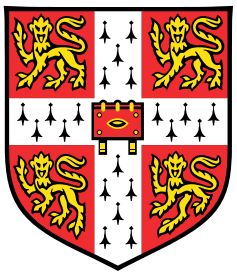
Integrated Workload Allocation and Condition-based Maintenance Threshold Optimisation



Hao Li

Department of Engineering
University of Cambridge

This thesis is submitted for the degree of
Doctor of Philosophy

St Catharine's College June 2019

Declaration

I hereby declare that, the contents of this dissertation, entitled *Integrated Workload Allocation and Condition-based Maintenance Threshold Optimisation*, are the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation contains fewer than 65,000 words including appendices, footnotes, tables and equations and has fewer than 150 figures. It does not exceed the prescribed word limit for the Degree Committee of the Engineering Department.

Hao Li
June 2019

Abstract

Title: Integrated Workload Allocation and Condition-based Maintenance Threshold Optimisation

By: Hao Li

Effective asset management is considered key to reducing total costs of asset ownership while enhancing machine availability, guaranteeing security, and increasing productivity. Amongst all the activities involved in asset management, maintenance has been one of the major focus areas of academic research due to its potential in helping manufacturers to generate the most value from their assets. The emergence of condition-based maintenance (CBM) in which decisions are made based on the real-time condition of assets, has opened up new possibilities in developing more comprehensive approaches to improve the performance of production systems. For instance, a trend has been observed where attempts are made to couple CBM decisions with those on other production-related factors such as inventory control, spare parts management, and labour routing. The intrinsic link between the degradation behaviour of and the workload allocated to an asset, however, has not been sufficiently studied. Consequently, the potential benefits of intervening in machine degradation, either in the context of a single asset or a fleet of assets, are rarely explored. It is therefore essential that a systematic approach is at hand to improve system performance by exploiting the inter-relationship between production and maintenance.

This thesis is dedicated to developing a dynamic integrated decision-making model to improve the system-level performance of a fleet of parallel assets. The aim of the model is to realise the potential benefits, mainly in the form of lower maintenance costs and reduced penalty costs incurred due to loss of production, by simultaneously optimising workload allocation and the CBM threshold. The decision-making model is implemented using an agent-based system involving two types of agents - 1) machine agents that reside within each individual machine; and 2) a coordinator agent that oversees the entire system. The integrated decision-making model is constituted of two components - 1) a workload-dependent condition-based maintenance optimisation model based on Gamma Process at the asset level through a machine agent; and 2) a workload allocation strategy at the system level implemented by a coordinator agent. Numerical analysis is performed to demonstrate the rationale behind the decision-making process, which is to reach the most desirable balance between maintenance costs and penalty costs incurred by loss of production. The capability of the model to reduce total costs is demonstrated via comparison with traditional strategies such as uniform and random workload allocation. Additionally, the sensitivity analysis conducted has helped to reveal the respective factors that impact the potential reduction in maintenance costs and that in penalty costs, which include the sensitivity of asset degradation to workloads, heterogeneity of assets, penalty cost for a unit of production loss, redundancy of the system, etc.

The model presented in this study not only assists operation and maintenance managers to make decisions on the optimal combination of workload allocation and maintenance plans for assets in a production system, but also provides guidance on whether they should invest in workload control capabilities. Furthermore, the proposed approach allows practitioners to evaluate the long-term impacts of sudden events such as an increase in demand, a decrease in the number of redundant machines, and a change in the cost of maintenance actions.

Acknowledgements

The past four years are, and always will be one of the most unforgettable experience in my life, thanks to all the people who have been part of it and to Cambridge, the little town that has made everything happen.

First and foremost, my deepest gratitude goes to my supervisor, Dr. Ajith Kumar Parlikad. I would always be thankful for his continuous contribution of time, ideas, and support to my Ph.D. I am also grateful for his help while I was struggling with my case studies, and for all the insightful discussions we had. Furthermore, I truly appreciate the freedom he gave me to explore and to work at a comfortable pace. Without his constant guidance and feedback, this Ph.D would not have been achievable. I would also like to thank Professor Andy Neely for kindly agreeing to be my advisor in my first year.

My thanks also go out to the support and inspiration I received during the three months as an exchange research student to the Department of Industrial Systems and Engineering at Rutgers University. I would like to thank Professor Mohsen A. Jafari for hosting me, and Professor David W. Coit for kindly sharing with me his knowledge and expertise, which has provided invaluable inputs to my work.

