar recursive filtering is obtainable as shown in equation (3-1—3-3) in Chapter 3. Arulampalam et al. (2001) noted that a grid-based method would give the optimal recursion of $P(x_i \mid \mathcal{Z}_i)$ with the assumption that the state space is discrete and consists of a finite number of states. However, in the continuous case, the grid-based method could still be applied by partition of the space into small intervals, which needs more computation. This approach is similar to the HMM method introduced in Chapter 3, except that this approach introduces a concept of weight for the conditional probability of a state, given a measurement.

To demonstrate how a grid-based algorithm works in a discrete case, a recursive formula for obtaining $P(x_i \mid \mathcal{Z}_i)$ from $P(x_{i-1} \mid \mathcal{Z}_{i-1})$ is given by:

$$P(x_i^k \mid \mathcal{Z}_i) = \frac{p(y_i \mid x_i^k) \sum_{k=1}^{N} P(x_i^k \mid x_{i-1}^k) P(x_{i-1}^k \mid \mathcal{Z}_{i-1})}{\sum_{k=1}^{N} p(y_i \mid x_i^k) \sum_{k=1}^{N} P(x_i^k \mid x_{i-1}^k) P(x_{i-1}^k \mid \mathcal{Z}_{i-1})} \quad (5-2)$$

Suppose that for the state at time $t_{i-1}$, we have $x_{i-1}^k$, where $k = 1,...,N$ defines the number of states involved. For each state $x_{i-1}^k$, let the conditional probability of that state given measurements up to time $t_{i-1}$ be denoted by $w_{i-1}^k$, shown as $P(X_{i-1} = x_{i-1}^k \mid \mathcal{Z}_{i-1}) = w_{i-1}^k$. This implies the probability that the system is at state $k$ at time $t_{i-1}$ given all observations up to time $t_{i-1}$. Then, the posterior density at time $t_{i-1}$ can be written as:

$$P(x_{i-1} \mid \mathcal{Z}_{i-1}) = \sum_{k=1}^{N} w_{i-1}^k \delta(x_{i-1} - x_{i-1}^k) \quad (5-3)$$

where $\delta(\bullet)$ is the Dirac delta measure, defined by the following property:
\[ \delta(x_{i-1} - x_i^k) = \begin{cases} 1 & x_{i-1} = x_i^k \\ 0 & x_{i-1} \neq x_i^k \end{cases} \]  

(5-4)

At current checking time, \( t_i \), the posterior density is re-weighted using the state transition probabilities \( P(x_i^k \mid x_{i-1}^j) \). Hence the prediction step yields:

\[
P(x_i \mid \mathcal{Z}_{i-1}) = \sum_{k=1}^{N} w_{j,i-1}^k \delta(x_i - x_i^k) \]

where

\[
w_{j,i-1}^k \approx \sum_{j=1}^{N} w_{j,i-1}^j P(x_i^k \mid x_{i-1}^j) \]

(5-5)

To obtain the new posterior, \( P(x_i \mid \mathcal{Z}_i) \), the prior weight is updated using the monitored information and then normalized, which can be written as

\[
P(x_i \mid \mathcal{Z}_i) = \sum_{k=1}^{N} w_{j,i}^k \delta(x_i - x_i^k) \]

where

\[
w_{j,i}^k \approx \frac{w_{j,i-1}^k P(y_i \mid x_i^k)}{\sum_{j=1}^{N} w_{j,i-1}^j P(y_i \mid x_i^j)} \]

(5-6)

A trellis diagram can be used to visualize the grid-based method, as shown in Figure 5-1 below.
Each column in the trellis shows the possible states of the machine at a certain time $t_i$.

From Figure 5-1, each state in one column is connected to each state in the adjacent column by the transition probability. At every monitoring time $t_i$, the posterior density is predicted and then updated.

Applying this approach to our problem, it is necessary to understand the two key elements on which the method is used. First, we need to formulate the evolution of the state with time, $P(x_i^t \mid x_{i-1}^t)$ and then we need a formula to describe the relationship between the state and the condition monitoring information, $p(y_i \mid x_i^t)$. These formulations for $P(x_i^t \mid x_{i-1}^t)$ and $p(y_i \mid x_i^t)$ have already been discussed in Chapter 3, and we now insert them into our grid-based approximation model. The numerical result of this approach is shown in Section 5.3.
### 5.3 Particle Filter for a Discrete State Case

Another approach which can be interpreted as an extension of the grid-based filter is the sequential Monte Carlo method, also referred to as particle filtering (Doucet et al., 2000; Chen et al., 2004). The key idea of this approach is to represent the required posterior density function by a set of random samples with associated weights and compute the approximation based on these samples and weights. Therefore, the particle filter method is more dynamic than the grid-based approach; the particles are assumed to be uniformly distributed over the space in the former, rather than fixed, as in the grid-based method. Particle filters have developed rapidly in recent years and have been successfully applied in a number of different areas such as polymers, robotics, physics and engineering (Djuric et al., 2003), and in a large number of applications (Sarkka, 2006). However, there is no evidence that they have been used in the area of maintenance or reliability, which motivates us to carry out this research.

#### 5.3.1 Overview of the Particle Filtering Method

The explanation of particle filters that follows is based on a published tutorial (Arulampalam et al., 2001). The key element of particle filters is to approximate the posterior distribution $P(x_i | \mathcal{Z}_i)$ using a set of random samples/particles specified by

$$\mathcal{X}_i = \{x_i^k, w_i^k\}_{k=1}^N,$$

where $N$ is the number of particles denoted by $\{x_i^k\}_{k=1}^N$ and $\{w_i^k\}_{k=1}^N$ are the associated weights, in which the weights are then normalized such that $\sum_{k=1}^N w_i^k = 1$. Using a similar presentation from equation (5-3), the posterior distribution for the particle filter approach at any time $t_i$ can be approximated as:

$$P(x_i | \mathcal{Z}_i) \approx \sum_{k=1}^N w_i^k \delta(x_i - x_i^k) \quad (5-7)$$
where \( w_i^k \) is the weight for the \( k^{th} \) state and \( \delta(\bullet) \) is the Dirac delta function. The key element now is how we can rationalize the weight discussed above, which leads us to the concept of importance sampling.

Suppose we want to generate \( N \) particles from a distribution \( P(x_i \mid \mathcal{Z}_i) \) in which each random particle will be assigned an equal weight, \( 1/N \). However, in reality the value of distribution \( P(x_i \mid \mathcal{Z}_i) \) is not available, but it is possible to generate particles \( \{x_i^k\}_k \) from a tractable distribution \( Q(x_i \mid \mathcal{Z}_i) \), which is proportional to \( P(x_i \mid \mathcal{Z}_i) \) and is called importance density. Given the distribution \( Q(x_i \mid \mathcal{Z}_i) \), we assign the proportional weight as:

\[
w_i^k \propto \frac{P(x_i^k \mid \mathcal{Z}_i)}{Q(x_i^k \mid \mathcal{Z}_i)} \tag{5-8}
\]

Then the normalized weight of the \( k^{th} \) is defined by

\[
w_i^k = \frac{w_i^k}{\sum_{k=1}^{M} w_i^k} \tag{5-9}
\]

Now, let us assume that the posterior distribution \( P(x_{i-1} \mid \mathcal{Z}_{i-1}) \) is approximated by the discrete random measures \( \{x_i^{k_{i-1}}, w_i^{k_{i-1}}\}_k \). Given the particles at \( P(x_{i-1} \mid \mathcal{Z}_{i-1}) \) and a new observation \( y_i \), the objective is to exploit \( \{x_i^{k_{i-1}}, w_i^{k_{i-1}}\}_k \) in obtaining \( \{x_i^k, w_i^k\}_k \). To achieve this, we use a sequential importance-sampling (SIS) algorithm, which generates the particles \( \{x_i^k\} \) and the weight \( \{w_i^k\} \) at every time \( t_i \) using the following formulation

\[
w_i^k \propto w_i^{k_{i-1}} \frac{p(y_i \mid x_i^k)P(x_i^k \mid x_{i-1}^k)}{Q(x_i^k \mid x_{i-1}^k, y_i)} \tag{5-10}
\]
See Arulampalam et al. (2001) for how this formulation was established. Generally, the SIS algorithm is implemented by performing the following two steps for every time \( t_i \):

1. Draw particles \( x_i^k \sim Q(x_i \mid x_{i-1}^k, y_i) \) where \( k = 1, 2, \ldots \)

2. Compute the weight of \( w_i^k \) according to equation (5-10), and then normalize them according to equation (5-9).

Hence, it is shown that the importance function plays an important role in the operation of the particle filter approach. Needless to say, the closer the proposed distribution is to the distribution, the better the approximation will be. Arulampalam et al. (2001), Doucet et al. (2000) and Djuric et al. (2003) reported that there are two most frequently used importance densities, namely prior importance and optimal importance densities. In the case of the former, it is convenient to choose the importance density as:

\[
Q(x_i \mid x_{i-1}^k, y_i) = P(x_i \mid x_{i-1}^k)
\]  

(5-11)

Then, by substituting equation (5-11) into equation (5-10), the update equation can be shown to be

\[
w_i^k \propto w_{i-1}^k P(y_i \mid x_i^k)
\]  

(5-12)

which can easily be calculated.

The optimal importance density minimizes the variance of the importance weight condition on \( x_{i-1}^k \) and \( y_i \) is shown in Doucet et al. (2000) as

\[
Q(x_i \mid x_{i-1}^k, y_i)_{opt} = P(x_i^k \mid x_{i-1}^k, y_i)
\]  

(5-13)
When substituting equation (5-13) into equation (5-10) the update equation can be shown as

\[ w_i^k \propto w_{i-1}^k p(y_i \mid x_{i-1}^k) \]
\[ = w_{i-1}^k \int_0^\infty p(y_i \mid x_i) P(x_i \mid x_{i-1}^k) dx_i \]  \hspace{1cm} (5-14)

Applying this in our discrete case, equation (5-14) can be re-written as

\[ w_i^k \propto w_{i-1}^k \sum_{k=1}^{N} p(y_i \mid x_i^k) P(x_i^k \mid x_{i-1}^k) \]  \hspace{1cm} (5-15)

Note that the implementation of particle filters with prior importance densities is much easier than for those with optimal importance densities. This is because the computation of \( p(y_i \mid x_{i-1}^k) \) requires more time to solve the integration. Prior importance density, on the other hand, is independent of the current monitoring data, which means that the state estimation is explored without the knowledge of the observations, hence making the SIS filter less responsive.

The next implementation issue that needs attention is that the distribution of the importance density may exhibit unwanted phenomena, where after a few iterations, all the particles except for a very few are assigned negligible weights. This phenomenon, called degeneracy, implies that the performance of the particle filter will deteriorate. One suitable measure of the degeneracy of the algorithm is the introduction of the effective sample size \( N_{eff} \) defined as

\[ N_{eff} = \frac{1}{\sum_{k=1}^{N} \left( w_i^k \right)^2} \]  \hspace{1cm} (5-16)
where $w_i^k$ is the normalized weight. Whenever $N_{eff}$ is below a predefined threshold $N_T$, it indicates severe degeneracy and to reduce it resampling is performed. Resampling is an idea that eliminates particles with small weights and concentrates on particles with large weights. In principle, it is implemented as follows:

- Generate a new set of particles $\{x_i^l\}$ by resampling from the previous set $\{x_i^k\}$ with probabilities $P(x_i^l = x_i^k) = w_i^k$
- Reset the weight $w_i' = \frac{1}{N}$ to the particles

Next, we proceed to showing how these steps are applied to resolve the problem of identifying the initial point of a random defect from vibrations of bearing data.

### 5.3.2 Implementation of Sequential Importance Sampling (SIS) for a Discrete Case

In this case study, both of the methods explained above are applied to our state prediction model. Note that $P(x_i \mid x_{i-1}^k)$ and $p(y_i \mid x_i^k)$ used in both algorithms follow the equations (3-8 – 3-16) and (3-17) respectively. First, we present an algorithm for the implementation of SIS (sub-optimal) to our problem.

#### Algorithm 1: SIS (prior-importance)

For time steps $i = 0,1,2,\ldots$ do the following:

**Step 1:** Start with the non-parametric prior distribution of the state $x_0 \sim \{x_0^1, x_0^2\}$, $w_0 \sim \{w_0^1, w_0^2\}$ represented by a set of 2 discrete states.

**Step 2:** Generate $x_i^k \sim P(x_i \mid x_{i-1}^k)$, hence $x_i \sim \{x_i^1, x_i^2\}$.

**Step 3:** Assign the particle weight for each state, $w_i^k$, where $w_i^k = w_{i-1}^k p(y_i \mid x_i^k)$.

**Step 4:** Normalize the importance weights, $w_i = \frac{w_i^k}{\sum_{k=1}^{N} w_i^k}$. 

Step 5: Calculate $P(x_i \mid \mathcal{Z}_i) \approx \sum_{k=1}^{N} w_i^k \delta(x_i - x_i^k)$.

Step 6: Repeat step 2 for $i = 1,2,t_n$, where $t_n$ is the last monitoring point.

End

Algorithm 2: SIS (optimal importance)

For time steps $i = 0,1,2,\ldots$ do the following:

Step 1: Start with the non-parametric prior distribution of the state $x_0 \sim \{x_0^1,x_0^2\}$, $w_0 \sim \{w_0^1,w_0^2\}$ represented by a set of 2 discrete states.

Step 2: Generate $x_i^k \sim p(x_i \mid x_{i-1}^k)$ hence $x_i \sim \{x_i^1,x_i^2\}$.

Step 3: Assign the particle weight for each state, $w_i^k = w_{i-1}^k p(y_i \mid x_{i-1}^k)$, where $p(y_i \mid x_{i-1}^k) = \sum_{k=1}^{N} p(y_i \mid X_i = k)P(X_i = k \mid x_{i-1}^k)$. In the case of the initiation of a random defect, the transition matrix $P(x_i \mid x_{i-1}^k)$ used in Chapter 3, is normalized into 2 states.

Step 4: Normalize the importance weights, $w_i = \frac{w_i^k}{\sum_{k=1}^{N} w_i^k}$.

Step 5: Calculate $P(x_i \mid \mathcal{Z}_i) \approx \sum_{k=1}^{N} w_i^k \delta(x_i - x_i^k)$.

Step 6: Repeat step 2 for $i = 1,2,t_n$ where $t_n$ is the last monitoring point.

End

In the case of the initiation of a random defect, we have only two states, hence, degeneracy is not an issue; rather, it helps us to identify when the second stage starts. Thus, the resampling approach is not applicable here. The grid-based approach and the simple version of the SIS approach described above are conceptually straightforward to implement. However, because of the computational requirements of the method increased exponentially with the number of states, it can be very inefficient, except for problems with fewer states (Arulampalam et al., 2001; Doucet et al., 2000).
5.4 Parameter Estimation

It is desirable to apply the model developed above to the vibration data used in Chapter 4, with the main objective being to identify the starting point of a fault. However, the unknown parameters in $P(x_i \mid Z_i)$ need to be estimated. In the case of both the grid-based and particle filter approaches, the only information observed was $y_i, i = 1,2,\ldots,n$ where $n$ is the last observation. Given that we have the information that the system failed at time $t$, where $t > t_n$, we could apply the technique of the maximum likelihood estimator (MLE) by establishing the likelihood function discussed in equation (3-29) in Chapter 3.

Both techniques require an initial value for the weight of each state. Furthermore, we know that the item has been monitored from new, hence at time $t_0$, we are confident that the system is in a normal state. In the case of the grid-based approach, the initial weight for the normal is set as 1 and defect is set as 0; we carried out the parameter estimation using equation (3-29) and produced the results shown in Table 5-1 below.

<table>
<thead>
<tr>
<th>$\hat{\alpha}$</th>
<th>$\hat{\beta}$</th>
<th>$\hat{b}$</th>
<th>$\hat{\lambda}_1$</th>
<th>$\hat{\lambda}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4314</td>
<td>3.635</td>
<td>0.0983</td>
<td>0.0089</td>
<td>0.0093</td>
</tr>
</tbody>
</table>

Table 5-1: Estimated parameters using the grid-based approach

The result from the grid-based approach is similar to those obtained by HMM in Chapter 3. This is because the structure of the problem and its solution via the filtering approach are similar to HMM. However, as a comparison, by giving an initial weight for each state, these approximation approaches make the posterior density straightforward and allow a quick calculation.

In the case of SIS, the initial value is set to $w_0 = \{w_0^1, w_0^2\} = \{0.9,0.1\}$ to generate $N$ random samples of $\{x_0^1, x_0^2\}$. The formulation of $P(x_i \mid x_{i-1})$ established in Chapter 3 is based upon the 3-states formulation; hence to generate 2 states (normal and defect) we normalized the normal state and the defect state according to
\[ P(x_i^j | x_{i-1}^k) = \frac{P(x_i^j | x_{i-1}^k)}{\sum_{j=1}^{2} P(x_i^j | x_{i-1}^k)}, \quad j, k = 1, 2, \quad j > k \]  

(5-17)

Hence, the estimated parameter is obtained and shown in Table 5-2 and Table 5-3 below.

<table>
<thead>
<tr>
<th>( \hat{\alpha} )</th>
<th>( \hat{\beta} )</th>
<th>( \hat{b} )</th>
<th>( \hat{\lambda}_1 )</th>
<th>( \hat{\lambda}_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4469</td>
<td>3.7332</td>
<td>0.0614</td>
<td>0.0353</td>
<td>0.0129</td>
</tr>
</tbody>
</table>

Table 5-2: Estimated parameters using a particle filter with prior importance density

<table>
<thead>
<tr>
<th>( \hat{\alpha} )</th>
<th>( \hat{\beta} )</th>
<th>( \hat{b} )</th>
<th>( \hat{\lambda}_1 )</th>
<th>( \hat{\lambda}_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.47366</td>
<td>3.9990</td>
<td>0.0600</td>
<td>0.0427</td>
<td>0.0115</td>
</tr>
</tbody>
</table>

Table 5-3: Estimated parameters using a particle filter with optimal importance density

As a result, in the case of particle filtering, the estimated parameters from difference importance density produce nearly similar results. At this stage, it is very difficult for us to determine which set of estimates is better than the other until some technique is used to measure them. Hence, using the estimated parameters, we calculated the result of posterior distribution \( P(x_i | \mathcal{Z}_i) \) with the aim of identifying the starting point of a random defect. These results are shown in Figure 5-2 and Figure 5-3 below.

Figure 5-2: Identification of a random defect of bearing Gu-b3 using SIS algorithm (optimal density)
The results from the SIS approach with different importance sampling show that they produced results similar to the grid-based or HMM methods. It is concluded that there is no such importance sampling that is better than the others in terms of the result produced for $P(x_i \mid Z_t)$. However, if we consider an effort to calculate $P(x_i \mid Z_t)$, it is suggested that particle filtering with prior sampling density is the best option, due to its simplicity as explained in Section 5.2.2.

Since the case of identifying the initial point of a random fault consists of only two states, the degeneracy problem is not a great concern. It should be noted here that use of the particle filtering also has some drawbacks such as:

1. The particle filtering is computationally intensive, which implies that the number of particles used is proportional to the computational time. Hence, increasing the number of particles will need more time.

2. The re-sampling procedure tends to improve the efficiency of importance sampling, but the replication of particles with large weights would introduce correlation.
among those particles, whereas the particles should behave independently of each other.

3. A solution to the particle filter depends on the choice of the proposal distributions, and not knowing the underlying distribution $p(x_i^k | x_{i-1}^k, y_i)$ may make this problem more difficult.