I am thankful to Chris Riches for providing an industrial perspective to this study, and to all other industrial partners who contributed useful data and information to this study.

I gratefully acknowledge the funding received towards my Ph.D. from China Scholarship Council and Cambridge Commonwealth, European, and International Trust. My work was also supported by Henry Lester Trust and Lundgren Research Award.

Heartfelt thanks go to all DIAL members, past and present. The whole group has been more than merely a place to seek for advice and collaboration. It is also a source of joy and friendship. In particular, I would like to thank Raj for his words of encouragement when I started to write up, and Alena for helping me with optimisation problems.

I would also like to thank my best friend, Yi, for making time for answering all sorts of statistics questions I encountered in my research, and more importantly, for being my best friend.

Although many may say that Ph.D is inherently a lonely journey, mine has been one that is rather pleasant, thanks to all my friends from St Catharine's College and the IfM. In particular, I truly appreciate the time spent with Yuankun and Qingxin over the last two years. Their company made the difficult times I had during this period much less terrifying.

Lastly, I am sincerely grateful for having the love and encouragement of my family. A special thanks goes to my mum, who has done an absolutely amazing job of being both a loving mother and a trustworthy friend. Words cannot explain how blessed I feel to always have her support and understanding of all decisions I have made in my life. If it was not for her, I would never have made it this far.

Abstract

Effective asset management is considered key to reducing total costs of asset ownership while enhancing machine availability, guaranteeing security, and increasing productivity. Amongst all the activities involved in asset management, maintenance has been one of the major focus areas of academic research due to its potential in helping manufacturers to generate the most value from their assets. The emergence of condition-based maintenance (CBM) in which decisions are made based on the real-time condition of assets, has opened up new possibilities in developing more comprehensive approaches to improve the performance of production systems. For instance, a trend has been observed where attempts are made to couple CBM decisions with those on other production-related factors such as inventory control, spare parts management, and labour routing. The intrinsic link between the degradation behaviour of and the workload allocated to an asset, however, has not been sufficiently studied. Consequently, the potential benefits of intervening in machine degradation, either in the context of a single asset or a fleet of assets, are rarely explored. It is therefore essential that a systematic approach is at hand to improve system performance by exploiting the inter-relationship between production and maintenance.

This thesis is dedicated to developing a dynamic integrated decision-making model to improve the system-level performance of a fleet of parallel assets. The aim of the model is to realise the potential benefits, mainly in the form of lower maintenance costs and reduced penalty costs incurred due to loss of production, by simultaneously optimising workload allocation and the CBM threshold. The decision-making model is implemented using an agent-based system involving two types of agents - 1) machine agents that reside within each individual machine; and 2) a coordinator agent that oversees the entire system. The integrated decision-making model is constituted of two components - 1) a workload-dependent condition-based maintenance optimisation model based on Gamma Process at the asset level through a machine agent; and 2) a workload allocation strategy at the system level implemented by a coordinator agent. Numerical analysis is performed to demonstrate the rationale behind the decision-making process, which is to reach the most desirable balance between maintenance costs and penalty costs incurred by loss of production. The capability of the model to reduce total costs is demonstrated via comparison with traditional strategies such as uniform and random workload allocation. Additionally, the sensitivity analysis conducted has helped to reveal the respective factors that impact the potential reduction in maintenance costs and that in penalty costs, which include the sensitivity of asset degradation to workloads, heterogeneity of assets, penalty cost for a unit of production loss, redundancy of the system, etc.

The model presented in this study not only assists operation and maintenance managers to make decisions on the optimal combination of workload allocation and maintenance plans for assets in a production system, but also provides guidance on whether they should invest in workload control capabilities. Furthermore, the proposed approach allows practitioners to evaluate the long-term impacts of sudden events such as an increase in demand, a decrease in the number of redundant machines, and a change in the cost of maintenance actions.

Table of contents

List of figures

List of tables

Nomenclature

Coordinator Strategy Symbols

$(\delta')_l^m$	Symbol denoting whether unit m is 'considered' operational at the beginning of the l^{th} decision epoch from now on
$\bar{U}(S_l)$	Total maximum production of machines under S_l
δ_k^m	Symbol denoting whether machine m is operational at the k^{th} decision epoch
ε_k	The k^{th} decision-making epoch
$\mathbf{S}_l$	Set of all possible combinations of the status of machines at the beginning of the l^{th} decision epoch from now on
B	Buffer size used to control the conservativeness of the coordinator-level strategy
D_k	Production demand of the k^{th} time period
L	Length of time horizon considered at the coordinator level
l_k	The amount of unmet production demand in the k^{th} time period
M	Number of parallel machines in the production system
n	Number of load ratios used in generating regression functions at the machine level
$P(S_l)$	Probability of S_l happening
q_k	Penalty costs resulted from unmet production demand in the k^{th} time period
S_l	One combination of the status of machines at the beginning of the l^{th} decision epoch from now on
U_k	Total production from all machines in the k^{th} time period

$y(\cdot)$ The function that defines the relationship between q_k and $[U_k, D_k]$, $q_k = y(U_k, D_k)$

Individual Asset Model Symbols

$\Delta X(t)$ Degradation increment of an asset from time t to $t+\Delta t$, $\Delta X(t) = X(t+\Delta t) - X(t)$

$\Gamma(\cdot)$ Gamma function

κ_t Shape parameter at time t for the Gamma degradation process, $\kappa = \kappa_0 \Delta t$, where κ_0 is the shape parameter that corresponds to the unit-time Gamma distribution

$\lambda(S_i^-)$ Shock intensity of the asset immediately after the i^{th} shock and minimal repair

$\lambda(S_i^-)$ Shock intensity of the asset immediately before the i^{th} shock and minimal repair

$\lambda(t)$ Shock intensity of the asset at time t with degradation level $X(t)$, $\lambda(t) = \lambda_0 \exp(\lambda_1 \frac{X(t)}{F})$, where λ_0 represents the rate of shock failures happening to an asset in perfect condition ($X(t) = 0$), and λ_1 represents the significance of influence of asset degradation on vulnerability to shocks

θ_t Scale parameter at time t for the Gamma degradation process, $\theta_t = \theta_0(\exp(r_t))^s$, where θ_0 is a constant

$C_{mr,i}$ Minimal repair costs incurred between the i^{th} and $(i+1)^{th}$ preventive maintenance

C_{mr,t_0,N_0+1} Minimal repair costs incurred between t_0 and the $(N_0+1)^{th}$ preventive maintenance

F Failure threshold

$f_{X(R_i^+)}(\cdot)$ PDF of $X(R_i^+)$

$g(x; \kappa_0 \Delta t, \theta_t)$ Gamma distribution pdf for a degradation increment from time t to $t+\Delta t$

H Preventive maintenance threshold for individual assets

N Number of preventive maintenance carried out before a replacement takes place

N_0 Number of preventive maintenance that an asset has gone through at the initial state

n_{mr,t_0,N_0+1} Number of minimal repairs incurred between t_0 and the $(N_0+1)^{th}$ preventive maintenance

$Q(H, N)$ Expected remaining maintenance costs averaged over the expected life-cycle of individual machines

r_t Asset load ratio at time t, $r_t = \frac{u_t}{W}$

s Coefficient quantifying the sensitivity of asset degradation behaviour to load ratio

t_0 Initial time state

T_i Operating time period between the i^{th} and $(i+1)^{th}$ preventive maintenance

T_{t_0,N_0+1} The first operating time period of the asset from t_0 and before the $(N_0+1)^{th}$ preventive maintenance

u_t Workload assigned to the asset at time t

$v^m(\cdot)$ Regression function of optimal maintenance cost rate of unit m under r

W Production capacity of the asset

$X(R_i^+)$ Asset degradation immediately after the i^{th} imperfect preventive maintenance

$X(R_i^-)$ Asset degradation level at the instance when the i^{th} preventive maintenance is triggered

$X(S_i^+)$ Asset degradation immediately after the i^{th} shock and minimal repair

$X(S_i^-)$ Asset degradation immediately before the i^{th} shock and minimal repair

$X(t)$ Stochastic degradation process of assets

x_0 Initial degradation level of the asset

$z^m(\cdot)$ Regression function of the time to next preventive maintenance or replacement of unit m under workload ratio r

Maintenance-related Symbols

α_i, β_i Shape parameters of $f_{X(R_i^+)(x)}$

μ_i Mean value of the rectification effect by the i^{th} imperfect preventive maintenance

σ_i^2 Variance of the rectification effect by the i^{th} imperfect preventive maintenance

b Coefficient quantifying the deterioration rate of imperfect maintenance rectification

c_{mr} Cost per minimal repair task

c_{pm} Cost per preventive maintenance task

c_{rp} Cost per replacement task

t_{mr}^{m} Time taken to perform minimal repair on machine m

t_{pm}^{m} Time taken to perform preventive maintenance on unit m

t_{rp}^{m} Time taken to replace machine m

Chapter 1

Introduction

The thesis aims to develop an approach that simultaneously optimises the condition-based maintenance threshold and workload allocation among a fleet of parallel assets having the same production functions. This introductory chapter intends to provide an overview of the background and motivation of this research. A brief description is also given on the research problems this study attempts to address as well as the methodology adopted. The organisation of the thesis is given at the end of this chapter.