### 5.5 Summary

In this chapter, approaches were sought to solve the computing problem of state estimation. Approximation techniques were considered, from which we identified the two most applicable approaches for a small number of states. Both grid-based and particle filter methods were applied to our problem and produced satisfactory results. Two problems that commonly arise while implementing particle filters were discussed. The degeneracy phenomenon was shown not to be a problem in our case; rather it helps us to find a solution. The second problem, determining the importance sampling, is just an option to calculate the current weight of each state, which in our case study shows little difference in the results between them. Thus, we conclude that the use of approximation techniques makes our model computationally feasible and easier to implement. However, it is noted that these approaches are only applicable if we have a small number of states.
CHAPTER 6: CONDITION BASED MAINTENANCE MODELLING BASED UPON OIL ANALYSIS DATA

6.1 Introduction

The analysis of oil samples taken from an engine gives an indication of the suitability of the oil for continued use and provides important information about the condition of the engine. This could allow the identification of wearing components before severe failure could occur without dismantling the engine. Given this condition-monitoring data, maintenance decisions may be taken as required and, most importantly, maintenance may be done in an effective and efficient way. This chapter describes a modelling development for maintenance decision-making using oil analysis data as the observed monitoring information. To be specific, Spectrometric Oil Analysis Programme (SOAP) is used as a condition-monitoring measure for marine diesel engines to determine the residual time of the engine at a specific monitoring point.

This chapter starts with some analysis, assumptions and techniques necessary to gain insight into SOAP data that will be useful to our modelling development. Several issues regarding the consistency, incompleteness and dimensions of the data used are discussed. Next, the modelling methodology is presented, adopting the concept of conditional residual time as discussed in Chapter 2. At the same time we proposed an approach called the ‘total metal concentrations’ calculation, which is used to explain the relationship between the observed metal concentrations data, measured by the oil analysis programme and the residual time of an engine.

The formula to find a conditional residual time distribution using a filtering technique is presented and a method for estimating the model parameters is also discussed. The model is then fitted to a set of ‘clean’ SOAP data to test the methodology. Once the distribution of the conditional residual time is known, we are able to establish a decision model to recommend the optimal maintenance decision-making. Model validation is also discussed at the end of the chapter.
6.2 Background

The Spectrometric Oil Analysis Programme (SOAP) is a technique for identifying the elemental composition of particles up to approximately 10 microns (Edwards et al., 1998) entrained in machinery oil samples. The rationale behind this technique is that as mechanical components wear, they shed small metallic particles that become entrained in the oil. Furthermore, particles over 10 microns are likely to exit the oil circulation via some filtration and this could lead the small particles which are less prone to the filter to remain suspended within the engine. If this measure is obtainable it could provide an indication of machine condition (Edwards et al., 1998). Wear metals such as iron (Fe), aluminium (Al), chromium (Cr), copper (Cu), tin (Sn), lead (Pb), silver (Ag), titanium (Ti) and nickel (Ni) are measurable, as well as lubricant additives such as calcium (Ca), barium (Ba), zinc (Zn), phosphorus (P), magnesium (Mg), boron (Be) and molybdenum (Mo). Other contaminants such as silicon (Si), sodium (Na) and potassium (K) are also detectable. By running periodic sampling and testing, SOAP enables the observation of trends in the metal concentrations of the engine oil. An increase in the concentration of particular elements can be used to identify the impending failure of a specific component. In this process, knowledge of the metal elements in a particular system is used to determine the likely origin of the wear particles. It is noted here that the differences in used oil and different engine models may cause a variation in the analysis results. Apart from identifying wear metal in oil analysis, SOAP also provides a physical property analysis that can be used to detect property changes in used oil. Changes in viscosity, water content, total based number and insoluble content may be indicative of faulty equipment or operating conditions.

6.3 Data Collection

At present, we have a set of SOAP data from diesel engines used in ships. The dataset consists of three types of data, as follows:

1. lifetime data (failure replacement and preventive replacement).
2. condition indicators (measured SOAP data).
3. maintenance events.
6.3.1 Lifetime Data

The lifetime data tells us the age in operating hours of the engine before it has been replaced. Such replacement may be carried out for two reasons. The first is a failure, in which the engine is replaced as it occurs; and the second refers to a preventive replacement in which the decision to replace the engine is carried out due to several considerations such as cost, scheduling and logistics aimed at the prevention of breakdowns or failures.

6.3.2 Condition Indicators

The condition indicators obtained from observed SOAP data reveal that there are 28 element indicators which can be broken down into three categories: lubricant condition, contamination and metal concentrations. Lubricant condition assesses whether the oil itself is fit for further service or is ready for a change. Assessment of contaminants measures the dirt, water, etc., which could degrade the oil. Metal concentrations measures several wear particles that become entrained in the oil due to component wear. At every check, the oil sample is analysed and all the elements quantified as parts per million (ppm). Generally, if the quantity of any element is higher than the tolerable level, maintenance actions may be performed, such as repair or replacement of a component, or topping-up or changing the oil.

6.3.3 Maintenance Events

The third dataset indicates how the maintenance event is performed at a specific time with some maintenance actions. Such events are changing lube oil, changing lube oil filters, changing filters, piston and cylinder inspection, etc.

The data for the SOAP analysis, from a third party company, was unsorted, contained missing values, was incomplete and inconsistent, and needed further explanations, as is common in practice (Ascher et al., 1995; Mathur et al., 2001). However, much effort has been put into understanding and sorting the data, to give us a ‘clean’ dataset that is appropriate for our model.
The data that we received was stored in a row-wise format, which needs to be transformed into a column-wise format for easy manipulation. The difference between these formats is illustrated in Figure 6-1 and Figure 6-2 below.

<table>
<thead>
<tr>
<th>SAMPLE</th>
<th>MACHINE</th>
<th>DATE</th>
<th>HOURS</th>
<th>OVERHAUL</th>
<th>ELEMENT</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>437</td>
<td>75001/2</td>
<td>31/1/1994</td>
<td>00:00:00</td>
<td>9829</td>
<td>1 Al</td>
<td>5</td>
</tr>
<tr>
<td>437</td>
<td>75001/2</td>
<td>31/1/1994</td>
<td>00:00:00</td>
<td>9829</td>
<td>1 Ba</td>
<td>0</td>
</tr>
<tr>
<td>437</td>
<td>75001/2</td>
<td>31/1/1994</td>
<td>00:00:00</td>
<td>9829</td>
<td>1 Cr</td>
<td>1</td>
</tr>
<tr>
<td>437</td>
<td>75001/2</td>
<td>31/1/1994</td>
<td>00:00:00</td>
<td>9829</td>
<td>1 Cu</td>
<td>0</td>
</tr>
<tr>
<td>437</td>
<td>75001/2</td>
<td>31/1/1994</td>
<td>00:00:00</td>
<td>9829</td>
<td>1 Fe</td>
<td>1</td>
</tr>
<tr>
<td>437</td>
<td>75001/2</td>
<td>31/1/1994</td>
<td>00:00:00</td>
<td>9829</td>
<td>1 Ni</td>
<td>0</td>
</tr>
<tr>
<td>437</td>
<td>75001/2</td>
<td>31/1/1994</td>
<td>00:00:00</td>
<td>9829</td>
<td>1 Pb</td>
<td>0</td>
</tr>
<tr>
<td>437</td>
<td>75001/2</td>
<td>31/1/1994</td>
<td>00:00:00</td>
<td>9829</td>
<td>1 Sb</td>
<td>0</td>
</tr>
<tr>
<td>437</td>
<td>75001/2</td>
<td>31/1/1994</td>
<td>00:00:00</td>
<td>9829</td>
<td>1 Sn</td>
<td>0</td>
</tr>
<tr>
<td>437</td>
<td>75001/2</td>
<td>31/1/1994</td>
<td>00:00:00</td>
<td>9829</td>
<td>1 Ti</td>
<td>0</td>
</tr>
<tr>
<td>437</td>
<td>75001/2</td>
<td>31/1/1994</td>
<td>00:00:00</td>
<td>9829</td>
<td>1 Ti</td>
<td>0</td>
</tr>
<tr>
<td>617</td>
<td>75001/1</td>
<td>18/2/1993</td>
<td>00:00:00</td>
<td>8936</td>
<td>1 Cu</td>
<td>1</td>
</tr>
<tr>
<td>617</td>
<td>75001/1</td>
<td>18/2/1993</td>
<td>00:00:00</td>
<td>8936</td>
<td>1 Fe</td>
<td>1</td>
</tr>
<tr>
<td>617</td>
<td>75001/1</td>
<td>18/2/1993</td>
<td>00:00:00</td>
<td>8936</td>
<td>1 Ni</td>
<td>0</td>
</tr>
<tr>
<td>617</td>
<td>75001/1</td>
<td>18/2/1993</td>
<td>00:00:00</td>
<td>8936</td>
<td>1 Pb</td>
<td>0</td>
</tr>
<tr>
<td>617</td>
<td>75001/1</td>
<td>18/2/1993</td>
<td>00:00:00</td>
<td>8936</td>
<td>1 Sb</td>
<td>0</td>
</tr>
<tr>
<td>617</td>
<td>75001/1</td>
<td>18/2/1993</td>
<td>00:00:00</td>
<td>8936</td>
<td>1 Sn</td>
<td>0</td>
</tr>
<tr>
<td>617</td>
<td>75001/1</td>
<td>18/2/1993</td>
<td>00:00:00</td>
<td>8936</td>
<td>1 Ti</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 6-1: Row format for monitoring data

<table>
<thead>
<tr>
<th>HOURS</th>
<th>OVERHAUL</th>
<th>Ag</th>
<th>Al</th>
<th>Ba</th>
<th>Ca</th>
<th>Mg</th>
<th>Fe</th>
<th>Al</th>
<th>Cr</th>
<th>Cu</th>
<th>Pb</th>
<th>Si</th>
<th>Sn</th>
<th>Ti</th>
<th>Zn</th>
</tr>
</thead>
<tbody>
<tr>
<td>18307</td>
<td>1 0 2 0</td>
<td>0 0 0 0 570 0 0 91 1</td>
<td>1 0 3 1 7 0 0 0 0 0 0 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1 1 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18486</td>
<td>1 0 1 1 0 3 740 0 43 0 76 0</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1 1 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18622</td>
<td>1 0 2 0 0 0 680 0 2 90 6 7 0</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18900</td>
<td>1 0 2 0 0 0 740 0 45 0 68 0</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1 1 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19267</td>
<td>1 0 2 1 0 3 670 1 1 1 6 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1 1 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19430</td>
<td>1 0 1 2 0 5 70 0 5 90 6 5 8</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1 1 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 6-2: Monitoring data after column manipulation

6.4 Data Analysis

In order to correlate the data with the residual life, we had to model the relationship between them. This needs more explanation, as we did not use the information from the SOAP as it was collected, but transformed it into another measurement that we called ‘total metal concentrations’ that represents the cumulative metal concentrations since new. The rationale behind this transformation is that, to establish the relationship between the deterioration process from oil analysis and the residual life, we need some quantification for the deterioration process influenced by condition-monitoring variables. Here, the deterioration process is called wear, and we believe that wear is a
non-decreasing process accumulated since new. If we want to use metal concentration to represent a function of wear, then the calculation of the cumulative metal concentration will provide an indirect measure of the cumulative wear since new with random noise. The other reason that we did not use the raw data is that oil top-ups or changes will distort the metal concentration content within the oil sample since a newly flushed engine will have very little metal concentration in the oil. The transformation was performed using the following formula.

Total Wear Metal,

\[
y_i = y_{i-1} + [C*(e_i - e_{i-1})] + [a*(e_i + e_{i-1})/2]
\]

where

\( y_i = \) Total Wear Metal at time \( t_i \), where \( t_i \) is the \( ith \) checking time since new.

\( e_i = \) Element Concentration (ppm) in the \( ith \) oil sample

\( C = \) Oil Capacity of component (litres)

\( a = \) Oil Added (litres)

In this study, we first used metal concentration measures only as monitoring indicators, because their characteristics directly provide important information on the wear condition of internal engine parts (Lukas et al., 1996; Barraclough et al., 2003). Noted however, we also used other indicators provide by SOAP, which will be explained in Chapter 7.

Using equation (6-1) and considering only metal concentration measures, we plotted the value of the total concentration of each metal against operating hours, as shown in Figure 6-3. Only six metal concentrations from the metal group were available from the dataset that could be taken to characterize the residual time of the system.
At this early stage, we have two major issues in analysing the dataset. The first is the incomplete nature of the data. This refers to the recorded monitoring indicators, which were not from new as we required. As an example, Figure 6-3 shows the observed data started around 10,000 hours before the engine was replaced to prevent failure, or at failure. Despite the problem mentioned above, we do have a small sample of SOAP data, which was collected from new. However, this type of data is monitored up to 10,000 hours only. To make a complete history of the SOAP for those engines with missing early data, we used a simple linear regression based on the availability of the small dataset of early SOAP from other engines, by extrapolating these values. The transformation resulted in a complete set of total metal concentration from wear elements, which could be taken to model the residual time of the system.

As mentioned above, we have six element indicators of the total metal concentration that could be used to predict the residual life. We had difficulty in choosing which of these indicators are really useful and can be expected to produce the best results. Hence, we had two options regarding the dimensions of total metal concentration indicators to be used in our model. The dimension of a model is the number of independent or input indicators used by the model. The first option was to use all the total metal concentrations as given, but this would result in a complex model where all metal concentrations could be correlated with each other; we would have to use a joint probability density function for them, which is difficult and requires more parameters to be estimated. The second option was to reduce the correlation and the dimensions of the total metal concentrations, but at the same time we may lose some of our original information.

Figure 6-3: A sample of total metal concentration after the transformation
For model simplification, we chose the second option, and used the widely known data decomposition technique called Principal Component Analysis (PCA) to simplify the data. Generally, PCA is a useful procedure dealing with dimension reduction techniques, especially when we have a set of sample measurements that are highly correlated (Jolliffe, 1986).

6.4.1 Principal Component Analysis

PCA encodes the most relevant information contained in a sample in a set of orthonormal vectors. This set defines a characteristic subspace that contains the main features of the sample, and the number of selected vectors defines the amount of variance that can be explained by the PCA model. To start with, we defined $\mathbf{y}_i$ where $i=1,2,\ldots,m$ is a variable vector that represents sample data, and $\mathbf{v}_i$, $i=1,2,\ldots,n$ are the principal components; the linear relationship between $\mathbf{y}_i$ and $\mathbf{v}_i$ is given by Williams et al. (1995) as:

$$
\mathbf{v}_1 = u_{11}\mathbf{y}_1 + u_{12}\mathbf{y}_2 + u_{13}\mathbf{y}_3 + \ldots + u_{1m}\mathbf{y}_m \\
\mathbf{v}_2 = u_{21}\mathbf{y}_1 + u_{22}\mathbf{y}_2 + u_{23}\mathbf{y}_3 + \ldots + u_{2m}\mathbf{y}_m \\
\mathbf{v}_3 = u_{31}\mathbf{y}_1 + u_{32}\mathbf{y}_2 + u_{33}\mathbf{y}_3 + \ldots + u_{3m}\mathbf{y}_m \\
\vdots \\
\mathbf{v}_n = u_{n1}\mathbf{y}_1 + u_{n2}\mathbf{y}_2 + u_{n3}\mathbf{y}_3 + \ldots + u_{nm}\mathbf{y}_m \\
$$

(6-2)

where $\mathbf{u}_k = [u_{k1}, u_{k2}, u_{k3}, \ldots, u_{km}]^T$ is the $k$th eigenvector of the correlation or covariance matrix. The principal component $\mathbf{v}_i$’s are uncorrelated and their variances are given by the corresponding eigenvalues. Basically, principal component analysis will generate the same dimension from the original data and rank them according to the value of the variance. The question of how many variables should be retained needs to be answered. As a solution, a number of procedures have been suggested (Green et al., 1978). One flexible approach is to use the Kaiser criterion, which recommends that only principal components of the correlation or covariance matrix with eigenvalues greater than 1 need to be retained. Another technique, called the scree test, allows us to plot the order of eigenvalues and then look for elbows in the curves. A more convenient
approach is to retain only those eigenvalues that account, on a cumulative basis, for some higher proportion of the total variance, such as 75 or 80%.

Using the latter techniques and the scree plot, the dimension of the variable was chosen to be 1, as most of the dataset produces similar results, as depicted in Figure 6-4 and Figure 6-5 below.

<table>
<thead>
<tr>
<th>Eigenvalues 830001/19</th>
<th>Scree Plot</th>
<th>%Total Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2.83E+08</td>
<td></td>
<td>98.21%</td>
</tr>
<tr>
<td>2 2.68E+06</td>
<td></td>
<td>0.93%</td>
</tr>
<tr>
<td>3 1.77E+06</td>
<td></td>
<td>0.61%</td>
</tr>
<tr>
<td>4 4.80E+05</td>
<td></td>
<td>0.17%</td>
</tr>
<tr>
<td>5 1.90E+05</td>
<td></td>
<td>0.07%</td>
</tr>
<tr>
<td>6 3.00E+04</td>
<td></td>
<td>0.01%</td>
</tr>
</tbody>
</table>

**Figure 6-4: Case 1 – Choosing the dimensions of principal component analysis for engine 830001/19**

<table>
<thead>
<tr>
<th>Eigenvalues 830001/26</th>
<th>Scree Plot</th>
<th>%Total Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1.0532E+08</td>
<td></td>
<td>98.71%</td>
</tr>
<tr>
<td>2 9.9000E+05</td>
<td></td>
<td>0.93%</td>
</tr>
<tr>
<td>3 2.0000E+05</td>
<td></td>
<td>0.19%</td>
</tr>
<tr>
<td>4 8.0000E+04</td>
<td></td>
<td>0.07%</td>
</tr>
<tr>
<td>5 6.0000E+04</td>
<td></td>
<td>0.06%</td>
</tr>
<tr>
<td>6 5.0000E+04</td>
<td></td>
<td>0.05%</td>
</tr>
</tbody>
</table>

**Figure 6-5: Case 2 – Choosing the dimensions of principal component analysis for engine 830001/26**

In short, the transformation procedure in this approach consists of the following steps:

1. Obtain a sample set of raw data of total metal concentrations. The sample data will form a matrix that consists of observed variables at every monitoring point.
2. Compute the correlation or covariance matrix.
3. Compute the eigenvalues and eigenvectors of the correlation matrix above.
4. Order the eigenvalues and corresponding eigenvectors from greater to smaller. Note that the number of eigenvectors is equal to the number of variables in the sample.
5. Choose principal components and form a matrix of vectors.
6. Derive a new dataset. This can be done by taking the transpose of the vector and multiplying it with the original data set transposed.

After carrying out the analysis, we concluded that a single dimension of variables should simplify our model to represent overall total metal concentration. In subsequent analyses, we shall use the first principal component of total metal concentration as our monitoring information, $y_i$, unless otherwise specified.

The next problem that we encountered is irregular monitoring intervals. In fact, we could use the total metal concentration as it is, but this would increase modelling complexity as extra parameters may be needed to take the irregular interval into consideration and cause difficulty in parameter estimation. Since our interest is the total metal concentration, we need to re-organise the data and set up an imaginary regular monitoring interval for all data sets. In Figure 6-6 the $\ast$, represents the original value of the total metal concentrations at irregular time checking points. To set up a regular interval, we find out the mean interval from the data as our approximation value, which is 220 hours. Using this as our regular interval for all engines over their lifetime, we have a representation shown in Figure 6-6, where $\circ$ denotes the imagined monitoring points.
Figure 6-6: Re-organizing condition-monitoring data from original reading

The difference before and after re-organising the dataset is shown in Figure 6-7 below.

![Figure 6-7: The difference before and after re-organising](image)

Having solved these problems, we now have a dataset, which contained information from new, and had equal monitoring intervals for ease of our modelling development. An attempt to formulate the residual time prediction model using the cleaned data of the diesel engines was then carried out. A 1st PCA of the final 35 datasets for regular observed monitoring information from new, used in the subsequent analysis, is given in Figure 6-8 below.
The next section demonstrates the methodology for our proposed model.

### 6.5 Modelling Methodology

First, we define the state variables of the system monitored as the residual time conditional on current and past observed monitoring parameters. In this case we have the total metal concentration that indicates the state of the system at every checking time. It must be noted here that we do not know when failure will occur, so the prediction of the residual time is made in a probability distribution.

To estimate the distribution of the residual life of the engines, we seek to apply the filtering model developed by Wang (2002) to establish a model of residual time predictions. For the purpose of model building, we use the following notation:

1. $X_i$ is a random variable used to represent the conditional residual time of the system at time $t_i$ from new and $x_i$ is its realization.