1.1 Background, Motivation, and Problem Description

As manufacturing has always been a capital-intensive industry, consistent efforts have been made to efficiently utilise resources within the organisations. Effective asset management is then considered key to reducing the total cost of asset ownership while improving machine availability, guaranteeing security, and increasing productivity. Multiple activities are included in asset management, ranging from the design in the very beginning until the decommissioning of physical capital. Amongst all the activities, the maintenance of assets has attracted noticeable attention from both the industry and the scientific community.

Early practice of maintenance was either reactive or time-based. Such practice, however, soon fell behind the requirements imposed by the rapidly developing manufacturing technologies and the increasing complexity of production systems. The growing needs to reduce unplanned maintenance as well as the proliferation of sensors and communication infrastructure has shifted the traditional practice towards condition-based maintenance (CBM). Based on the information collected through condition monitoring, CBM is a maintenance scheme that attempts to give recommendations on maintenance actions only when there are implications of abnormal behaviours of a physical asset, thus avoiding unnecessary maintenance tasks while at the same time keeping machine reliability at the required level

[53]. Consequently, there have been growing needs for mathematical models that can help with finding the best CBM threshold. For such models to be developed and implemented successfully, it is of course vital to choose the most appropriate health index to represent the degradation level of assets as well as to correctly describe asset deterioration behaviour under various operating conditions. One important factor characterising an operating condition is the type of task or the amount of workload assigned to an asset. It has been pointed out by Celen and Djurdjanovic [13] that different operations degrade the manufacturing cells of a cluster tool differently, and by Liao and Elsayed [60] that LED bulbs fail faster if higher light intensity is required from them. Although treating the deterioration of assets as an independent self-evolving process largely simplifies the situation, sub-optimal solutions may be implemented and costs might be incurred which could have been saved. The fact that the workload-dependent degradation behaviour of assets has not been sufficiently studied and modelled forms one corner of the motivation for this research project.

Until the recent past, it has often been the case that machines in a manufacturing plant are supplied by different original equipment manufacturers (OEMs) that provide maintenance service for their own products. As a result, a significant amount of work has been dedicated to finding the best maintenance practice for a single asset, ranging from models for determining the optimal inspection interval to data analysis techniques for abnormality detection. However, implementing a single-asset maintenance model in real-life situations is both unwise and impractical as nowadays most production goals can only be achieved with satisfactory overall performance of a fleet of assets. Having one asset performing at its best does not naturally lead to the optimal performance of the production system. The fundamental reason lies in the inter-dependencies existing between assets belonging to the same fleet. One category of dependency that is of particular relevance is named performance dependence [77], which states that the system performance is determined by the performance of assets in the system as well as their configuration as a fleet. A typical example of optimisation model developed on the assumption of performance independence is given in Van Horenbeek et al. [93]. The same downtime cost is incurred by each unexpected failure of assets, which is not the case in the so-called 'k-out-of-N' systems where no cost is induced at the system-level as long as k or more assets are functioning as normal. It is thus crucial that asset owners have a clear understanding of the performance dependency while making maintenance and production decisions.

Manufacturing has evolved from its infancy where simple tasks were usually carried out by a single skilled labour to a complex process involving multiple aspects, such as raw material preparation, production, maintenance, quality control, and customer service on finished products. Treating these areas independently with separate models is likely to yield