2. The first principal of the total metal concentration at time $t_i$ is denoted by random variable $Y_i$. Its measure of $y_i$ is obtained by transformation from observed metal concentration and is assumed to be correlated with the residual time. Note that $Y_i$ could be a vector if several principal components are used. In
the following we simply use $y_i$ to represent realization of the condition information at $t_i$.

3. The condition monitoring history obtained up to time $t_i$ is denoted by $\mathcal{I}_i$, where
$$\mathcal{I}_i = \{y_1, y_2, \ldots, y_i\}.$$ 

4. We seek $p(x_i \mid \mathcal{I}_i)$ the pdf of $X_i$ given $\mathcal{I}_i$.

The assumptions made were as follows:

1. Condition monitoring takes place at time $t_i$, $i = 1, \ldots, n$, where the interval between consecutive monitoring checks is equal.
2. The equipment considered here is a single-component system subject to one dominant failure mode, i.e. wear-related failure.
3. The state of the system, $x_i$, influences its monitored condition information, $y_i$, via $p(y_i \mid x_i)$ and not vice versa. This relationship requires more explanation.
   The underlying process of $X_i$ is independent of $Y_i$, but dependent on $X_{i-1}$. However since $X_i$ is not observable but influences $Y_i$ through $p(y_i \mid x_i)$, then the estimated pdf, $p(x_i \mid \mathcal{I}_i)$ becomes dependent on $\mathcal{I}_i = \{y_1, y_2, \ldots, y_i\}$.

This allows us to formulate the model as briefly discussed in the next section.

6.6 Model Formulation

According to Wang et al. (2000), the distribution of the residual time, $X_i$, given the history of the monitoring information, $\mathcal{I}_i$, is given by $p(x_i \mid \mathcal{I}_i)$. Mathematically $p_i(x_i \mid \mathcal{I}_i)$ can be shown as

$$p_i(x_i \mid \mathcal{I}_i) = \frac{\int p(y_i \mid x_i)p(x_i \mid \mathcal{I}_{i-1}) \, dx_i}{\int_0 \int p(y_i \mid x_i)p(x_i \mid \mathcal{I}_{i-1}) \, dx_i}$$ (6-3)
Here we define the remaining time $x_i$ at $t_i$ as the remaining time at $t_{i-1}$ minus the interval between $t_i$ and $t_{i-1}$ provided the item has survived to $t_i$ and no maintenance action has been taken since. In mathematical notation it can be written as follows (Wang et al., 2000):

$$x_i = \begin{cases} 
  x_{i-1} - (t_i - t_{i-1}) & \text{if } x_{i-1} > t_i - t_{i-1}, \text{ no maintenance intervention,} \\
  \text{not defined} & \text{otherwise}
\end{cases} \quad (6-4)$$

From equation (6-4), $p(x_i \mid \mathcal{X}_{i-1})$ can be written as

$$p(x_i \mid \mathcal{X}_{i-1}) = p_{i-1}(x_{i-1} = x_i + t_i - t_{i-1} \mid x_{i-1} > t_i - t_{i-1}, \mathcal{X}_{i-1})$$

$$= \frac{p_{i-1}(x_{i-1} = x_i + t_i - t_{i-1} \mid \mathcal{X}_{i-1})}{\displaystyle \int_{t_i-t_{i-1}}^{\infty} p_{i-1}(z = x_{i-1} \mid \mathcal{X}_{i-1})dz} \quad (6-5)$$

Here we assumed that $x_0$ is the residual time measured since new and all measurements of the condition monitoring parameter $y_i$ used are also since new. Since $\mathcal{X}_0$ is not available at $t_0$ in most cases, we could set $p(x_0 \mid \mathcal{X}_0) = p(x_0)$, which is the pdf. of the engine life.

The next stage is to establish the relationship between the observed information, $y_i$, and the residual time, $x_i$, which can be established by a probability distribution (Wang et al., 2000). We know that our monitored information $y_i$ is an increasing function (see Figure 6-8); hence, at every checking time, the value of $y_i$ must be greater than or equal to $y_{i-1}$. This can be modelled by a 3-parameter pdf of $y_i \mid x_i$ with $y_{i-1}$ as the location parameter, so $y_i$ is always $\geq y_{i-1}$.

Generally, we expect that a short residual time, $x_i$, will generate a high reading in $y_i$. It can be shown that equation (6-3) can be determined recursively if $p(x_0 \mid \mathcal{X}_0)$ and $p(y_i \mid x_i)$ are known. The forms of $p(x_0)$ and $p(y_i \mid x_i)$ can be chosen from any distribution and finally assessed against the goodness-of-fit. In the case of $p(x_0)$, its
estimated parameter is provided by engine failure data or by subjective assessment of experienced experts who are familiar with the engine. In contrast, the model parameters in $p(y_i \mid x_i)$ have to be estimated from observed data of $y_i$.

To demonstrate the basic model, we begin by looking at the distribution of $p(x_0)$. Here, we consider a Weibull distribution, as it is one of the most commonly used distributions in reliability engineering applications. A Weibull distribution is characterized by its shape and scale parameters, and by changing the shape parameter, the Weibull distribution can be made to have many different shapes, from highly skewed like an exponential distribution to nearly bell-shaped like a normal distribution. Because of the many shapes it attains for various values, it can model a great variety of data and life characteristics. Thus, the pdf of engine life, $x_0$, is given as

$$p(x_0) = \alpha\beta(\alpha(x_0))^{\beta-1} e^{-(\alpha x_0)^\beta}$$  \hspace{1cm} (6-6)$$

where $\alpha$ and $\beta$ are the scale and shape parameters, respectively.

Next, we proceed with modelling the increment of total metal concentrations given the residual time. Figure 6-8 reveals a trend of $y_i$ (almost linear) with a constant variance. This implies that we need a pdf. of increasing $y_i$ in $t_i$ and a constant variance for the random variable in $Y_i \mid X_i$. Here, we also choose a Weibull distribution to describe the relationship of $p(y_i \mid x_i)$, which is written as

$$p(y_i \mid x_i) = \lambda_i\eta(\lambda_i(\Delta y_i))^{\eta-1} e^{-(\lambda_i\Delta y_i)^\eta}$$  \hspace{1cm} (6-7)$$

where $\lambda_i$ and $\eta$ are the scale and shape parameters, respectively. $\Delta y_i = y_i - y_{i-1}$ and $y_{i-1}$ is the location parameter.

Based upon the data that we have observed, the conditional relationship between $y_i$ and $x_i$ is established by function $\lambda_i$, where $E(Y_i) \propto \frac{1}{\lambda_i}$. At $t_i$, when $x_i$ decreases, $\lambda_i$ decreases as well, which will result in an increase in $y_i$, so $x_i$ and $y_i$ are negatively
correlated. It is noted that as \( t_i \) increases, \( y_i \) also increases, influenced by \( x_i \). Surely, the increment of \( y_i \) may also produce a large variance of \( y_i \) which we did not want it to happen. Hence, since \( x_i + t_i \) is the engine life, by carefully choosing \( \lambda_i = ae^{b(x_i+t_i)} \), it implies a constant variance and mean for a particular engine if the life is assumed to be fixed but unknown, according to a Weibull distribution. This can be seen as

\[
E(Y_i) = \frac{\Gamma(1+\frac{1}{\beta})}{\lambda_i} \quad \text{and variance as} \quad Var(Y_i) = \frac{\Gamma(\frac{2}{\beta} + 1) - (\Gamma(\frac{1}{\beta} + 1)^2)}{\lambda_i^2}
\]

Substituting equation (6-5) to equation (6-3), it can be updated as

\[
p_i(x_i \mid \mathcal{F}_i) = \frac{p(y_i \mid x_i)p_{i-1}(x_i + t_i - t_{i-1} \mid \mathcal{F}_{i-1})}{\int_0^\infty p(y_i \mid x_i)p_{i-1}(x_i + t_i - t_{i-1} \mid \mathcal{F}_{i-1})dx_j}
\]

(6-8)

To show the recursive steps, we used equations (6-6) and (6-7) and substitute them into equation (6-8) above. Starting with \( i = 1 \), that is

\[
p_i(x_i \mid \mathcal{F}_i) = \frac{p(y_i \mid x_i)p_0(x_i + t_i)}{\int \int_0^\infty p(y_i \mid x_i)p_0(x_i + t_i)dx_i}
\]

\[
= \lambda^\beta \eta \Delta y_i^{\gamma-1} \frac{e^{-\Delta y_i \gamma} * \alpha^\beta \beta (x_i + t_i)^{\beta-1} e^{-(\alpha(x_i+t_i))^\gamma}}{\int \int_0^\infty \lambda^\beta \eta \Delta y_i^{\gamma-1} \frac{e^{-\Delta y_i \gamma} * \alpha^\beta \beta (x_i + t_i)^{\beta-1} e^{-(\alpha(x_i+t_i))^\gamma}}{dx_i}
\]

\[
= \frac{(ae^{b(x_i+t_i)})^\eta e^{-(ae^{b(x_i+t_i)})(\Delta y_i)^\gamma} * (x_i + t_i)^{\beta-1} e^{-(\alpha(x_i+t_i))^\gamma}}{\int_0^\infty (ae^{b(x_i+t_i)})^\eta e^{-(ae^{b(x_i+t_i)})(\Delta y_i)^\gamma} * (x_i + t_i)^{\beta-1} e^{-(\alpha(x_i+t_i))^\gamma} dx_i}
\]

(6-9)

When \( i = 2 \), we have

\[
p_2(x_2 \mid \mathcal{F}_2) = \frac{p(y_2 \mid x_2)p_1(x_2 + t_2 - t_1 \mid \mathcal{F}_1)}{\int_0^\infty p(y_2 \mid x_2)p_1(x_2 + t_2 - t_1 \mid \mathcal{F}_1)dx_2}
\]
\[
\lambda^n \eta \Delta y_2^{n-1} e^{-\lambda_2 y_2^n} \frac{p(\Delta y_1 \mid x_i) p(x_i + t_i)}{\int_0^\infty p(\Delta y_1 \mid x_i) p(x_i + t_i) dx_i} = \int_0^\infty \lambda^n \eta \Delta y_2^{n-1} e^{-\lambda_2 y_2^n} \frac{p(\Delta y_1 \mid x_i) p(x_i + t_i)}{\int_0^\infty p(\Delta y_1 \mid x_i) p(x_i + t_i) dx_i} dx_2
\]

(6-10)

\[
= \frac{(a e^{b_1 x_2 + t_2}) \eta e^{-\alpha(\lambda_1 + \lambda \alpha) \Delta y_1^n} * (a e^{b_1 x_1 + t_1}) \eta e^{-\alpha(\lambda_1 + \lambda \alpha) \Delta y_1^n} * (x_i + t_i) \beta^{-1} e^{-\alpha(\lambda_1 + \lambda \alpha) \Delta y_1^n} \int_0^\infty \lambda^n \eta \Delta y_2^{n-1} e^{-\lambda_2 y_2^n} \alpha \beta (x_i + t_i) \beta^{-1} e^{-\alpha(\lambda_1 + \lambda \alpha) \Delta y_1^n} dx_2
\]

(6-11)

Generalizing this to the \(i\)th observation point, the conditional probability of residual time can be written as

\[
p(x_i \mid \mathcal{Z}_i) = \begin{cases} 
(x_i + t_i) \beta^{-1} e^{-\alpha(\lambda_1 + \lambda \alpha) \Delta y_1^n} \prod_{k=1}^i \psi_k (x_i, t_i), & i > 1 \\
\int_0^\infty (z + t_i) \beta^{-1} e^{-\alpha(\lambda_1 + \lambda \alpha) \Delta y_1^n} \prod_{k=1}^i \psi_k (z, t_i) dz \\
p(x_0) = \alpha \beta (\alpha x_0) \beta^{-1} e^{-\alpha x_0}, & i = 1
\end{cases}
\]

(6-12)
To apply the above model, the values of model parameters in \( p(x_0) \) and \( p(y_i | x_i) \) must be estimated. Here we used the maximum likelihood estimation technique to achieve our goal.

### 6.7 Parameter Estimation

The model parameters are estimated in two steps. The first is to estimate the parameters in \( p(x_0) \), namely \( \alpha \) and \( \beta \), the scale and shape parameters of the Weibull distribution as a failure distribution. This can be done because of our assumption that \( y_i \) is dependent on \( x_i \), but \( x_i \) is independent of \( y_i \). The lifetime data is used to find for this estimation, which is shown in section 6.8.

Given \( \hat{\alpha} \) and \( \hat{\beta} \) are available, we then progress to the second step to estimate parameters \( a \), \( b \) and \( \eta \) which is in \( p(y_i | x_i) \) from the observed data. The observed monitoring data gives two pieces of information that are valuable to us at every monitoring time \( t_i \). The first is the observed monitoring information, \( y_i \), with its history, \( \mathcal{I}_{i-1} \), and the second is that if the engine has survived over \( t_i \), that is the residual time, \( t_{i-1} > t_i - t_{i-1} \). Furthermore, if the last observation is a failure at time \( t_f \) (assuming we have this information, which in practice is hard to obtain), \( t_f > t_n \), where \( t_n \) is the time of the last monitoring check, its contribution to the likelihood function is \( p(x_n = t_f - t_n | \mathcal{I}_n) \). Hence, the likelihood function for a single item becomes

\[
L(\theta, \mathcal{I}_n) \propto \left( \prod_{i=1}^{n} p(y_i | \mathcal{I}_{i-1}) \right) \int_{t_{i-1}}^{\infty} p(x_{i-1} | \mathcal{I}_{i-1}) p(x_n = t_f - t_n | \mathcal{I}_n) \right) \]  \tag{6-13}

where \( p(y_i | \mathcal{I}_{i-1}) \) can be written as

\[
p(y_i | \mathcal{I}_{i-1}) = \int_{0}^{\infty} p(y_i | x_i) p(x_i | \mathcal{I}_{i-1}) dx_i
\]

\[
= \int_{0}^{\infty} p(y_i | x_i) \frac{P_{i-1}(x_i + t_i - t_{i-1} | \mathcal{I}_{i-1})}{\int_{t_{i-1}}^{\infty} p_{i-1}(z | \mathcal{I}_{i-1}) dz} dx_i \tag{6-14}
\]
To acquire a generalized function of \( p(y_1 \mid \mathcal{I}_{i-1}) \), we use equation (6-14) above and start with \( i = 1 \)

\[
p(y_1 \mid \mathcal{I}_0) = \int_0^\infty p(y_1 \mid x_i) \frac{p_0(x_i + t_i - t_0 \mid \mathcal{I}_0)}{t_i - t_0} \, dx_i \]

Using a Weibull distribution for both \( p(x_0) \) and \( p(x_i \mid \mathcal{I}_i) \), we have

\[
p(y_1 \mid \mathcal{I}_0) = \int_0^\infty \lambda \eta (\Delta y_1)^{\gamma-1} e^{-(\lambda \Delta y_1)} \frac{\alpha^\beta \beta (x_i + t_i)^{\beta-1} e^{-(\alpha (x_i + t_i))^{\beta}}}{\int_0^\infty \alpha^\beta \beta x_0^{\beta-1} e^{-(\alpha x_0)^{\beta}} \, dx_0} \, dx_i
\]

\[
= \int_0^\infty \eta \Delta y_1^{\gamma-1} (ae^{\beta (x_i + t_i)})^{\gamma} e^{-(ae^{\beta y_1})^{\gamma}} (x_i + t_i)^{\beta-1} e^{-(\alpha (x_i + t_i))^{\beta}} \, dx_i \int_0^\infty x_0^{\beta-1} e^{-(\alpha x_0)^{\beta}} \, dx_0
\]

When \( i = 2 \), we have

\[
p_2(y_2 \mid \mathcal{I}_1) = \int_0^\infty \lambda \eta (\Delta y_2)^{\gamma-1} e^{-(\lambda \Delta y_2)^{\gamma}} \frac{p_1(x_2 + t_2 - t_1 \mid \mathcal{I}_1)}{t_2 - t_1} \, dx_2 \int_0^\infty p_1(x_1 \mid \mathcal{I}_1) \, dx_1
\]

\[
= \int_0^\infty \eta \Delta y_2^{\gamma-1} \lambda^2 e^{-(\lambda \Delta y_2)^{\gamma}} \int_0^\infty \eta \Delta y_1^{\gamma-1} \lambda^2 e^{-(\lambda \Delta y_1)^{\gamma}} \frac{\alpha^\beta \beta (x_i + t_i)^{\beta-1} e^{-(\alpha (x_i + t_i))^{\beta}}}{\int_0^\infty \eta \Delta y_1^{\gamma-1} \lambda^2 e^{-(\lambda \Delta y_1)^{\gamma}} \alpha^\beta \beta (x_i + t_i)^{\beta-1} e^{-(\alpha (x_i + t_i))^{\beta}} \, dx_i} \, dx_2
\]

\[
= \int_0^\infty \eta \Delta y_2^{\gamma-1} (ae^{\beta (x_2 + t_2)})^{\gamma} e^{-(ae^{\beta \gamma})^{\gamma}} (ae^{\beta (x_1 + t_1)})^{\gamma} e^{-(ae^{\beta \gamma})^{\gamma}} (x_i + t_i)^{\beta-1} e^{-(\alpha (x_i + t_i))^{\beta}} \, dx_2 \int_0^\infty (ae^{\beta (x_1 + t_1)})^{\gamma} e^{-(ae^{\beta \gamma})^{\gamma}} (x_i + t_i)^{\beta-1} e^{-(\alpha (x_i + t_i))^{\beta}} \, dx_i
\]
Generalizing this to the \( i \)th observation point, \( p(y_i \mid \mathcal{I}_{t-1}) \) becomes

\[
\begin{aligned}
&= \left\{ \begin{array}{l}
\int_{0}^{\infty} \frac{\eta \Delta y_{i}^{\eta-1} (ae^{b(x_i + t_i)})\eta e^{-(\alpha e^{b(x_i + t_i)})\beta_i} (x_i + t_i)\beta_i e^{-(\alpha (x_i + t_i))\beta_i} dx_1}{\int_{t_2 - t_1}^{\infty} (ae^{b(x_i + t_i)})\eta \frac{e^{-(\alpha e^{b(x_i + t_i)})\beta_i} (x_i + t_i)\beta_i e^{-(\alpha (x_i + t_i))\beta_i} dx_1} {dx_2}} \\
0 \leq i \leq \infty
\end{array} \right.
\end{aligned}
\]
When \( n = 2 \), we have

\[
L(\theta, \mathcal{F}_2) \propto \left\{ \prod_{j=1}^{2} p(y_j \mid \mathcal{F}_j) \int_{t_{j-1}}^{\infty} p(x_{j-1} \mid \mathcal{F}_{j-1}) dx_{j-1} \right\} \left( p(x_2 = t_f - t_2 \mid \mathcal{F}_2) \right) = p(y_2 \mid x_2) p(y_1 \mid x_1) p(x_1 + t_1)
\]

\[
= (ae^{b(x_2 + t_1)})^\eta \eta \Delta y_2^{\eta - 1} e^{-(ae^{b(x_2 + t_1)})^\eta} (ae^{b(x_1 + t_1)})^\eta \eta \Delta y_1^{\eta - 1} e^{-(ae^{b(x_1 + t_1)})^\eta} \alpha^\beta \beta(x_1 + t_1)^{\beta - 1} e^{-(\alpha(x_1 + t_1))^\beta}
\]

\[
= \alpha^\beta \beta f e^{-(\alpha t_f)^\beta} \prod_{k=1}^{2} \eta \Delta y_k^{\eta - 1} (ae^{b(t_f)})^\eta e^{-(ae^{b(t_f)})^\eta} (\eta^{\beta - 1})
\]

(6-19)

Generalizing this to the \( nth \) observation point, the likelihood function becomes

\[
L(\theta, \mathcal{F}_n) \propto \alpha^\beta \beta f e^{-(\alpha t_f)^\beta} \prod_{k=1}^{n} \eta \Delta y_k^{\eta - 1} (ae^{b(t_f)})^\eta e^{-(ae^{b(t_f)})^\eta} (\eta^{\beta - 1})
\]

(6-20)

Parameter estimates that maximize the likelihood of equation (6-20) are the most likely values for the parameters. For numerical convenience, the log of the likelihood is used. Therefore, equation (6-20) can be rewritten as

\[
LogL(\theta, \mathcal{F}_n) = \beta \log \alpha + \log \beta + (\beta - 1) \log t_f - (\alpha t_f)^\beta + \sum_{k=1}^{n} \left\{ \log(\eta) + (\eta - 1) \log \Delta y_k + \eta \log(ae^{b(t_f)}) - (ae^{b(t_f)})^\eta \right\}
\]

(6-21)

where \( n \) is the number of condition monitoring checks, \( t_f \) is the actual failure time, \( \alpha, \beta \) are the parameters from \( p(x_0) \) and \( a, b, \eta \) are parameters which we want to estimate.