sub-optimal solutions since in fact these areas are interrelated [41]. For instance, lack of maintenance may lead to a higher probability of non-conforming parts being produced, which is likely to push more investment into quality control. Ignoring the intrinsic connections between various elements in the manufacturing process may incur additional costs on shareholders. As a matter of fact, a growing research interest has been seen in the development of integrated decision-making models that consider more than one production-related factor [41]. That being said, the decision making of condition-based maintenance threshold with workload and task allocation has not been addressed until very recently. Moreover, few models consider the varying degradation rates caused by assets performing different types of tasks or taking on different workloads. Consequently, the potential benefits of intervening in machine degradation have rarely been explored. For example, controlling the degradation of an asset by adjusting its workload would enable asset owners to postpone or bring forward maintenance to a more preferable time. Stakeholders can gain more flexibility and control over their assets by using this option, as now they may not have to wait for maintenance requests that can be triggered by any machine at any time. Furthermore, it is not uncommon nowadays to see manufacturing plants in possession of a fleet of parallel assets capable of conducting the same type of tasks. The redundancy in the system then raises new questions: how do we know which assets should be put to use, and how do we allocate the workload among the assets? A straightforward argument might be that the newer the machine, the more workload it should take on. While this is intuitively correct, a counter-intuitive solution is found in [42], where it is preferable to allocate the most workload to the most degraded machine to ensure a satisfactory overall long-term performance.

The implication of the discussions above is that, when a decision has been made on the allocation of workload among a fleet of assets in a production system, it will affect both every individual assets and the entire system due to performance dependency between assets. For a single asset, its deterioration behaviour, and consequently its maintenance plans, will change as the workload varies. For the production system, a specific workload allocation not only has impact on the maintenance costs that will be incurred, but also affects the system availability in the long term. It requires a holistic view over the dynamics between workload allocation, machine degradation, condition-based maintenance optimisation, and system availability in the long term in order to realise the most benefits from assets.

This thesis aims to formulate an integrated condition-based maintenance threshold and workload allocation decision-making model for a fleet of parallel assets that produce the same type of products and attempt to meet production demands together. The model developed in this research project is intended to realise the potential benefits by exploiting the dependency between assets as well as the interactions between workload and deterioration within an

individual asset. The key findings from both the process of developing the model as well as analysis of modelling results will also assist companies in deciding whether the proposed model is suitable for their own business and assets. The research questions that will be addressed by this study are presented in the next section.

1.2 Research Questions

It follows from the discussion above that additional benefits might be realised with a dynamic optimisation model that ties together the decision making of condition-based maintenance threshold and workload allocation to a fleet of similar assets. For the purpose of addressing this gap, here we formally propose the following three key research questions:

- **Research Question 1:** What are the constitutive components for a dynamic optimisation model for joint decision making of CBM threshold and workload allocation for an asset fleet? Discussion in a later chapter will reveal the two important components of such a model: a machine-level load-dependent CBM optimisation model and a system-level task/workload allocation strategy.
- **Research Question 2:** How can we quantify the impact of workload on the degradation behaviour of individual assets, and how can such knowledge be used to set the most appropriate CBM threshold?
- **Research Question 3:** How can we quantify the impact of a specific workload allocation on maintenance and production at the system-level, and what type of information is needed in order to do so?

The first research question requires the constitutive components of such a decision-making model to be identified and the structural relationship between these components to be defined. The second question concerns the development of a load-dependent condition-based maintenance optimisation model for individual assets. The third research question is focused on developing a workload allocation strategy that incorporates the interactions between production and maintenance to seek optimal system-level performance.

1.3 Research Methodology

In order to address the research questions brought up in the previous section, we followed the steps shown in Figure 1.1.

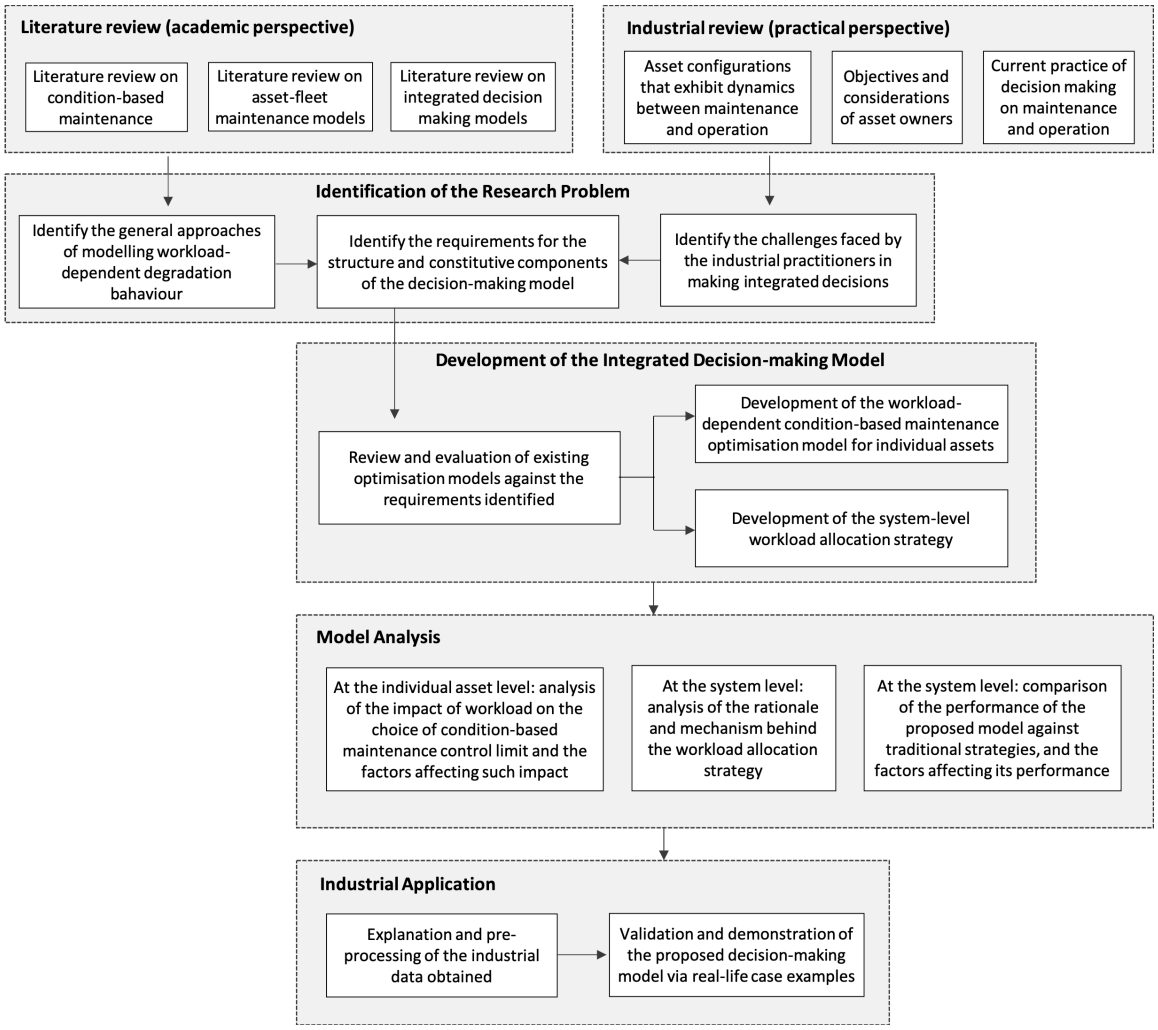


Fig. 1.1 Research methodology followed by the thesis

First of all, the research gap to be addressed by this study is identified from both an academic and a practical perspective. Specifically, the literature review is conducted along three different dimensions: 1) review on condition-based maintenance optimisation models for individual assets, with an emphasis on operation/workload-dependent degradation models; 2) review on asset-fleet maintenance models; and 3) review on integrated decision-making models, with an emphasis on operation/workload allocation models. The industrial reviews include both case studies from the academic literature as well as exploratory case studies with manufacturering companies. The purpose here is to gain practical insights into the current practice in maintenance and operation planning as well as to identify key challenges faced by the industry regarding the decision-making process. These two qualitative reviews led to the identification of the research problem this study attempts to solve.

Further discussions on the research gap then lead to the identification of the requirements for the condition-based maintenance and workload allocation integrated decision-making model to be developed later in the thesis. These requirements concern both the structure as well as the constitutive components of the decision-making model. With the help of review and evaluation of existing approaches against these requirements, an innovative integrated decision-making model is then developed. Closer examinations into the proposed model using numerical examples are conducted on the proposed approach to demonstrate the rationale and mechanism behind the decision-making process. The performance superiority of the proposed model is demonstrated via comparison with traditional strategies, and how various factors affecting such superiority is discussed via sensitivity analysis. Finally, the model is applied to real industrial cases in order to demonstrate the applicability and validity of the proposed approach.

It can be noticed that this research project involves both qualitative and quantitative analysis, and consequently combines the advantages of the two: the qualitative part ensures that a holistic view on maintenance and operation practice is obtained in a proper context; while the quantitative part enhances the reliability and credibility of the study.