Similarly, for \( m \) life datasets the likelihood function becomes

\[
L(\theta, \mathcal{F}_m) \propto \left\{ \prod_{j=1}^{m} \alpha^\beta \beta f e^{-(\alpha t_f)^\beta} \prod_{k=1}^{n} \eta \Delta y_k^{\eta - 1} (ae^{b(t_f)})^\eta e^{-(ae^{b(t_f)})^\eta} (\eta^{\beta - 1}) \right\}
\]

(6-22)

Taking logs on both sides, the log-likelihood is written as
\[ \text{LogL}(\theta, J) = \sum_{j=1}^{m} \left\{ \beta \log \alpha + \log \beta - (\beta - 1) \log t_j - (\alpha t_j)^\beta + \sum_{k=1}^{n_j} \left\{ \log(\eta) + (\eta - 1) \Delta y_{jk} + \eta \log (ae^{bt_j}) - (ae^{bt_j})^\eta \right\} \right\} \] (6-23)

where \( n_j \) is the number of condition monitoring checks for the \( j \)th item, \( t_j \) is the actual failure time for the \( j \)th item, \( \alpha, \beta \) are the parameters from \( p(x_0) \) and \( a, b, \eta \) are parameters we want to estimate.

Further work on the model developed above was carried out with actual data from our analysis, described in the next section.

6.8 The Case Study

This section describes attempts to fit the model we had developed with 54 sets of life data of marine diesel engines. According to the lifetime data obtained, 11 diesel engines were replaced due to failure while in operation and the remaining 43 were replaced through preventive maintenance. It is noted here that all engines were re-conditioned rather than replaced. Here, we assumed re-conditioning is equivalent to replacement, since during re-conditioning the engine is examined for any damage or major defects including breakage, cracks, corrosion, wear, etc. Though some of the re-conditioned engines may have some minor cosmetic imperfections, they are operationally perfect and contain the same components that a new engine would have.

Our aim at this stage was to ascertain the lifetime distribution of the data and compute respective parameters from it. At the outset, it was assumed that the lifetime distribution of the data was a Weibull with scale parameter \( \alpha \) and shape parameter \( \beta \). Here, we treat replacement due to failure as a complete data and preventive replacement as a censored data. Complete data means that the failure time of an engine is observed, whereas censored data indicates that the engine had not yet failed when the data was obtained. In order to find \( p(x_0) \) we planned to use the complete data, representing the engines’ complete life. However, when we looked in detail, only 10 engines from the group of a complete dataset, provides a measure concerning the complete life of the
engines. Also, as our assumption referred to failure because of wear, one engine had to be eliminated, because it was replaced as a result of alternator failure that was not wear related. Thus, it reduces the sample of complete data, which could affect our parameter estimation if we still want to use only complete life data.

One way to carry out parameter estimation is by using the complete data with censored data as well. In our case, most of the engines were replaced preventively at around 20,000 hours. Consideration of censored data is important for a realistic estimation, since it contains valuable information about the survival times of the engine. In the case of using all the complete and censored data, the likelihood function for parameter \( \theta = \{\alpha, \beta\} \) is given by

\[
L(\theta, t) \propto \prod_{i \in U} f(\theta, t) \prod_{i \in C} S(\theta, t)
\]  

(6-24)

where \( f(\theta, t) \) and \( S(\theta, t) \) are the probability density and survival probability of the chosen distribution. \( U \) is the set of items which are observed to fail and \( C \) is the set of right-censored observations. The log likelihood function can be written as

\[
\log L(\theta, t) = \sum_{i \in U} \log f(\theta, t) + \sum_{i \in C} \log S(\theta, t)
\]  

(6-25)

Using our data, Table 6-1 and Table 6-2 below show the estimated parameters and their variances respectively.

<table>
<thead>
<tr>
<th>( \hat{\alpha} )</th>
<th>( \hat{\beta} )</th>
<th>( E(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0005586</td>
<td>2.298</td>
<td>38,064 hours</td>
</tr>
</tbody>
</table>

Table 6-1: Estimated parameters

<table>
<thead>
<tr>
<th>( \text{var}(\hat{\alpha}) )</th>
<th>( \text{var}(\hat{\beta}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.837E-8</td>
<td>0.466</td>
</tr>
</tbody>
</table>

Table 6-2: Variances of estimated parameters
According to the estimates we made, the expected life of an engine was too far (38,064 hours) from the recommended engine life (20,150 hours) given by the personnel dealing with this data. We could therefore argue about what is the definition of an engine failure: it either fail completely or fail to perform its intended functions. It became clear at this stage that the censored data we had in terms of preventive replacement was not random but purposely censored because of the definition of functional failure. Due to limited failure data and purposely-censored data, we proposed two approaches that could work to estimate the lifetime distribution given the incomplete life data as follows:

1. Using failure and interval-censored information.
2. Predicting failures of those censored using a regression model based on existing failure data.

Interval-censored data arises when a failure time cannot be observed, but can be estimated to lie within an interval. This assumes that the life of a censored engine can only survive up to the maximum point, \( t_{\text{max}} \), hence the likelihood function can be written as

\[
\log L(\theta, t) = \sum_{i \in U} \log f(\theta, t) + \sum_{i \in C} \log S(\theta, t) - S(\theta, t_{\text{max}}) \tag{6-26}
\]

In our case, we could choose \( t_{\text{max}} = \max(t_{i \in C}) + K \), where \( K = 100 \) is an arbitrary constant value. The estimation result for this approach is shown in Table 6-3 with the variance of the estimation shown in Table 6-4.

<table>
<thead>
<tr>
<th>( \hat{\alpha} )</th>
<th>( \hat{\beta} )</th>
<th>( E(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0009478</td>
<td>6.256</td>
<td>23.54 hours</td>
</tr>
</tbody>
</table>

Table 6-3: Estimated parameters for failure and interval-censored information

<table>
<thead>
<tr>
<th>var(( \hat{\alpha} ))</th>
<th>var(( \hat{\beta} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.792E-10</td>
<td>0.651</td>
</tr>
</tbody>
</table>

Table 6-4: Variances of estimated parameters for failure and interval-censored information
The result shows that the expected failure time is not far from the recommended mean value, which provides evidence of the validity of this approach. In addition, the variance for $\hat{\alpha}$ and $\hat{\beta}$ is relatively small compared with the values of estimated parameters, which shows the accuracy of the estimation.

The next method that we believed would give us a better result was that of predicting the engine failures from the incomplete data by using the relationship of total metal concentration values and failure time of complete life data. Using the 10 datasets of failure data that we have, we need to have their failure times and total metal concentrations. However, the metal concentrations from three datasets could not be calculated due to little or no observed condition-monitoring data. Hence, we computed seven datasets that established a linear regression from this relationship, as shown in Figure 6-9 below.

![Figure 6-9: Regression of total metal concentration and failure time](image)

Using this regression technique, we calculated the 95% prediction level for total metal concentration values, which we expected would yield the predicted value for the time to failure. Next, we mapped the total metal concentration from the incomplete replacement dataset into this 95% prediction level, shown in Figure 6-10 below.
As we assume that a shorter residual life will produce higher metal concentration, \( y_i \), we add some extra hours to those censored lives based on the \( y_i \) at their last check, which must be in the 95% prediction interval. To do this, we draw a straight line from the censored time until it touches either the regression line or one of 95% prediction intervals. A lower \( y_i \) will produce a longer residual life. Using this approach provided more failure data, allowing the estimates of the parameters within \( p(x_g) \) to be performed using the likelihood function written below.

\[
L(\theta, t) \propto \prod_{i \in U} f(\theta, t)
\]

(6-27)

where \( f(\theta, t) \) is the probability density of a chosen distribution and \( U \) is the items which are observed to fail. The log likelihood function can be written as

\[
\log L(\theta, t) = \sum_{i \in U} \log f(\theta, t)
\]

(6-28)

The estimated result for this approach is shown in Table 6-5 and the variances of the estimated parameters are shown in Table 6-6 below.
<table>
<thead>
<tr>
<th>( \hat{\alpha} )</th>
<th>( \hat{\beta} )</th>
<th>( E(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0009762</td>
<td>5.333</td>
<td>22,657.68 hours</td>
</tr>
</tbody>
</table>

Table 6-5: Estimated parameters for 95% prediction interval from regression of failure data

<table>
<thead>
<tr>
<th>( \text{var}(\hat{\alpha}) )</th>
<th>( \text{var}(\hat{\beta}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.055E-9</td>
<td>0.753</td>
</tr>
</tbody>
</table>

Table 6-6: Variances of estimated parameters for 95% prediction interval from regression of failure data

The result using this approach for estimating parameters \( \hat{\alpha} \) and \( \hat{\beta} \) also shows that the expected failure was close to the recommended average, which gives confidence in the developed method. Similar confidence is also given by the variances of the estimated parameters, suggesting that the estimation is good. However, it can be observed that the result obtained from this approach is not much different from the result obtained by using the failure and interval-censored data. This scenario takes place because both approaches are seeking a method for adding some extra hours to predict when the failure will occur. As it is hard to justify which one of these approaches could give better results, both approaches were used for further analysis.

The next stage was to find a method to estimate the parameters of \( a, b \) and \( \eta \) from the observed data \( y_i \). In our case, among the 54 sample engines for which data was provided, three engines from the failure replacement group had no information about monitoring checks. The same problem applied in the preventive replacement group, where 15 engines had to be eliminated for having very little or no information about the monitoring checks. Our final dataset for the estimating purpose was thus seven engines in the failure replacement group and 24 in the preventive replacement category.

An analysis was performed on this data with the main objective of transforming the actual monitoring data to obtain understandable condition-monitoring information, i.e. total metal concentrations. The procedure of this analysis was explained in Section 6.4 above. We randomly chose half (22) transformed datasets to undertake parameter
estimation, and reserved the remaining dataset to test the fit of the model. This approach is carried out due to the fact that we need to test and validate the model.

In order to estimate $a$, $b$ and $\eta$, two alternative approaches appeared useful in our model to best describe the data, depending on the method of estimating $p(x_i)$. The first approach, referred to as the ‘anticipated approach’ relied on a prediction of failure time for those censored engines, while the second was called the ‘interval censored approach’ since we put on an interval for the failure time for those preventive replaced engines. For the ‘anticipated approach’ it was possible to use the likelihood function formulated in equation (6-21), but for the ‘interval-censored approach’ the likelihood function had to be adapted to take account of the interval-censored information. Here, we show how the likelihood function for the ‘interval-censored approach’ is established.

If the last observation is censored at time $t_s$, $t_s > t_n$, where $t_n$ is the last monitoring check, and $t_{max+n}$ is the maximum time before which we assume the engine will fail, its contribution to the likelihood function is $\int_{t_s}^{t_{max+n}} p(x_i \mid x_{t_i}) dt_i$. The likelihood function based on the interval-censored approach can be written as

$$L(\theta, \bar{x}_i) \propto \prod_{i=1}^{n} (p(y_i \mid x_{t_i}) \int_{t_s}^{t_{max+n}} p(x_{t_i} \mid x_{t_{i-1}}) dx_{t_{i-1}}) \int_{t_s}^{t_{max+n}} p(x_n \mid x_{n}) dx_n$$

(6-29)

Similarly, generalizing this to the $nth$ observation point, the likelihood function becomes

$$L(\theta, \bar{x}_i) \propto \alpha^\beta \beta \int_{t_s}^{t_{max+n}} \left[ \prod_{i=1}^{n} (ae^{b(x_n + x_{t_i})})^\eta \eta \Delta y_i ^\beta - 1 e^{-(ae^{b(x_n + x_{t_i})})^\eta \Delta y_i ^\beta} \right] (x_n + t_n)^{\beta - 1} e^{-(\alpha + x_n)^\beta} dx_n$$

(6-30)

Taking logs on both sides,
\[ \log L(\theta, \mathcal{I}) = \beta \log \alpha + \log \beta + \]

\[
\sum_{j=1}^{m} \log \int_{t_{j-1}}^{t_{j}} \left\{ \prod_{i=1}^{n_{j}} \left( ae^{b(x_{i} + t_{n})} \right)^{\eta} e^{-\left( ae^{b(x_{i} + t_{n})}\Delta y_{i} \right)\eta} \right\} \left( x_{n} + t_{n} \right)^{\beta - 1} e^{-\left( \alpha \left( x_{n} + t_{n} \right) \right)^{\eta}} dx_{n} \]

(6-31)

For \( m \) datasets, the likelihood function becomes

\[ L(\theta, \mathcal{I}) \propto \alpha \beta \prod_{j} \int_{t_{j-1}}^{t_{j}} \left\{ \prod_{i=1}^{n_{j}} \left( ae^{b(x_{i} + t_{n})} \right)^{\eta} e^{-\left( ae^{b(x_{i} + t_{n})}\Delta y_{i} \right)\eta} \right\} \left( x_{n} + t_{n} \right)^{\beta - 1} e^{-\left( \alpha \left( x_{n} + t_{n} \right) \right)^{\eta}} dx_{n} \]

(6-32)

Taking logs on both sides,

\[ \log L(\theta, \mathcal{I}) = \beta \log \alpha + \log \beta + \]

\[
\sum_{j=1}^{m} \log \int_{t_{j-1}}^{t_{j}} \left\{ \prod_{i=1}^{n_{j}} \left( ae^{b(x_{i} + t_{n})} \right)^{\eta} e^{-\left( ae^{b(x_{i} + t_{n})}\Delta y_{i} \right)\eta} \right\} \left( x_{n} + t_{n} \right)^{\beta - 1} e^{-\left( \alpha \left( x_{n} + t_{n} \right) \right)^{\eta}} dx_{n} \]

(6-33)

This log likelihood is difficult to maximize, because if one of the parameters is too large or too small, the function will be zero or infinity. Hence we take logs on the inner function of the product term in equation (6-33) and then exponentiate it again, that is

\[
\sum_{j=1}^{n_{j}} \left( \bullet \right) = \sum_{j=1}^{n_{j}} \log \left( ae^{b(x_{i} + t_{n})} \right) + \log(\eta) + (\eta - 1) \log(\Delta y_{i}) - \left( ae^{b(x_{i} + t_{n})}\Delta y_{i} \right)\eta \]

(6-34)

This is purely for computation purposes. The log likelihood of equation (6-34) becomes

\[ \log L(\theta, \mathcal{I}) = \beta \log \alpha + \log \beta + \sum_{j=1}^{m} \log \int_{t_{j-1}}^{t_{j}} \left( \prod_{i=1}^{n_{j}} \left( \bullet \right) \right) \left( x_{n} + t_{n} \right)^{\beta - 1} e^{-\left( \alpha \left( x_{n} + t_{n} \right) \right)^{\eta}} dx_{n} \]

(6-35)

Calculating the log-likelihood function in equation (6-35) using the previously estimated \( \hat{\alpha} \) and \( \hat{\beta} \) based on both approaches gives the results in Table 6-7 and Table 6-8 below.
This provides two sets of estimated parameters for $a$, $b$ and $\eta$ that could be used in our model. To decide which set of estimates should be used, we calculated the simulated $\tilde{y}_i$ based on each set of estimates and compared it with the observed $y_i$. The simulated $\tilde{y}_i$ can be calculated using a two-stage process, where the first stage calculates the expected failure time, using formula $E(t) = \frac{\Gamma(1 + \frac{1}{b})}{\hat{\beta}}$ from a Weibull distribution and the second stage calculates $E(\tilde{y}_i) = \frac{\Gamma(1 + \frac{1}{\hat{\eta}})}{\hat{\lambda} e^{(b \hat{a} \hat{\eta})}}$ also from a Weibull distribution, where we use $x_i + t_i \approx E(t)$. The value of simulated $\tilde{y}_i$ is accumulated at each monitoring time $t_i$. However, using this graphical comparison, it is hard for us to choose the finest set, as it does not indicate any difference between the two methods. Thus, to select the best approach, we used the method of mean square error (MSE), which minimizes the squares of the deviations from the observed $y_i$. The result of the comparison is depicted in Table 6-9 below.

<table>
<thead>
<tr>
<th>MSE using anticipated data</th>
<th>MSE using interval data</th>
</tr>
</thead>
<tbody>
<tr>
<td>3263.2</td>
<td>3758.939</td>
</tr>
</tbody>
</table>

Table 6-9: MSE of anticipated and interval data

The smaller the MSE, the better of the fit, so we accepted the parameters estimated from anticipated data to fit with our model and performed the subsequent analysis. For the
anticipated approach, we calculated the variance and covariance matrix in Table 6-10 as follows.

<table>
<thead>
<tr>
<th></th>
<th>( \hat{a} )</th>
<th>( \hat{b} )</th>
<th>( \hat{\eta} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{a} )</td>
<td>3.43E-10</td>
<td>-3.23E-10</td>
<td>-9.78E-09</td>
</tr>
<tr>
<td>( \hat{b} )</td>
<td>-3.23E-10</td>
<td>3.11E-10</td>
<td>-1.86E-07</td>
</tr>
<tr>
<td>( \hat{\eta} )</td>
<td>-9.78E-09</td>
<td>-1.86E-07</td>
<td>2.52E-04</td>
</tr>
</tbody>
</table>

Table 6-10: Variance and covariance results

The variance (diagonal) of each parameter shows a small value, which implies that our estimation is not far from its mean, while the covariance is small, implying that the parameters are independent of each other.

6.9 Check the Goodness-Of-Fit to Data

The remaining datasets were then used to check the fit of our model to the data. We aimed to predict the residual life of the diesel engines given the observed information. Two datasets that illustrate typical monitoring information are presented in Figure 6-11 below.

![Figure 6-11: Typical monitoring observation](image)
Engine 830001/28 showed a steady increase in monitored information and failed at 27,500 hours, while engine 830001/30 had similar readings at the early stages but showed a sharp increase beginning at 14,300 hours of operation and failed at 22,200 hours. The same $y_i$ on the earlier stages is because they are extrapolated from other engines as the earlier information for both engines is not available to us. This evidence supports our assumption that the higher the total metal concentration the shorter the residual time.

Two examples of the prediction model of residual time in terms of probability distributions and actual residual times are shown in Figure 6-12 and Figure 6-13 below.

Figure 6-12: Case 1 – pdf and actual residual time with monitoring information for engine 830001/28
Both results show that the actual residual time is in the predicted region. To test the hypothesis that the monitored information does influence the model prediction, we calculated the pdf. of the residual time of the diesel engine without the observed information as a comparison. The formulation for the residual time that depends on age without condition-monitoring data can be written using $p_0(x_0)$, as at time $t_i$

$$p_0(x_0) = \frac{p_0(x_i + t_i - t_{i-1})}{\int_0^\infty p_0(x_i + t_i - t_{i-1})dx_i}$$

$$= \frac{(x_i + t_i)^{(\beta-1)} e^{-\alpha(x_i + t_i)^\delta}}{\int_0^\infty (x_i + t_i)^{(\beta-1)} e^{-\alpha(x_i + t_i)^\delta} dx_i}$$  \hspace{1cm} (6-36)

The actual residual time was also plotted. Results are shown in Figure 6-14 below.
To make a more straightforward judgement, we plotted the pdf of each distribution at the last observation point before a failure occurred. A comparison of both approaches is illustrated in Figure 6-15 and Figure 6-16 below.
Both cases show a significant difference but at different scales, which reflects the influence of monitoring data. Figure 6-15 shows that the increment of $\Delta v_i$ for engine 830001/28 is small, which indicates that the influence of condition-monitoring information is small. In contrast, Figure 6-16 shows a very significant difference between the pdf of residual life with and without monitoring information. This is because the higher increment of $\Delta v_i$ does have some effect on the residual time. This shows that the monitoring information plays an important role in prediction.

6.10 Testing the Model

Generally, two tests of statistical inferences can be made with any developed model. The first test is to find the confidence interval of estimated residuals from observed data, which can be used to infer the value of one or more population modes such as mean or standard deviation. To obtain a confidence interval that the true mean value is within the pre-determined interval, we first specify a confidence level, usually 90%, 95% or 99%. Then, using the variance, we calculate the standard deviations of the distribution. A typical confidence interval for the expected value can be written as
\[ \bar{x} \pm (1 - \alpha) \times SE(\bar{x}) \]  

(6-37)

where \( \bar{x} \) is the expected value, \((1 - \alpha)\%\) is the percentage of the confidence level and \( SE(\bar{x}) \) is the standard deviation of the mean. Using a standard \( t \) distribution gives the value for a specified confidence level. It should be noted that as the confidence level increases, the confidence interval becomes wider, showing that the confidence level has a clear effect on confidence-interval length. In this study, some failure times are available to us, so the observed residual time can be calculated. Using the difference between expected residual time and the observed residual time as the length of the confidence interval, we can propose a method to evaluate the model’s accuracy. The shorter the difference is, the better is the estimate. This means that our estimation is close to the observation. Therefore, if \( \bar{x}_i \) is the expected residual time and \( \tilde{x}_i \) is the observed residual time at time \( t_i \), the significant confidence level, \( \alpha \), can be defined as a probability of

\[
P = \begin{cases} 
P(\bar{x}_i < x_i < 2\bar{x}_i - \tilde{x}_i) & \text{if } \tilde{x}_i < \bar{x}_i \\
P(2\bar{x}_i - \tilde{x}_i < x_i < \bar{x}_i) & \text{if } \bar{x}_i < \tilde{x}_i 
\end{cases}
\]  

(6-38)

A similar approach can also be found in Zhang (2004). Hence, we calculated equation (6-38) with the remaining dataset and one example of this calculation are depicted in Table 6-11 below.

<table>
<thead>
<tr>
<th>Sampling time point</th>
<th>Observed residual time</th>
<th>Expected residual time</th>
<th>Confidence level P</th>
</tr>
</thead>
<tbody>
<tr>
<td>770</td>
<td>375.8</td>
<td>281.31</td>
<td>46.00%</td>
</tr>
<tr>
<td>779.2</td>
<td>366.7</td>
<td>275.75</td>
<td>45.00%</td>
</tr>
<tr>
<td>788.3</td>
<td>357.5</td>
<td>269.02</td>
<td>44.00%</td>
</tr>
<tr>
<td>797.5</td>
<td>348.3</td>
<td>262.49</td>
<td>43.00%</td>
</tr>
<tr>
<td>806.7</td>
<td>339.2</td>
<td>260.11</td>
<td>40.00%</td>
</tr>
<tr>
<td>815.8</td>
<td>330</td>
<td>257.82</td>
<td>37.00%</td>
</tr>
<tr>
<td>825</td>
<td>320.8</td>
<td>253.34</td>
<td>35.00%</td>
</tr>
<tr>
<td>834.2</td>
<td>311.7</td>
<td>248.87</td>
<td>33.00%</td>
</tr>
</tbody>
</table>
The calculation for the confidence level using the remaining dataset gave values below 50%, which are good and provide evidence in terms of a good fit to the data. Next, using the remaining dataset, we wondered how well the model actually reflects the data, or in the other words is how ‘close’ the observed values to those, which would be expected under the fitted model. One statistical test that addresses this issue is the chi-square, $\chi^2$, goodness-of-fit test, which determines whether there is a significant difference between the fitted model and the observed. We first formulate the conservative hypothesis, called the null hypothesis ($H_0$), which states that there is no difference between the observed and the expected value. The test procedure begins by arranging the $N$ observations into a set of $k$ class intervals or cells. Generally, the formulation of the chi-squared statistic is written as:

$$
\chi^2 = \sum_{i=1}^{k} \frac{(N_i - Np_i)^2}{Np_i}
$$

(6-39)

where $N_i$ is the number of observations included in category $i$, $p_i$ is the probability that the random variable $X$ of the population falls into category $i$ and $Np_i$ is the expected value of category $i$. The $\chi^2$ statistic is then compared with the $\chi^2$ distribution with degree of freedom $k-1$. If the $\chi^2$ distribution is greater than the $\chi^2$ statistic, we can say that the observed distributions are similar.

Applying it in our remaining dataset, we randomly used five datasets, which includes 492 observations. In each observation, there is a probability distribution of $p(x \mid \mathcal{Z})$.
which are not identical. They do share common parameters and similar pattern, however the only difference are the observed and expected values at different observation points. To carry out the chi-square test, we could transform \( p(x_i | \mathcal{X}_i) \) into a uniform distribution of \( M \) intervals with an equal probability; see Figure 6-17 below; and Chapter 4.

![Figure 6-17: Partition the pdf of \( p(x_i | \mathcal{X}_i) \) into equal probability and transform to uniform distribution](image)

Thus, we partition \( x_i \) into \( M \) cells so that each cell has an equal probability, and then we examine which cell each observation fell into, and count the number of observations in each cell. By doing this transformation, our distribution would follow a uniform pattern; so it was necessary to test whether the data in fact follows a uniform distribution. Chi-square tests allow an arbitrary number of intervals to be chosen, but it is recommended that \( N_i \) values be approximately equal, and that \( N \) is at least five (Law, 2000). In our case, we chose the number of cells to be 10. We then constructed **Table 6-12** to show the calculation of this process.
<table>
<thead>
<tr>
<th>Cell (i)</th>
<th>$N_i$</th>
<th>$p_i$</th>
<th>$Np_i$</th>
<th>$(N_i - Np_i)$</th>
<th>$\left(\frac{(N_i - Np_i)^2}{Np_i}\right)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31</td>
<td>0.1</td>
<td>49.2</td>
<td>-18.2</td>
<td>6.73252</td>
</tr>
<tr>
<td>2</td>
<td>61</td>
<td>0.1</td>
<td>49.2</td>
<td>11.8</td>
<td>2.830081</td>
</tr>
<tr>
<td>3</td>
<td>37</td>
<td>0.1</td>
<td>49.2</td>
<td>-12.2</td>
<td>3.025203</td>
</tr>
<tr>
<td>4</td>
<td>46</td>
<td>0.1</td>
<td>49.2</td>
<td>-3.2</td>
<td>0.20813</td>
</tr>
<tr>
<td>5</td>
<td>56</td>
<td>0.1</td>
<td>49.2</td>
<td>6.8</td>
<td>0.939837</td>
</tr>
<tr>
<td>6</td>
<td>48</td>
<td>0.1</td>
<td>49.2</td>
<td>-1.2</td>
<td>0.029268</td>
</tr>
<tr>
<td>7</td>
<td>52</td>
<td>0.1</td>
<td>49.2</td>
<td>2.8</td>
<td>0.15935</td>
</tr>
<tr>
<td>8</td>
<td>58</td>
<td>0.1</td>
<td>49.2</td>
<td>8.8</td>
<td>1.573984</td>
</tr>
<tr>
<td>9</td>
<td>52</td>
<td>0.1</td>
<td>49.2</td>
<td>2.8</td>
<td>0.15935</td>
</tr>
<tr>
<td>10</td>
<td>51</td>
<td>0.1</td>
<td>49.2</td>
<td>1.8</td>
<td>0.065854</td>
</tr>
<tr>
<td>Total</td>
<td>492</td>
<td>1</td>
<td></td>
<td></td>
<td>15.72358</td>
</tr>
</tbody>
</table>

Table 6-12: Calculating the goodness-of-fit

The computation of the test is shown in Table 6-12 and the value of the test statistics is $\chi^2 = 15.7235$. Referring to the chi-square table, we see that $\chi^2_{0.05} = 16.919$, which is not exceeded by $\chi^2$, so we cannot reject the null hypothesis when $\alpha = 0.05$. Thus, this test gives us a reason to conclude that the distribution of sample data follows its population distribution.

### 6.11 The Decision Model

The essential decision to make at different monitoring points is whether we should replace the engine or not, given all monitored information available. If the answer is ‘yes’, then what is the best time for such replacement, and should we wait for a suitable production window, such as a scheduled shutdown? Suppose that we need to develop a decision model: the decision variables themselves might depend on several criteria such as cost, reliability, safety or any other criterion of interest. All of them require input, that is, the residual time probability distribution. As an example, we may consider the decision model developed by Wang et al (2000) as
\[
C(t) = \frac{C_f P_i(X_i < T - t_i \mid \mathcal{F}_i) + C_p (1 - P_i(X_i < T - t_i \mid \mathcal{F}_i))}{t_i + (T - t_i)(1 - P_i(X_i < T - t_i \mid \mathcal{F}_i)) + \int_0^{T-t_i} z p_i(z \mid \mathcal{F}_i)dz}
\]

(6-40)

where

- \(C(t)\) is the total expected cost per unit time
- \(t_i\) is the current monitoring point measured since new
- \(T\) is the planned replacement time
- \(P_i(X_i < T - t_i \mid \mathcal{F}_i)\) is the probability that \(X_i < T - t_i\) time \(t_i\)
- \(C_f\) is a failure cost
- \(C_p\) is a preventive cost.

The decision model above assumes that the cost of monitoring is negligible. Suppose that the mean cost of a failure is 10 times higher than for preventive maintenance; we can use the estimated parameters and equation (6-40) above to compute the decision model based on engines life data. By plotting the result of the calculation of equation (6-40) at every time \(T\) and assuming it is discrete, we could identify the best time to carry out preventive replacement, \(T\). Two examples are shown from our case studies. In the first case, the engine is monitored for 102 data points, the interval between monitoring points is 9.2 days, and the engine failed after 944.45 days of operation. We see that from Figure 6-18 the cost of preventive replacement decreases at the early stage, which indicates that any effort to replace the engines is unnecessary until it starts to increase at the 80th checking point. This suggests that a preventive replacement should be undertaken to reduce the cost.
Figure 6-18: Case 1 – Expected cost per day in terms of planned replacement at time $T$ given that the current monitoring check is time $t_i$

A similar plot is obtained in Figure 6-19, where in this case the engine was monitored for 72 data points from new, the interval between monitoring points was 9.2 days and it failed at 658.2 days. The expected cost per day is decreasing at an early stage, until it starts to increase at the 69th monitoring point. This suggests that if preventive replacement were carried out, the cost would be minimized. Both cases show that our model works well in recommending preventive replacement before failure occurs.

Figure 6-19: Case 2 – Expected cost per day in terms of planned replacement at time $T$ given that the current monitoring check is time $t_i$
6.12 Summary

This chapter has presented a model to aid maintenance decision making by introducing a measure called the residual time of a marine engine, based upon monitoring data such as wear metal concentration obtained from SOAP. An analysis of such data was carried out to obtain the format required for the model. We then explained how the model was formulated and estimations of its parameters. The model was shown to fit with the actual data, producing a significant result. Next, we carried out two statistical tests to measure the goodness-of-fit of the data with the model. Lastly, we showed how the model developed, can be used in a decision model which determines the appropriate time to carry out maintenance in terms of cost.
CHAPTER 7: CONDITIONAL RESIDUAL TIME MODELLING USING BOTH RESPONSIVE AND REFLECTIVE VARIABLES

7.1 Introduction

The residual time of a machine tends to be highly uncertain and may depend upon several variable factors. In the case of oil analysis, the moment lubricant oil enters an operating machine it begins to deteriorate as it becomes contaminated with wear metal debris, oxidation by-products and external substances such as water, dirt and/or process material. A high level of oil contamination is usually an indication of the presence of a mechanical or lubricant deterioration, which if left unattended could lead to extensive damage or failure of the machine. Generally, using Spectrometric Oil Analysis Programme (SOAP) data provides us with three types of condition-monitoring data, i.e. metal concentration data, lubricant condition data and contaminant condition data. In theory, metal concentration is a good indicator that reflects the wear processes occurring within the engines. By defining the residual lifetime as a representation of the wear process, the relationship between the residual time and metal concentration is established by assuming that the total metal concentration is a function of the residual life. In view of the fact that wear is the root cause of the observed metal concentrations, the relationship between wear (residual lifetime) and metal concentrations is a one-way relationship which implies that the amount of wear determines the amount of metal concentration, with some random noise, but not vice versa. Thus, we could call metal concentration data as reflective or concomitant variables.

The monitoring of the lubricant condition and the contaminants, on the other hand, determines if the lubricant itself is fit for a continued use, based on certain performance measures. Changes in the operating state of a machine, such as stress and loads, can have a substantial effect on the results. Even more importantly, the lubrication schedule within the plant and the types of additions and changes made to the working lubricant within the machine will have a profound effect on the lubricant analysis. Therefore, this information by itself has some role in determining the condition of both the lubricant and the machine, to be used in maintenance decision-making. If we assume that the
lubricant and contaminants will influence the wear rate or residual life, then they are types of what we call responsive variables, which implies that they are, to some extent, responsive or contributing to the wear process.

In a parallel study, Chapter 6 showed how the residual time of marine engines can be modelled using total metal concentrations based on SOAP analysis. However, SOAP analysis can detect metal particles only up to 10 microns in diameter, which implies that larger metal particles are not detected. Hence, it might not capture information that could tell us about the exact accumulated metal concentration in engines. However, it does provide some indication of machine condition. SOAP provides lubricant performance and contaminants measures as indicators of lubricant performance, which could influence the engine wear. Hence, we are interested to develop a residual prediction model that combines the information given by metal concentration, the lubricant data, and contaminants, i.e. both responsive and reflective variables.

7.2 Model Developments

This development of the model is similar to other work (Wang et al., 2000, 2002, 2005; Wang, 2002, 2003, 2004; Zhang, 2004) but with some new extensions, as we embed lubricant and contaminant data as a responsive factor that could increase or decrease the residual time, which has not been studied before in this context. The notation used in our model building is as follows:

7.2.1 Notation

1. $X_i$ is a random variable representing the residual time at time $t_i$ with $x_i$ as its realization.
2. $Y_i$ is a type of reflective monitoring information which reflects the condition of the engine state. $Y_i$ is random and its realization is $y_i$. The relationship between $X_i$ and $Y_i$ is described by $p(y_i | x_i)$.
3. $z_i$ is not a random variable hence it is treated as a covariable, which indicates a type of responsive monitoring information which may influence $X_i$, but is otherwise assumed not to have an influence on $Y_i$. This is because, if the measurements are
correlated, it is sufficient to use only one measurement. Thus, at this early stage, a simple analysis has been carried out that supports our assumption that there is no correlation between \( Y_i \) and \( z_i \).

4. \( X^+_i \) is the residual time before information \( z_i \) is taken at time \( t_i \). In our case, both information \( y_i \) and \( z_i \) are available at the same time.

5. \( X^-_i \) is the residual life after the information \( z_i \) is taken at time \( t_i \).

6. \( \mathcal{Z}_i \) and \( \mathcal{N}_i \) are the monitoring history for \( y_i \) and \( z_i \), that is \( \mathcal{Z}_i = \{y_1, y_2, \ldots, y_i\} \) and \( \mathcal{N}_i = \{z_1, z_2, \ldots, z_i\} \).

### 7.2.2 Model Formulations

Our objective is to find the residual time distribution, given both responsive and reflective data, i.e. \( p(x^+_i \mid \mathcal{Z}_i, \mathcal{N}_i) \) at time \( t_i \). To start with, we defined the residual time adopted from Wang and Christer (2000) as

\[
x^-_i = \begin{cases} x^+_{i-1} - (t_i - t_{i-1}) & \text{if } x^+_{i-1} > t_i - t_{i-1}, \\ \text{not defined} & \text{otherwise} \end{cases} \tag{7-1}
\]

The relationship between \( x^-_i \) and \( x^+_i \) is assumed to be

\[
x^+_i = x^-_i e^{-B\Delta z_i} \tag{7-2}
\]

where \( B \) is a parameter that defines the influence of covariate \( \Delta z_i \) on \( x^-_i \) and \( \Delta z_i = (z_i - z_{i-1}) \). Note that if \( \Delta z_i \) is a vector, then this relationship can be presented as

\[
x^+_i = x^-_i e^{-\sum B_{i,j} \Delta z_{i,j}} \tag{7-3}
\]

where \( M \) is the size of the vector. For the sake of simplicity, we will use a scalar vector to explain the influence of \( \Delta z_i \) on the model. We will deal with this vector while fitting the model to data. Hence, equations (7-2) and (7-3) can be explained as follows:

1. if \( \Delta z_i = 0 \), then \( x^+_i = x^-_i \), which means the residual time remains the same.
2. If $\Delta z_i > 0$, then $x_i^+ < x_i^-$, which means the residual time becomes shorter and the system is deteriorating.

3. If $\Delta z_i < 0$, then $x_i^+ > x_i^-$, which means the residual life increases as the system improves.

We seek to establish the $p(x_i^+ | \mathcal{I}_i, \mathcal{N}_i)$. From Wang and Christer (2000), $p(x_i^+ | \mathcal{I}_i, \mathcal{N}_i)$ can be written as

$$p(x_i^+ | \mathcal{I}_i, \mathcal{N}_i) = \frac{\int_0^\infty p(y_i | x_i^+) p(x_i^+ | \mathcal{I}_{i-1}, \mathcal{N}_{i-1}) dx_i^+}{\int_0^\infty p(y_i | x_i^+) p(x_i^+ | \mathcal{I}_{i-1}, \mathcal{N}_{i-1}) dx_i^+}$$

(7-4)

By directly using the relationship of $p(y_i | x_i^+)$ and equations (7-1) and (7-2), equation (7-4) can have the following form

$$p(x_i^+ | \mathcal{I}_i, \mathcal{N}_i) = \frac{\int_0^\infty p(y_i | x_i^+) p(x_i^+ = x_i^+ e^{Bz_i} + t_i - t_{i-1} | \mathcal{I}_{i-1}, \mathcal{N}_{i-1}) dx_i^+}{\int_0^\infty p(y_i | x_i^+) p(x_i^+ = x_i^+ e^{Bz_i} + t_i - t_{i-1} | \mathcal{I}_{i-1}, \mathcal{N}_{i-1}) dx_i^+}$$

(7-5)

Equation (7-5) is important since it explains how the condition-monitoring data is taken into account in the model, which needs further clarification. $p(y_i | x_i^+)$ is a measurement model specifying the relationship between observed reflective and responsive data and residual time immediately after time $t_i$. The rationale behind this approach is that we observe the information $y_i$ and $z_i$ at the same time. Two key issues must be discussed here. The first is, what is the distribution of $p(x_0^+ | \mathcal{I}_0, \mathcal{N}_0)$ at time $t_0$. Since $\mathcal{I}_0$ and $\mathcal{N}_0$ are not available at $t_0$ in most cases, we could set-up $p(x_0^+ | \mathcal{I}_0, \mathcal{N}_0) = p(x_0)$, which is the $pdf$ of the machine life.

The second issue is to establish the relationship between the observed information, $y_i$ and $z_i$, with the residual time, $x_i^+$, which can be established by a probability distribution (Wang et al., 2000). Generally, we expect that a short residual time, $x_i^-$, will generate a
high reading in $y_i$. However, the covariate, $z_i$, could increase or decrease the residual time, $x_i^+$. This relationship is modelled in equation (7-2). Thus, the forms of $p(x_0)$ and $p(y_i \mid x_i^+)$ can be chosen from a distribution and it can be shown that (7-5) can be determined recursively if $p(x_0 \mid \mathcal{I}_0)$ and $p(y_i \mid x_i^+)$ are known. As an example, the Weibull distribution is chosen to represent both $p(x_0)$ and $p(y_i \mid x_i^+)$. Given

$$p(y_i \mid x_i^+) = \eta \lambda^\gamma y_i^{\gamma - 1} e^{-(\lambda y_i)^\gamma}$$  \hspace{1cm} (7-6)

where $\lambda = ae^{b(x_i^+ + \epsilon_i)}$ and

$$p(x_0 \mid \mathcal{I}_0, \mathcal{U}_0) = p(x_0) = \beta \alpha x_0^{\beta - 1} e^{-(\alpha x_0)^\beta}$$  \hspace{1cm} (7-7)

we start with $i = 1$,

$$p(x_i^+ \mid \mathcal{I}_1, \mathcal{U}_1) = \frac{p(y_i \mid x_i^+)p(x_i^+ = x_i^+ e^{B \alpha x_i} + t_1 - t_0)}{\int_0^\infty p(y_i \mid x_i^+)p(x_i^+ = x_i^+ e^{B \alpha x_i} + t_1 - t_0)dx_i^+}$$

$$= \frac{\left(e^{b(x_i^+ + \epsilon_i)}\right)\left(e^{-\left(a^{x_i^+ + \epsilon_i\gamma}\right)}\right)\left(x_i^+ e^{B \alpha x_i} + t_1\right)^{b-1} e^{-\left(a^{x_i^+ + \epsilon_i\gamma}\right)}dx_i^+}{\int_0^\infty \left(e^{b(x_i^+ + \epsilon_i)}\right)\left(e^{-\left(a^{x_i^+ + \epsilon_i\gamma}\right)}\right)\left(x_i^+ e^{B \alpha x_i} + t_1\right)^{b-1} e^{-\left(a^{x_i^+ + \epsilon_i\gamma}\right)}dx_i^+}$$  \hspace{1cm} (7-8)

then $i = 2$

$$p(x_i^+ \mid \mathcal{I}_2, \mathcal{U}_2) = \frac{p(y_i \mid x_i^+)p(x_i^+ = x_i^+ e^{B \alpha x_i} + t_2 - t_1 \mid \mathcal{I}_1, \mathcal{U}_1)}{\int_0^\infty p(y_i \mid x_i^+)p(x_i^+ = x_i^+ e^{B \alpha x_i} + t_2 - t_1 \mid \mathcal{I}_1, \mathcal{U}_1)dx_i^+}$$

$$= \frac{\left(e^{b(x_i^+ + \epsilon_i)}\right)\left(e^{-\left(a^{x_i^+ + \epsilon_i\gamma}\right)}\right)\left(x_i^+ e^{B \alpha x_i} + t_1\right)^{b-1} e^{-\left(a^{x_i^+ + \epsilon_i\gamma}\right)}dx_i^+}{\int_0^\infty \left(e^{b(x_i^+ + \epsilon_i)}\right)\left(e^{-\left(a^{x_i^+ + \epsilon_i\gamma}\right)}\right)\left(x_i^+ e^{B \alpha x_i} + t_1\right)^{b-1} e^{-\left(a^{x_i^+ + \epsilon_i\gamma}\right)}dx_i^+}$$  \hspace{1cm} (7-9)
Generalizing to the \( i \)th term, \( p(x_i^+ \mid \mathcal{Z}_i, \mathcal{N}_i) \) can be written as

\[
p(x_i^+ \mid \mathcal{Z}_i, \mathcal{N}_i) = \frac{\left(e^{b_i(x_i^+ - t_i)}\right)}{\int_0^{\infty} \left(e^{b_i(x_i^+ - t)}\right) dx_i^+} \prod_{j=1}^{d_i} \left(1 + \frac{\sum_{k=j+1}^{d_i} I_{k,j} (t_k - t_{k-1}) e^{\alpha_i (x_i^+ - t_{k-1})}}{\sum_{k=j+1}^{d_i} I_{k,j} e^{\alpha_i (x_i^+ - t_{k-1})}}\right)
\]

where

\[
\psi_j(x_i^+) = x_i^+ e^{\beta_j} + \sum_{k=j+1}^{d_i} (t_k - t_{k-1}) I_{k,j} e^{\alpha_i (x_i^+ - t_{k-1})} + t_j
\]

and \( I_{k,j} = 0 \) if \( k > i \)

\( I_{k,j} = 1 \) otherwise

This completes our formulation for predicting the conditional residual time, given that we have all the information.

### 7.3 Parameter Estimation

To be able to use equation (7-10) we need to know the parameters within it. The model parameters are estimated in two steps. The first is to estimate the parameters in \( p(x_0) \), namely \( \alpha \), the scale, and \( \beta \), the shape parameters of the Weibull distribution as a failure distribution. This can be done from our assumption that \( y_j \) is dependent on \( x_j \), but \( x_j \) is independent of \( y_j \). The second step is to estimate \( a, b, \eta \) and the coefficients of the covariates \( B_{\text{INS}}, B_{\text{KV}}, B_{\text{TBN}} \) and \( B_{\text{WC}} \) which are in \( p(y_j \mid x_i^+) \) from the observed data.

At every time check \( t_i \), two pieces of information are valuable to us. The first is the observed monitoring information, \( y_i \) with its history, \( \mathcal{Z}_{i-1} \); and the second is if the engine has survived over \( t_i \), the residual time is \( x_{i-1}^+ > t_i - t_{i-1} \). Furthermore, if the last observation is the failure at time \( t_f \) (assuming we have this information, which in practice is hard to obtain), \( t_f > t_n \), where \( t_n \) is the time of the last monitoring check, its
contribution to the likelihood function is $p(x_n = t_f - t_n \mid \mathcal{J}_n, \mathcal{N}_n)$. The likelihood function for a single item using this approach becomes

$$L(\theta, \mathcal{J}) \propto \prod_{i=1}^{n} p(y_i \mid \mathcal{J}_{i-1}, \mathcal{N}_{i-1}) \int_{t_i-t_{i-1}}^{\infty} p(x_{i}^{+} \mid \mathcal{J}_{i-1}, \mathcal{N}_{i-1}) dx_{i}^{+} \left\{ p(x_{n} = t_f - t_n \mid \mathcal{J}_n, \mathcal{N}_n) \right\} (7-11)$$

where $p(y_i \mid \mathcal{J}_{i-1}, \mathcal{N}_{i-1})$ can be written as

$$p(y_i \mid \mathcal{J}_{i-1}, \mathcal{N}_{i-1}) = \int_{0}^{\infty} p(y_i \mid x_i^{+}) p(x_i^{+} e^{B_{i1}} \mid \mathcal{J}_{i-1}, \mathcal{N}_{i-1}) dx_i^{+}$$

$$= \int_{0}^{\infty} p(y_i \mid x_i^{+}) \frac{p(x_i^{+} e^{B_{i1}} + t_i - t_{i-1} \mid \mathcal{J}_{i-1}, \mathcal{N}_{i-1})}{\int_{t_i-t_{i-1}}^{\infty} p(x_{i}^{+} \mid \mathcal{J}_{i-1}, \mathcal{N}_{i-1}) dx_{i}^{+}} dx_{i}^{+} (7-12)$$

Starting with $n = 1$, the likelihood function becomes

$$L(\theta, \mathcal{J}) \propto \left\{ p(y_1 \mid \mathcal{J}_0, \mathcal{N}_0) \int_{0}^{\infty} p(x_{1}^{+} \mid \mathcal{J}_0, \mathcal{N}_0) dx_0 \right\} \left\{ p(x_{1}^{+} = t_f - t_1 \mid \mathcal{J}_1, \mathcal{N}_1) \right\}$$

$$= \left\{ \int_{0}^{\infty} p(y_1 \mid x_1^{+}) p(x_1^{+} e^{B_{11}} \mid \mathcal{J}_0, \mathcal{N}_0) dx_1^{+} \int_{0}^{\infty} p(x_{0} \mid \mathcal{J}_0, \mathcal{N}_0) dx_0 \right\} \left\{ p(x_{1}^{+} = t_f - t_1 \mid \mathcal{J}_1, \mathcal{N}_1) \right\}$$

$$= \left\{ \int_{0}^{\infty} p(y_1 \mid x_1^{+}) \frac{p(x_1^{+} e^{B_{11}} + t_1)}{\int_{0}^{\infty} p(x_{0} \mid \mathcal{J}_0, \mathcal{N}_0) dx_0} \int_{0}^{\infty} p(x_{0} \mid \mathcal{J}_0, \mathcal{N}_0) dx_0 \right\} \frac{p(y_1 \mid x_1^{+}) p(x_1^{+} e^{B_{11}} + t_1)}{\int_{0}^{\infty} p(y_1 \mid x_1^{+}) p(x_1^{+} e^{B_{11}} + t_1) dx_1^{+}}$$

$$= p(y_1 \mid x_1^{+}) p(x_1^{+} e^{B_{11}} + t_1)$$

$$= (ae^{B_{11}x_1^{+} + t_1})^\beta \eta^{\beta-1} e^{-\eta} \alpha^\beta \beta((x_1^{+} e^{B_{11}} + t_1))^\beta e^{-\alpha x_1^{+} e^{B_{11}}} (7-13)$$

138
For $n = 2$,

$$L(\theta, \mathcal{S}) \propto \left\{ \frac{p(x_1^+ | \mathcal{S}_0, \mathcal{S}_{N_0}) \int p(x_0 | \mathcal{S}_0, \mathcal{S}_{N_0}) dx_0}{p(x_2^+ = t_f - t_2 | \mathcal{S}_2, \mathcal{S}_{N_2})} \right\}^{*}$$

$$= \left\{ \int_0^\infty p(y_1 | x_1^+) \frac{p(x_1^+ e^{B_{N_1}} + t_1)}{p(x_0 | \mathcal{S}_0, \mathcal{S}_{N_0}) dx_0} \right\}^{*}$$

$$= \left\{ \frac{\int_0^\infty p(y_2 | x_2^+) \frac{p(x_2^+ e^{B_{N_2}} + t_2 - t_1 | \mathcal{S}_1, \mathcal{S}_{N_1})}{p(x_1^+ | \mathcal{S}_1, \mathcal{S}_{N_1}) dx_1^+} \right\}^{*}$$

$$= \int_0^\infty p(y_1 | x_1^+) p(x_1^+ e^{B_{N_1}} + t_1) dx_1^+ p(y_2 | x_2^+) p(x_2^+ e^{B_{N_2}} + t_2 - t_1 | \mathcal{S}_1, \mathcal{S}_{N_1})$$

$$= \int_0^\infty p(y_1 | x_1^+) p(x_1^+ e^{B_{N_1}} + t_1) dx_1^+ p(y_2 | x_2^+) p(x_1^+ e^{B_{N_1}} + t_1)$$

$$= p(y_2 | x_2^+) p(y_1 | x_1^+) p(x_1^+ e^{B_{N_1}} + t_1)$$
\[
\begin{align*}
    &= (ae^{b(x_i^++t_1)})^\eta \eta y_1^{\eta-1} e^{-(ae^{b(x_i^++t_1)})^\eta} (ae^{b(x_i^++t_1)})^\eta \eta y_1^{\eta-1} e^{-(ae^{b(x_i^++t_1)})^\eta} \\
    &= \alpha^\beta (x_i^+ e^{Bx_i} + t_1)^{\beta-1} e^{-(\alpha (x_i^+ e^{Bx_i} + t_1))^{\beta}}.
\end{align*}
\]

(7-14)

Generalizing this to the \( n \)th observation point, the likelihood function becomes

\[
L(\theta, \mathcal{S}) \propto (ae^{b(x_n^++t_n)})^\eta \eta y_n^{\eta-1} e^{-(ae^{b(x_n^++t_n)})^\eta} \alpha^\beta \beta(y_1(x_n^n))^{\beta-1} \prod_{i=2}^n \eta y_i^{\eta-1} (ae^{b\psi_i(x_n^n)})^\eta e^{-(ae^{b\psi_i(x_n^n)})^{\eta}}
\]

where

\[
\psi_j(x_n^n) = x_j^+ e^{Bx_j} + \sum_{k=j+1}^i (t_k - t_{k-1}) I_{k,i} e^{Bx_k} + t_j \quad j = 1, \ldots i
\]

and

\[
\begin{cases}
    I_{k,i} = 0 & \text{if } k > i \\
    I_{k,i} = 1 & \text{otherwise}
\end{cases}
\]

(7-15)

Hence, the log likelihood function of all observation points can be written as

\[
\log L(\theta, \mathcal{S}) = \eta \log (ae^{b(y_n^n)}) + \log \eta + (\eta - 1) \log y_n - (ae^{b(y_n^n)})^\eta + \beta \log \alpha + \log \beta + \log \psi_i(x_n^n) + \sum_{i=2}^n \log \eta + (\eta - 1) \log y_i^{-1} + \eta \log (ae^{b\psi_i(x_n^n)}) - (ae^{b\psi_i(x_n^n)})^\eta
\]

(7-16)

For the \( M \) datasets, the likelihood function is shown as

\[
L(\theta, \mathcal{S}) \propto \prod_{i=1}^M \left\{ (ae^{b(x_i^++t_i)})^\eta \eta y_i^{\eta-1} e^{-(ae^{b(x_i^++t_i)})^\eta} \alpha^\beta \beta(y_1(x_n^n))^{\beta-1} \prod_{i=2}^n \eta y_i^{\eta-1} (ae^{b\psi_i(x_n^n)})^\eta e^{-(ae^{b\psi_i(x_n^n)})^{\eta}} \right\}
\]

(7-17)
where

$$\psi_j(x_i^t) = x_i^t e^{\theta_{ij}} + \sum_{k=j+1}^i (t_k - t_{k-1}) I_{k,i} e^{\theta_{ij}} + t_j, \quad j = 1, \ldots, i$$

and

$$\begin{cases}
I_{k,i} = 0 & \text{if } k > i \\
I_{k,i} = 1 & \text{otherwise}
\end{cases}.$$

Then the log-likelihood for the $M$ datasets is written as

$$\log L(\theta, \mathcal{Z}_r) = \sum_{r=1}^M \left\{ \eta \log(ae^{b(x_i^r)}) + \log \eta + (\eta - 1) \log y_n - (ae^{b(x_i^r)})^q + \beta \log \alpha + \log \beta + (\beta - 1) \log(\psi_1(x_i^r)) + \sum_{r=2}^M \log \eta + (\eta - 1) \log y_{r-1} + \eta \log(\psi_r(x_i^r)) - (ae^{b(x_i^r)})^q \right\}$$

(7-18)

### 7.4 Fitting the Model to the Data

In order to fit the model to the data, we need to prepare data that is suitable for the model’s requirements. To organize the required reflective dataset, we used the previous analysis in Chapter 6 that yields the first principal component of total wear metal concentrations. For the responsive dataset, we use the original oil data values from oil performance and contaminant measurements to show their influences on the residual life. A decision needs to be made in order to determine the dimension of the responsive information. In our case, we have four variables from lubricant measurements, i.e. INS, KV, TBN and WC. INS measures the build-up of insoluble combustion-related debris and oxidation products.

High INS values will cause lacquer formation on hot surfaces, sticking of piston rings, and wear of the cylinder liner and bearing surfaces, which indicates a contamination of the system. Kinetic viscosity (KV) is the most important property of a lubricant. Increased oil viscosity will cause more wear of bearings and running surfaces, thus high KV indicates that the system has been contaminated. WC is a quantification of water contamination, where water in a lubricant will promote corrosion and oxidation. An increase again shows that the system has been contaminated. The total base number
(TBN) of a used lubricant is a measurement of its ability to neutralize acids. A low TBN will cause more corrosion and thus indicates oil contamination.

One way to reduce the size or the dimension of the input variable is by applying the PCA. However, it is notice that the measurement unit of each element in responsive data varies, which prevents us from carrying out the PCA. Hence, we decided to use all the information and assumed that all elements are independent of each other. We applied a concept called independent component analysis (ICA), for two main reasons; the first is to normalize the unit of each individual oil performance measurement, and the second is to identify the possibility of reducing the dimensions of the monitored variables for oil performance and contaminants as PCA is also used as a pre-processing step for ICA. Normalization of the measurement unit of the monitored variables is important, as each variable has its own unit and it was noted from the literature and discussions with persons familiar with similar types of data that these variables are very much correlated among themselves, (Hussin and Wang, 2006). Therefore, ICA can be used to calculate the contribution of each item of responsive data.

It is noted that we are applying PCA rather than ICA to the reflective data, even though ICA can also be used to reduce the dimensionality of the dataset, as all the elements in reflective data are measured on the same scale and are assumed to be correlated (Jolliffe, 1986). The literature also shows that the aim of PCA is variance maximization and dimension reduction (Fodor, 2002 ), while ICA is more suited to ensuring the statistical independence of the input variables (Lennon et al., 2001) and (Fodor, 2002 ).

### 7.4.1 Independent Component Analysis (ICA)

ICA of a random vector $Y$ consists of finding a linear transformation $\hat{S} = WY$ so that the components $\hat{S}_i$ are as independent as possible, in the sense of maximizing some function $F(\hat{s}_1, \ldots, \hat{s}_n)$ that measures independence. This is not very different from principal component analysis (PCA; see Chapter 6), used in finding the transformation $Z = UY$ so that elements of a new set of $z_i$ are linearly uncorrelated and the variances of $z_i$ are ranked from the first to the last. To begin the discussion of ICA, we can express the entire system of $n$ measured signals and $m$ observations as
We refer to each independent signal as $S$. Each observed signal $Y$ can be expressed as a linear combination of these independent signals. $A$ is an $m \times n$ mixing matrix that generates $Y$ from $S$. Hence, the goal of ICA, given the observation $Y$, is to calculate matrix $A$ and the set of independent $S$ values. In order to perform this task, our intention is to find a matrix $W$ (the inverse of $A$) such that, if it is applied to $Y$, it yields a dataset $\hat{S}$ that is as independent as possible and thus approaches the unknown signals $S$.

$$\hat{S} = WY : \begin{bmatrix} \hat{s}_1 \\ \vdots \\ \hat{s}_n \end{bmatrix} = \begin{bmatrix} w_{11} & \cdots & w_{1m} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nm} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$ (7-20)

It should be noted that each of these variables influences the others, so that applying ICA as a technique to separate linear mixed sources seems promising. Several approaches for making ICA estimations are available. In this research, the FastICA package developed for Matlab applications and designed by Aapo Hyvärinen at Helsinki University of Technology has been applied (Hyvärinen et al., 1997). Thus, for every set of data, the transformation is made from the original value of lubricant variables to a set of independent variables. These transformed data will be used for further analysis.

As noted above, ICA can reduce the dimensions of the variables, although this is not its main purpose. This analysis was conducted on our dataset, and while conducting the whitening process, the eigenvalues of the first principal component of uncorrelated variables showed a range of values from 55 to 95%. This strongly suggests that the dimensionality could not be reduced, hence we decided to use the entire set of variables in the model.
Having all the data required in hand, we first estimated the parameter values for our model. The 22 randomly selected dataset is again used for this purpose and the results are shown below in Table 7-1. The dataset chosen is similar to the data used in Chapter 6 for model comparison.

<table>
<thead>
<tr>
<th>( \hat{a} )</th>
<th>( \hat{b} )</th>
<th>( \hat{\eta} )</th>
<th>( \hat{B}_{INS} )</th>
<th>( \hat{B}_{KV} )</th>
<th>( \hat{B}_{TBN} )</th>
<th>( \hat{B}_{WC} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00106</td>
<td>0.00021</td>
<td>1.19712</td>
<td>0.03321</td>
<td>0.01672</td>
<td>0.03516</td>
<td>0.06309</td>
</tr>
</tbody>
</table>

Table 7-1: Parameter estimation for responsive and reflective variables

The variance and covariance matrix was subsequently calculated to provide an indication of the accuracy of the estimates. The variance and covariance matrix for the estimated parameters in shown in Table 7-2 below.

<table>
<thead>
<tr>
<th>( \hat{a} )</th>
<th>( \hat{b} )</th>
<th>( \hat{\eta} )</th>
<th>( \hat{B}_{INS} )</th>
<th>( \hat{B}_{KV} )</th>
<th>( \hat{B}_{TBN} )</th>
<th>( \hat{B}_{WC} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{a} )</td>
<td>3.809E-10</td>
<td>-3.42E-10</td>
<td>-2.38E-9</td>
<td>-2.34E-8</td>
<td>1.67E-8</td>
<td>-5.33E-10</td>
</tr>
<tr>
<td>( \hat{b} )</td>
<td>-3.42E-10</td>
<td>3.16E-10</td>
<td>-8.03E-8</td>
<td>-1.61E-8</td>
<td>-7.47E-8</td>
<td>-6.53E-8</td>
</tr>
<tr>
<td>( \hat{\eta} )</td>
<td>-2.38E-9</td>
<td>-8.03E-8</td>
<td>2.46E-4</td>
<td>1.24E-5</td>
<td>-2.44E-5</td>
<td>7.79E-6</td>
</tr>
<tr>
<td>( \hat{B}_{INS} )</td>
<td>-2.34E-8</td>
<td>-1.61E-8</td>
<td>1.24E-5</td>
<td>0.00234</td>
<td>-1.86E-4</td>
<td>1.35E-4</td>
</tr>
<tr>
<td>( \hat{B}_{KV} )</td>
<td>1.67E-8</td>
<td>-7.47E-8</td>
<td>-2.44E-5</td>
<td>-1.86E-4</td>
<td>0.00361</td>
<td>5.02E-4</td>
</tr>
<tr>
<td>( \hat{B}_{TBN} )</td>
<td>-5.33E-10</td>
<td>-6.53E-8</td>
<td>7.79E-6</td>
<td>1.35E-4</td>
<td>5.02E-4</td>
<td>0.00498</td>
</tr>
<tr>
<td>( \hat{B}_{WC} )</td>
<td>-7.11E-8</td>
<td>2.25E-9</td>
<td>4.35E-5</td>
<td>-2.25E-5</td>
<td>-7.35E-4</td>
<td>2.13E-3</td>
</tr>
</tbody>
</table>

Table 7-2: Variance and covariance of estimated parameters

The variances of the estimated parameters shown in Table 7-2 are relatively small for all parameters, which tells us something about variability around the mean. Thus, we can say that the parameter estimates were good in terms of variances. The covariance of each parameter is literally small, indicating that there are no relationships between estimated parameters.

Using the estimated parameters, we calculated our residual model \( P(x^* \mid \tilde{y}, N) \) using the remaining dataset. Figure 7-1 and Figure 7-2 below show two of the results for residual time distribution. The dotted straight line in the plot indicates the actual
residual life. Again, it is shown in both plots that the variance of the pdf becomes smaller as we have more information.

Figure 7-1: Case 1 – pdf and actual residual time with mixed monitoring information for engine 830001/28

Figure 7-2: Case 2 – pdf and actual residual time with mixed monitoring information for engine 830001/30
7.5 Model Comparison

To show the comparison of this new model with the model developed in Chapter 6, Figure 7-3 and Figure 7-4 below show the plots of distribution of residual life at the last monitoring point for each test dataset.

**Figure 7-3**: Case 1 – pdf residual time of engine 830001/28 at the last observation point

**Figure 7-4**: Case 2 – pdf residual time of engine 830001/30 at the last observation point
From these figures, it can be seen that the new model gives a new prediction that is significantly better than the previous one. Thus, it is concluded that the more condition-monitoring data is used, the better the model estimation becomes.

7.6 Summary

This chapter has presented a new development of conditional residual time prediction, in which the previous approach is enhanced with data that we assume will increase or decrease with the residual time. This was done by first assuming that the residual time is a function of observed reflective monitoring data, $y_i$. Then, we set up a relationship indicating how the residual time will change with the responsive monitoring data, $z_i$. Issues such as the possibility of the responsive variables being correlated with each other were solved by using ICA. By providing the data required by the model, we then estimated the model parameters using the maximum likelihood approach. The results from parameter estimation were sufficiently good to justify proceeding with fitting the model to the data. The results of this approach were satisfactory, showing better prediction than the previous model with the same dataset but without $z_i$. It is concluded that more information leads to a better prediction.
CHAPTER 8: A WEAR PREDICTION MODEL
BASED ON SPECTROMETRIC OIL ANALYSIS
PROGRAMME USED IN DIESEL ENGINES

8.1 Introduction

In a recent parallel study conducted by Wang (2006), the deterioration process of a system is presented by a generic term called “wear”. In general, the deterioration process is assumed to be described by discrete variables as shown in Chapter 3. Wang (2006) however, proposed that a continuous random variable could also be used to represent the deterioration process of a system and formulated its transitional probability. One of the advantages of this approach is that it eliminates the need of a threshold level of the wear that is commonly used in other wear prediction models (Pandey et al., 2005). In general, the model is formulated within a stochastic filtering approach and hidden Markov framework, which can be used to predict the current and future wear in a system given past observed information to date. Thus, having these predictions, it would enable us to develop appropriate maintenance decisions for a system subject to condition monitoring information.

In his recent work, Wang (2006) had used simulated data for the modelling development and later conducted a test for the model using an oil based monitoring dataset from aircraft engines. However, it has been reported that many factors could affect the wear of the system such as engine age, type of service, environmental conditions, engine metallurgy, etc., (Macian et al., 2003). Here, we have the opportunity to study the performance of this proposed wear prediction model by re-applying it to a different set of observed oil monitoring data. The monitoring information that is made available for this study are several sets of complete life data of diesel engines used in ships along with the history of their monitored SOAP information, see Chapter 6 for detail of the datasets.

The aim of the study is to test, validate and compare the wear prediction model as proposed and to see whether it could be used to generalize the deterioration process of oil based monitoring systems. A modelling methodology for the wear prediction model that is suited with the observed monitoring data will then be presented.
8.2 Modelling Development

This section outlines the assumptions and notation made for this study.

8.2.1 Notation

1. $t_i$ -- time of the $i$th monitoring check since new, $i = 1, 2, \ldots$.

2. $W_i$ -- a random variable that represents cumulative wear at time $t_i$, $W_i \in (W_{i-1}, 1)$, which is not directly observable and non-decreasing with $W_i$ of its realization.

3. $p(w_i | w_{i-1})$ -- the transition probability density function of the hidden Markov chain where $w_0 = 0$ denotes new condition and $w_i = 1$ denotes a failure. In this study it is assumed that the transition probability follows a generalized beta distribution with its probability distribution function given by

$$p(w_i | w_{i-1}) = \frac{(w_i - w_{i-1})^{i-1}(1 - w_i)^{q-1}}{B(p, q)(1 - w_{i-1})^{p+q-1}} \quad (8-1)$$

where

$$B(p, q) = \int_0^1 t^{p-1}(1 - t)^{q-1} dt \quad (8-2)$$

is the beta function.

4. $Y_i$ -- a random variable representing observed monitoring information with $y_i$ its observed value at time $t_i$. In this study, $y_i$ denotes the first principal component of the total metal concentration at time $t_i$, which is obtained by some transformation from original observed metal concentrations; see Chapter 6 for details of the transformation. Note that $Y_i$ could be a vector if several types of information are used.

5. The condition monitoring history obtained up to time $t_i$ is denoted by $\mathcal{Y}_i$, where $\mathcal{Y}_i = \{y_1, y_2, \ldots, y_i\}$

6. $Y_i | W_i$ follows a distribution with a probability density function, $p(y_i | w_i)$, which is used to describe the relationship between the observed monitoring data and the cumulative wear at time $t_i$. 
8.2.2 Assumptions

The assumptions made were as follows,

1. Condition monitoring takes place at time \( t_i , i = 1, \ldots, n \), where the intervals between consecutive monitoring checks are equal.
2. The equipment considered here is a single-component system subject to one dominant failure mode, i.e. wear related failure.
3. The observed monitoring data \( y_i \) is a function of wear but not vice versa.

This will lead us to formulate the model, which shall be discussed in the subsequent section.

8.3 Model Formulation

Wear usually is a process of deterioration, resulting from tear, fatigue and friction. When studying any failure where wear is suspected or in the process of occurring, its quantification is complex and difficult unless the wear is directly observed such as in the case of a brake pad. In previous wear related studies, age is the only factor that is used to calculate the wear and furthermore there is a need to define the threshold level used to describe the failure event. An example of such studies is presented by Pandey et al. (2005) where the cumulative deterioration or wear at time \( t \) follows a gamma distribution with the shape function, \( \lambda(t) > 0 \), and the scale parameter, \( \beta \), is a constant.

This model has a drawback in modelling wear since the development of wear is not only due to age but it could also depend on other factors such as the operational variables/process which could be observed using monitoring tools. Furthermore, the determination of a wear threshold level is important and needs to be done carefully. Wang (2006) proposed a model that the probability of the cumulative wear at \( t_i \) given its current and previous observed monitoring data \( \mathcal{Z}_i \), \( p(w_i | \mathcal{Z}_i) \) could be computed without any threshold value and provides the impetus for a more thorough investigation on this model.

According to Wang (2006), this method can be achieved by using a recursive filtering equation, which is written as
The key idea of equation (8-3) is sequentially to predict the cumulative wear of a dynamic system $w_i$ conditional on monitored data, $\mathcal{Z}_t$. Using this recursive scheme, it requires a few important formulations. Firstly, the function $p(w_i | w_{i-1})$, which describes the system dynamics, i.e. how the state/wear of the system changes, needs to be known. Hence, the theory of Markov chains is presented in this model. To do this, a beta distribution has been chosen as it has a finite range from 0 to 1, where a new condition is considered to have a wear of zero as $w_0 = 0$ and $w_1 = 1$ denotes a failure. The formulation of this transition probability is established in equations (8-1) and (8-2) above.

The second term that needs to be acknowledged is $p(y_i | w_i)$ which describes the relationship of having observation $y_i$ given that the wear is $w_i$. One of the ways of establishing this relationship is by using a probabilistic distribution as shown in (Wang et al., 2000). This modelling of $p(y_i | w_i)$ should reflect the nature of the wear given monitoring information. The detail for this formulation is explained in the next section.

It is noted that an analytical solution for $p(w_i | \mathcal{Z}_i)$ does not exist as mentioned by Wang (2006), thus any kind of numerical approach has to be used. Therefore, as suggested, two numerical approaches, which can be used to overcome the problem, namely approximated grid method and particle filtering, are applied in this study.

### 8.4 Model Approximations

In this section, we will study approximation methods for solving equation (8-3). We will use two algorithms proposed by Arulampalam et al. (2001) namely approximated grid method and particle filtering. The main difference between these algorithms is the way the grid is defined. In our case, the grid represents the state of the system.
Approximate grid, as the name reveals, approximates the posterior density by a set of points on a predefined grid. On the other hand, particle filter is also approximation by a set of points; however the grid is chosen in a stochastic way. The same algorithms were also used in Chapter 4 but they are targeted to solve the discrete case of the underlying state of a system to identify the initiation of a random defect.

8.4.1 Approximated Grid Method

In Arulampalam et al. (2001), the continuous state space can be used within a grid-based method to approximate the posterior density by composing it into \( N \) ‘cells’, where \( \{w^k_i : k = 1,2,\ldots,N\} \). With this approach, equation (8-3), can be rewritten as

\[
p(w_i | \mathcal{Z}_i) \approx \sum_{k=1}^{N} v^k_{ji} \delta(w_i - w^k_i)
\]  

(8-4)

where

\( \delta(\cdot) \) is the Dirac delta function and \( v^k_{ji} \) is given by

\[
v^k_{ji} = \frac{v^k_{j-1,i} \int_{w_{i}^{k},w_{i+1}^{k}} p(y_i | w_i)dw_i}{\sum_{j=1}^{N} v^j_{j-1,i} \int_{w_{i}^{j},w_{i+1}^{j}} p(y_i | w_i)dw_i}
\]  

(8-5)

and \( v^k_{j-1,i} \) is given by

\[
v^k_{j-1,i} = \sum_{j=1}^{N} v^j_{j-1,i} \int_{w_{i-1}^{j},w_{i}^{j}} p(w_i | \overline{w}_{i-1}^j)dw_i
\]  

(8-6)

where \( \overline{w}_{i-1}^j \) denotes the centre of the \( j \)th cell at time index \( i-1 \). The integration in equation (8-6) is reduced by another approximation to simplify the computation. This can be done by assuming that the weight is computed at the centre of the cell corresponding to \( v^k_{j} \) where it can be written as \( \overline{w}_i^j = (w_i^k - w^k_{i-1}) / 2 \). Thus, equation (8-5) and (8-6) can be rewritten as
\[
\begin{align*}
    v_{i|i-1}^k & \equiv \sum_{j=1}^{N} w_{i|i-1}^j P(\overline{w}_i^k | \overline{w}_i^{j}) \\
    v_{i|i}^k & \equiv \frac{v_{i|i-1}^k P(y_i | \overline{w}_i^k)}{\sum_{j=1}^{N} v_{i|i-1}^j P(y_i | \overline{w}_i^j)}
\end{align*}
\] (8-7)

In order to have a good approximation to the continuous state, \( N \) must be sufficiently large. However, in our case, depending on the parameters value \( p \) and \( q \), singularity may occur particularly when \( w_i \) approaches \( w_{i-1} \) or 1. One might consider numerical methods where the current value of the pdf. is divided by their cumulative probability, to overcome these limitations; hence the approximation can be calculated.

### 8.4.2 Particle Filtering

The overview of particle filtering has already been explained in Chapter 5 but it is applied to two states of a discrete case. Generally, the particle filter approximates the posterior density by using sets of state particles specified by \( X_i = \{x_i^k, w_i^k \}_{k=1}^{N} \), where \( N \) is the number of particles denoted by \( \{x_i^k\}_{k=1}^{N} \) and \( \{w_i^k\}_{k=1}^{N} \) are the associated weights. In the case of continuous states, \( N \) is chosen to have a large value, (e.g., \( N = 100 \)). Here, we present an algorithm for the implementation of SIS (sub-optimal) to the continuous case.

#### Algorithm 1: SIS (prior-importance)

For time steps \( i = 0,1,2,\ldots \) do the following:

**Step 1:** Start with the non-parametric prior distribution of the state \( w_0 \sim \{w_0^1, w_0^2, \ldots, w_0^N\} \), \( v_0 \sim \{v_0^1, v_0^2, \ldots, v_0^N\} \) represented by a set of \( N \) discrete states.

**Step 2:** Generate \( w_i^k \sim p(w_i \mid w_i^{k-1}) \), hence \( w_i \sim \{w_i^1, w_i^2, \ldots, w_i^N\} \).

**Step 3:** Assign the particle weight for each state, \( v_i^k \), where \( v_i^k = v_{i-1}^k P(y_i \mid w_i^k) \).

**Step 4:** Normalise the importance weights, \( v_i = \frac{v_i^k}{\sum_{k=1}^{N} v_i^k} \).
Step 5: Calculate $p(w_i | \mathcal{Z}_i) \approx \sum_{i=1}^{N} v_i^* \delta(w_i - w_i^*)$.

Step 6: Repeat step 2 for $i = 1, 2, t_n$, where $t_n$ is the last monitoring point.

End

It has already been explained in Chapter 5 that degeneracy is a common problem found in particle filtering. The solution to this problem is to use the re-sampling step, mainly to discard particles with small weights and multiply particles with large weights.

Unfortunately, the re-sampling step also introduces other problems. Theoretically, a particle that has large weights that will be selected many times could give rise to the problem known as sample impoverishment (Arulampalam et al., 2001), which implies that the effective number of samples is reduced. To explain this problem suppose we have a particle in a particular state before the probability of that state can be evaluated. If the state has no particles in it, the assumption is that its probability of being the true state of the system is zero. Without a particle in a certain state, that state will continue to be given zero probability of having occurred and this implies that the particles will not reflect the true density.

The simplest solution to the sample impoverishment problem is to increase the number of particles being used but this is probably unrealistic. In the literature, several ideas have been proposed to reduce the problem (Schon, 2003). One such idea is referred as roughening (Arulampalam et al., 2001) or jittering (Fearnhead, 1998), which introduces an additional noise to make the particles different from each other. With all the negative effects of re-sampling steps, one needs to provide a measure for degeneracy and avoid re-sampling if it is not necessary. Such measurement is already introduced in Chapter 5.

8.5 Parameter Estimation

In order to estimate the parameters of equation (8-4) the maximum likelihood approach is used. A likelihood function that explains the observed data is needed in order to utilize this method. To give details, suppose we have observed $y_i, i = 1, 2, \ldots, n$ where $n$ is the last observation, and the system failed at time $t_i, t_i > t_n$. Thus the likelihood function is given by
\[ L(\theta, \mathcal{F}) \propto \prod_{i=1}^{n} \{ p(y_i \mid \mathcal{F}_{i-1}) p(W_i < 0.99 \mid \mathcal{F}_i) p(W_i > 0.99 \mid \mathcal{F}_n) \} \quad (8-9) \]

where

\[ p(w_i) \text{ in equation (8-9) above is established by using the chain rule and shown as} \]

\[ p(w_i) = \int_0^w p(w_i \mid w_{i-1}) p(w_{i-1}) \, dw_{i-1} \quad (8-10) \]

and \( p(y_i \mid \mathcal{F}_{i-1}) \) is given by

\[
\begin{align*}
&= \int_0^1 p(y_i \mid w_i) p(w_i \mid \mathcal{F}_{i-1}) \, dw_i \\
&= \int_0^1 p(y_i \mid w_i) \int_0^w p(w_i \mid w_{i-1}) p(w_{i-1} \mid \mathcal{F}_{i-1}) \, dw_{i-1} \, dw_i \quad (8-11)
\end{align*}
\]

In equation (8-9), \( p(W_i < 0.99 \mid \mathcal{F}_i) \) represents the observed information that the system has not failed at time \( t_i \) and \( p(W_i > 0.99 \mid \mathcal{F}_n) \) is used to represent the failure probability at time \( t \) since \( p(W_i = 1 \mid \mathcal{F}_n) \) is not available due to the continuous nature of the beta distribution. For numerical convenience, the log of the likelihood is used, that is

\[
\text{Log}L(\theta, \mathcal{F}) = \sum_{i=1}^{n} \{ \log(p(y_i \mid \mathcal{F}_{i-1})) + \log(p(W_i < 0.99 \mid \mathcal{F}_i)) \} + \log(p(W_i > 0.99 \mid \mathcal{F}_n))
\quad (8-12)
\]

Maximizing equation (8-12) may give the estimated values that we wanted. However, it was observed this is not a clear-cut process as the cross correlation produced between the parameters within \( p(w_i \mid w_{i-1}) \) and those in \( p(y_i \mid w_i) \) may give us problems if both of them were estimated together. It has also been noted that the hidden Markov process of \( w_i \) is independent from \( y_i \); therefore, this allowed us to estimate the unknowns in a two-step process. The first step is to maximize
\[
\log L_1(\theta_1; t) = \sum_{i=1}^n \log(p(W_i < 0.99)) + \log(p(W_i > 0.99))
\] (8-13)

to get the estimated parameters values for \( p \) and \( q \), assuming these values are known to maximize

\[
\log L_2(\theta_2, \mathcal{X}_i) = \sum_{i=1}^n \log(p(y_i | \mathcal{X}_{i-1}))
\] (8-14)

and estimate the parameters within \( p(y_i | w_i) \). However, closed-form solution is not available and numerical schemes proposed in the sections 8.2 have to be applied.

### 8.6 Numerical Results

In order to apply to the oil based monitoring data used in the diesel engines detailed in Chapter 6, we used the information of wear metal elements observed in the SOAP samples. The process of cleaning and transforming this information into a format requested by the model has already been discussed in Chapter 6. To estimate \( p \) and \( q \), several sets of failure information were used. This requires some explanations. The wear increment from time \( t_{i-1} \) to \( t_i \) until it fails follows a beta distribution with \( p \) and \( q \) being parameters that must be greater than zero. At every monitoring check, one needs to know whether the machine fails or not, thus the likelihood of wear that explains this situation needs to be calculated. We coded equation (8-13) in FORTRAN95 and used their NAG optimisation routines, namely E04JYF to obtain the values of the estimated \( p \) and \( q \). Table 8-1 and 8-2 below summarize our parameter estimates for \( p \) and \( q \) with their variance and covariance matrix.

<table>
<thead>
<tr>
<th>( \hat{p} )</th>
<th>( \hat{q} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1968</td>
<td>5.6350</td>
</tr>
</tbody>
</table>

Table 8-1: Estimated parameters for beta distribution
Table 8-2: Variances and covariance for the estimated parameters

<table>
<thead>
<tr>
<th></th>
<th>( p )</th>
<th>( q )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p )</td>
<td>0.00021</td>
<td>0.0000014</td>
</tr>
<tr>
<td>( q )</td>
<td>0.0000014</td>
<td>0.034</td>
</tr>
</tbody>
</table>

The variances from this estimation are small implies that our estimation is good. Next, to estimate the parameters in equation (8-14), it is important to reflect the nature of the deterioration process (wear) according to the measured condition monitoring information. For the sake of clarity, consider the transformation of SOAP information from diesel engines, which shows a linear relationship between the first principal component of the total metal concentration and the time of failure, as shown in Figure 6-8 in Chapter 6.

Equation (8-14) describes the relationship or likelihood having observation \( y_i \) given that the state of the system is \( w_i \), which can be measured via a probability distribution function. In this study, a Weibull distribution to represent \( p(y_i | w_i) \) has been considered, but other distributions may also be used. The relationship can be written as a three parameters Weibull distribution as

\[
p(y_i | w_i) = \alpha_{w_i} \beta (\alpha_{w_i} (y_i - y_{i-1}))^{\beta-1} e^{-(\alpha_{w_i} (y_i - y_{i-1}))^\beta}
\] (8-15)

where \( y_{i-1} \) is served as the location parameter. More explanation is seen as necessary to show the relationship. The observed condition information increment from time \( t_{i-1} \) to \( t_i \), \( (y_i - y_{i-1}) \), is a non-negative quantity, which has a Weibull distribution with shape parameter \( \alpha_{w_i} \) and scale parameter \( \beta \). To model the relationship between \( E(Y_i) \) and \( w_i \) within time index \( t_i \), let the scale parameter \( \alpha_{w_i} \) as

\[
\alpha_{w_i} = a(w_i, t_i) = \frac{1 + At_i}{B + Cw_i}
\] (8-16)

This is motivated by Wang (2006), which used a similar set-up, but with four parameters. It is noted that the original set-up is found by observing \( y_i \) in aircraft
engine oil samples, where the slope of $y_i$ decreases, which indicates that the increment of $y_i$ is getting smaller as $t_i$ increases, but at the same time $t_i$, $y_i$ increases with $w_i$. However, it has been tested that, if we were given the right values of $A, B$ and $C$, this set-up can produce a constant increment $y_i$ as both time $t_i$ and $w_i$ increases, which produces a linear plot as shown Figure 6-8 in Chapter 6. The set-up revealed two important facts. The first is that the proportional of variance of $y_i$ is not increasing in time $t_i$ and the second is that $y_i$ increases in $w_i$.

To estimate the parameters $A, B, C$ and $\beta$ we coded equation (8-14) in FORTRAN95 and used the same NAG optimisation routines to get the estimated values of these parameters. However, for this estimation, the routines used did not converge with two possible causes. The first may be that equation (8-14) is not correctly formulated and the second may be the likelihood function produces many local minima, which resulted in NAG optimisations routine being unable to provide a solution. Equation (8-14) has been validated in Wang (2006), so our attention turns to the second problem. One way of solving the second type of problem is to use other alternative optimisation approaches. Genetic algorithm (GA) is one of the potential approaches that we would like to consider as it is claimed that GA is the most appropriate for complex non-linear models where location of the global optimum is a difficult task (Mardle and Pascoe, 1999).

GA follows the concept of an evolution solution by stochastically developing generations of solution populations using given fitness statistics of an objective function. This objective function may not contain the ‘nice’ properties such as continuity, differentiability and satisfaction of the Lipschitz Condition, (Houck et al., 1995) which in turn lets the GA not guarantee optimality. However, a GA solution is likely to be close to the global optimum and tends to be computationally intensive. Good news to us is that most GA software was readily available in the public domain, which can be implemented in solving our problem. In our case, a MATLAB tool box developed by Houck et al. (1995) has been selected to implement the GA algorithms. In details we coded equation (8-14) into M files and linked it with the main GA routine described above. The data that we used here is the monitoring data and since we have several sets
of data from the oil analysis, we used half of them to estimate the parameter and the remainder will be used to test the methodology. The GA routines worked well and produced the estimated parameters as shown in Table 8-3 below.

<table>
<thead>
<tr>
<th>( \hat{A} )</th>
<th>( \hat{B} )</th>
<th>( \hat{C} )</th>
<th>( \hat{\beta} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00000099</td>
<td>697.6646</td>
<td>0.1000</td>
<td>1.69181</td>
</tr>
</tbody>
</table>

Table 8-3: Estimated parameter values from data

The point of interest now is how well the model fits the data or in other words, how the model can be validated. One of the ways to find this out is by looking at the graphical figures that illustrate the actual \( y_i \) that we observed and compared with the expected \( \tilde{y}_i \) that can be calculated using all the estimated parameters. To generate the expected \( \tilde{y}_i \), we could simulate \( w_i \), a wear process using the beta distribution in equation (8-13) with the estimated \( p \) and \( q \). Then by using the simulated \( w_i \), we run another simulation to generate the signal process and calculate \( \tilde{y}_i \) according to our estimated parameters \( A, B, C \) and \( \beta \).

However, the general form of the probability distribution of beta distribution in FORTRAN95 routines is expressed in terms of the standard distribution, where \( p = 0 \) and \( q = 1 \), which is not suited with the equation used in (8-4). Hence we need to have some transformation while running the simulation. This can be done by transforming the value of random number from a standard beta distribution to the beta distribution in equation (8-1) following the algorithm in Figure 8-1 below.
Since \( p(y_{ij} | w_i) \) follows a Weibull distribution, the cumulative density function of a Weibull random variable, \( F(\Delta y) \) is given by

\[
F(\Delta y) = 1 - e^{-(\alpha \Delta y)^\beta}
\]  

(8-17)

After some manipulation,

\[
\Delta y = \frac{(-\log(1 - \text{rnd}))^{\frac{1}{\beta}}}{\alpha}
\]  

(8-18)

where \( \Delta y \) generated using this approach will follow a Weibull distribution.

We presented \( \tilde{y}_{ij} \) and their actual condition monitoring obtained from our data, which is randomly selected from engines 830000/5 and 901001/1. The result of this comparison is shown in the Figure 8-2 and Figure 8-3 below.
It can be seen from both figures that using estimated parameters produced results that significantly match our observed condition monitoring information. Numerical testing can also be carried out to test the relationship of the actual and the simulated results. Such relationship can be found with the Pearson product-moment correlation coefficient. The $R^2$ returns the square of the Pearson product moment correlation coefficient, which consists of a dimensionless index that ranges from -1 to 1 and reflects the extent of a linear relationship between two data sets. Once we have learned the correlation coefficient $R^2$, we need to test the significance of $R^2$. To do this, we need to set the critical alpha level that defines the error rate that we are willing to accept (normally 0.05 or 0.01) and it will compare with the same critical value from Pearson’s
correlation coefficient table. If our calculated $R^2$ statistic is higher than the critical value from the table, we could say our relationship is significant.

Next, we shown the value of the $R^2$ for comparison between observed $y_i$ and simulated $\tilde{y}_i$ and calculated the significance level at 5% and 1%. **Table 8-4** shown the relationship of our result.

<table>
<thead>
<tr>
<th></th>
<th>901001/1</th>
<th>830000/5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculate $R^2$</td>
<td>0.99956</td>
<td>0.99991</td>
</tr>
<tr>
<td>$R^2_{0.05, df}$</td>
<td>0.195</td>
<td>0.195</td>
</tr>
<tr>
<td>$R^2_{0.01, df}$</td>
<td>0.254</td>
<td>0.254</td>
</tr>
</tbody>
</table>

**Table 8-4: Null hypothesis for Pearson product-moment correlation coefficient**

The absolute value of our correlation coefficient is nearly to 0.999 in both cases, which is above the Pearson product-moment critical value that is 0.195 when $\alpha = 0.05$ and 0.254 when $\alpha = 0.01$. We reject our null hypothesis (there is no relationship) and retain the alternative hypothesis that there is a statistically significant relationship between the simulated and actual condition monitoring data.

The result of the graphical and numerical relationship has clearly indicated the strength of the model. With this result, it convinces us to calculate the current and the predicted wear distribution, which can be utilized in recommending maintenance actions. In the case of calculating the current wear distribution given the observed monitoring data, we would like to test the difference if no monitoring data is sought. Based upon the estimated parameter values, we use approximated grid approach to calculate $p(w_i)$ and $p(w_i \mid \mathcal{Z})$. The result is shown in Figure 8-4 and Figure 8-5 below.
Based upon the failure data, we calculate \( p(w_i) \) that follows a beta distribution using \( \hat{p} \) and \( \hat{q} \) but without observed monitoring data. It demonstrates a trend that the wear increases rapidly in the early stage of the engine life and slows down towards the end of the life.

![Figure 8-4: \( p(w_i) \) based on failure data only](image)

![Figure 8-5: \( p(w_i \mid \mathcal{Z}_i) \) based on failure data and condition monitoring information of engine 830001/28](image)
A similar pattern is also produced by \( p(w_i \mid \mathcal{I}_i) \) which shows the distribution of wear is narrower than \( p(w_i) \), which implies more accurate prediction. This is because with more available information, less variance is seen in our model. Using this model, it has shown that we are able to calculate the cumulative wear probability density at any time \( t_i \).

At this stage, we are also keen to find out if there is any relationship between the beta wear model and the residual time model proposed in Chapter 6 using the same data sets. This is mainly because that each approach uses different types of models but has a similar aim to determine the state of the system and ideally they should produce similar results. Hence a comparison study needs to be carried out to see if there is a significant difference between these two approaches. To do that, we need to transform both models into another measure that will show the similarity they share. One possible way of doing this is by determining the \( k \) -steps-ahead prediction density of failure, conditional on the available information \( \mathcal{I}_i \) at each time \( t_i \).

Using a beta wear model, the conditional probability of a failure in \((t_{i+k-1}, t_{i+k})\) could have the following form that calculates the failure probability density in \( t_{i+k} \) given it has not failed at \( t_{i+k-1} \) as

\[
p(w_{i+k} \mid \mathcal{I}_i, w_{i+k-1} < 0.99) = \frac{\int_0^{0.99} p(w_{i+k} \mid w_{i+k-1}, \mathcal{I}_i) p(w_{i+k-1} \mid \mathcal{I}_i) dw_{i+k-1}}{\int_0^{0.99} p(w_{i+k-1} \mid \mathcal{I}_i) dw_{i+k-1}}, k = 1, 2, \ldots
\]

(8-19)

Hence, the probability of a failure at time \( t_{i+k} \) can be established by using equation (8-19), which can be calculated as

\[
P(W_{i+k} > 0.99 \mid W_{i+k-1} < 0.99, \mathcal{I}_i) = \int_{0.99}^{1} p(w_{i+k} \mid \mathcal{I}_i, w_{i+k-1} < 0.99) dw_{i+k}
\]

(8-20)
To calculate the conditional failure probability for the residual time model in \((t_{i+k-1}, t_{i+k})\), given that it has survived at \(t_{i+k-1}\) has the following form

\[
P(t_{i+k-1} < X_i \leq t_{i+k} \mid X \geq t_{i+k-1}, \mathcal{Y}_i) = \frac{\int_{t_{i+k-1}}^{t_{i+k}} p(x_i \mid \mathcal{Y}_i) dx_i}{1 - \int_0^{t_{i+k-1}} p(x_i \mid \mathcal{Y}_i) dx_i}
\] (8-21)

Using the data from engine 830001/28, we calculated the conditional failure probability from both approaches with \(k = 1\) prediction step.

<table>
<thead>
<tr>
<th>(i)</th>
<th>(P(W_{i+k} &gt; 0.99 \mid W_{i+k-1} &lt; 0.99, \mathcal{Y}_i))</th>
<th>(P(t_{i+k-1} &lt; X_i \leq t_{i+k} \mid X \geq t_{i+k-1}, \mathcal{Y}_i))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0015045</td>
<td>0.00165013</td>
</tr>
<tr>
<td>10</td>
<td>0.0015126</td>
<td>0.00162982</td>
</tr>
<tr>
<td>70</td>
<td>0.0015209</td>
<td>0.00188199</td>
</tr>
<tr>
<td>96</td>
<td>0.0015210</td>
<td>0.00179508</td>
</tr>
</tbody>
</table>

**Table 8-5: Conditional failure probability from a beta wear model and the residual time model**

By observing **Table 8-5**, it can be concluded that the conditional failure probability using both approaches produces similar result. This implies that the system state/condition can be described by many measures, but they should provide similar failure predictions provided they are formulated correctly.

### 8.7 Summary

In conclusion, the work covered in this chapter details an application of a wear prediction model based on oil monitoring information used in diesel engines. The model is formulated within a stochastic filtering and hidden Markov framework, which is able to predict the current and future wear. In this study, the underlying state is assumed to have a continuous value. This is done by using a beta distribution to represent the dynamics of the underlying state of the system. The relationship between the wear and observed monitoring information is modelled by a probabilistic distribution where the scale parameter of the distribution of the observed monitoring information is a function
of the time and wear. Particular attentions have been devoted to numerical solutions as the analytical solution cannot be found. Two approximation approaches, namely approximation grid based and particle filtering, are used. The model proposed has been fitted to several sets of SOAP data for testing and validation. Also, an attempt to investigate the relationship between the beta wear model and the residual time model develop in Chapter 6 has come to a conclusion that both models produced similar results.
9  CHAPTER 9: CONCLUSION, DISCUSSION AND FUTURE RESEARCH

9.1  Conclusion of the Research

The primary objective of this thesis is to substantiate a maintenance decision-making by means of condition monitoring data. Thus, the thesis approached this idea by compiling the literatures of condition monitoring modelling and its effect on maintenance decision-making. Several condition-monitoring techniques have been discussed and all of them have concluded with the same question, that is, if the observed parameter is deviated from its normal value is it sure to know that something wrong has happened which could lead to a failure? Since the deterioration process is hidden and the monitored observations can only tell us something indirectly with noise, should we carry out maintenance actions now or wait for an opportunity window? Hence, it comes to a question on how we in a cost effective way, should capitalize the observed condition monitoring data to carry out maintenance. Analysis of the literature discovers that there exist a few models that could be used to explain the relationship between the observed monitoring information and the item/system condition. Then by using this information, a model for maintenance decision-making is developed.

In detail, the development of a model for maintenance decision-making consists of modelling the deterioration process and a decision strategy. In our study, a deterioration process model describes the condition of the system based on the age and the observed condition monitoring data. Also, we the deterioration process is modelled by using a concept called hidden Markov model approach, which assumes that an observed monitoring data could be used as a function of the underlying state or condition of the equipment. In developing these models, we used discrete and continuous measure to define the underlying state before it can be used to predict the condition of the item/system. Two types of condition monitoring information is used namely vibration and oil analysis data to show the examples of the modelling ideas. It shown that, the model solution is complex and time
consuming particularly when the underlying state is continuous, hence we explore approximation approaches namely grid based method and particle filtering. In short, here we demonstrate several important conclusions from the study.

1. Defining the underlying state of the system and their relationship with the condition monitoring towards a failure is important to quantify the condition of the system. If different variables are used to define the underlying state of the system, they should produce similar outcome in predicting the failure.

2. This research proves that monitoring information plays an important role in prediction the condition of the item/system concerned.

3. It is a crucial to explore the observed condition monitoring data before fitting it into the model. Any presence of trends or other patterns in data can provide valuable information in the study.

4. A simulation approach is useful to help us in establishing and testing a model before it can be applied with a real data.

5. Approximation by numerical techniques is used to overcome problems exist in the analytical form. It provides us with a balance between result and computational effort to gain the result.

6. This research also proves that subjective information is helpful while facing non-identifiable parameters estimation.

In the next section, the contributions of this study are presented.

9.2 Contributions of the Research

The aim of this thesis is to address the need of modelling to support maintenance decision-making given the condition monitoring data. For this purpose, we reveal that modelling
maintenance decisions based upon condition monitoring data is under explored. As the problem is addressed, our aim is to investigate how previous techniques are used to approach similar problems. One of the techniques that pay us a lot of interest is CRT modelling. In Chapter 3, a new idea for CRT model is presented. Here, we assumed that the CRT is no longer a continuous measure but it could take a set of discrete values, which could explain the state of the system. We consider this matter because in practice, after the monitoring took place, one would like to know the state of the system such as normal, defect or failed as it is easy to understand and then performs actions as necessary. Since each state can move to another state randomly, we formulated the transition probability that best describes the Markov process of our model. It is shown that the transition from one state to another is a time dependent. Using this model, we developed a model to identify the initiation of a random defect based upon the simulated data that was motivated from vibration monitoring.

In Chapter 4, the model developed in Chapter 3 is tested and validated with actual data obtained from a study of six rolling elements bearing (Wang, 2002). Also, an attempt to compare the result that we produce based upon the model developed in Chapter 3 with another study (Zhang, 2004) using the same data is carried out. In Chapter 5, we noticed that calculating the ‘hidden’ state of the system, as shown in our model in Chapter 3 is complex and time-consuming. Therefore, as a contribution to this thesis, Chapter 5 proposed the numerical approximation solutions for the model developed in Chapter 3, namely a grid based and particle-filtering approach. Both approaches are widely used to solve a problem related to Bayesian filtering approach, which has a similar representation with our Hidden Markov Model. In addition to this contribution, there is no evidence that both approaches have been used in any study in maintenance or reliability tasks.

In Chapter 6, a huge amount of oil analysis data from diesel engines is obtained, which required a lot of effort to clean, analyse and organize the data for our models. The contribution in this chapter is mainly on cleaning, analysing and organizing the data that need to be used in the CRT model. Issues such as approximate the complete data from incomplete data and how we regularize the monitoring interval from irregulars monitoring
interval is discussed. It is noted that, a technique called principal component analysis is used to reduce the dimension of the data input.

In Chapter 7, we extend the model developed in Chapter 6 with a few new extensions as we embed lubricant and contaminant data as responsive factors that could increase or decrease the residual time. The contribution of this chapter is the extension of the model and organizing data for model input. We applied independent component analysis to the dataset due to the fact that each element from the responsive dataset does influence the residual time, but correlated to each other.

As we mentioned in the thesis, residual time is only a measure used to describe the deterioration process of an item. In Chapter 8, another measure that we believe to be appropriate in describing the deterioration process is introduced. A general wear deterioration process is presented and the transition probability of this continuous state is modelled by a beta distribution. An attempt to investigate the relationship between the wear model and residual time model is carried out. The analysis concludes that both models produced similar results. One critical advantage of this model is that the threshold level normally needed is no longer used.

### 9.3 Future Research and Other Issues

During the course of this research, a few ideas have been identified as potential future research areas. It is noted that models developed in this research are solely based on condition monitoring data, which proved to be useful to predict the residual time of the item. However, many other types of information collected during the production process at the plant can also be useful such as maintenance actions, process control information, failure mode etc. In this thesis, it is assumed that the failure is known (using objective or subjective data) and this information is used in parallel with condition monitoring parameters to predict the residual time, while the maintenance actions and other effects are left out to simplify the modelling work. In reality, maintenance actions play an important role to influence the residual time of an item and this need to be investigated on how to
incorporate them into the residual time model or other models that were developed in this thesis.

Secondly, most of the failures that we used are assumed to be from one general failure mode. This is because the information that we have is limited only to the general failure mode of the system without any extra information about the cause of the failure element in the system. If this information is recorded properly, the failure modes could be detailed out into a specific failure mode and if it is used in our model, the prediction result may be better.

Thirdly, condition-monitoring data, gives us a lots of information regarding the condition of the system, see Chapter 6. This observable fact poses a question on how to choose the right and useful data or should we use all the data observed? This thesis discussed a technique called principal components analysis that aims to reduce the dimension of the data and at the same time retain as much of the actual information. It is noted that several other approaches could be used to overcome similar problems, hence testing other method with the data and finding a comparison among these approaches is a task that worth doing.

In Chapter 7, we extend our model by incorporating other variables that we assume will influence the residual time that we want to predict. These variables may be correlated, so we carried out an independent component analysis used to separate those variables. It is noted that, to reduce the dimension and correlation of the data, would simplify the modelling process, however we could also lose some valuable information. In the future, independent component analysis should also be tested with the data and the size of the independent components that might be used in the model can also be identified.

Finally, there is a trend that a group of condition monitoring techniques can be used together to determine the machine condition (Troyer, 1998) and (Smith, 2001). An attempt should be made to investigate the relationship of multi-sensor information with the underlying state of the system. Furthermore, if such relationship existed, the question now
is how we could use this valuable information to get a better understanding of the system and produce a more accurate decision.
REFERENCES