1.4 Organisation of Thesis

The main body (from Chapter 2 to 7) of the thesis is organised into six components:

1. Identification of the research problem
2. Development of an integrated workload allocation and CBM threshold decision-making model for a fleet of assets
3. Analysis of how the model captures the impact of workload on the choice of CBM threshold and the factors affecting such impact
4. Analysis of the advantage of the model over traditional strategies and the factors affecting the scale of such advantage
5. Demonstration of the model in practice
6. Conclusions

Figure 1.2 shows the relationship between the six components described above and the following six chapters.

Chapter 2: Literature Review

This chapter lays the academic background of the study by providing a detailed review of previous researches relevant to the field of this research project. The three goals this chapter attempts to achieve are: 1) to set the scene of this project by giving an introduction to asset management and key concepts involved; 2) to provide an in-depth analysis of existing literature on both individual and integrated condition-based maintenance models; 3) to identify the research problem and approaches that can potentially be extended to address the problem.

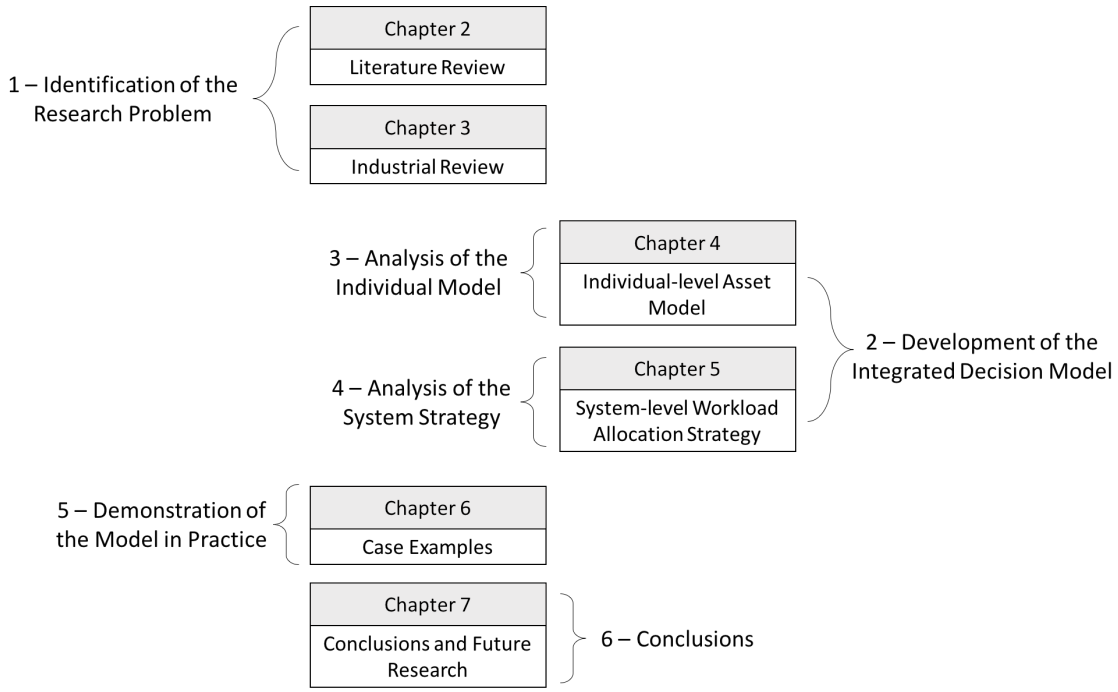


Fig. 1.2 Structure of the thesis

Chapter 3: Industrial Rationale

The third chapter aims to form an industrial understanding on the necessity of an integrated condition-based maintenance and workload allocation decision-making model. A review of current practice in the industry is provided. This chapter has gained insights into the specific industries and asset configurations that have the potential to benefit from such a decision-making model.

Chapter 4: Condition-based Maintenance Optimisation for Individual Assets

This chapter attempts to answer the first and second research questions defined in Chapter 2 by: 1) first identifying the constitutive components of an integrated condition-based

maintenance and workload allocation decision-making model; 2) and then building the first component - a workload-dependent condition-based optimisation model for individual assets. The second step is achieved by conducting an analysis of the requirements imposed on the individual model and comparing existing approaches against such requirements. The model is applied to numerical examples and discussions of the obtained results are provided.

Chapter 5: Coordinated Workload Allocation Strategy for Parallel Assets

The fifth chapter contributes to the third research question identified in Chapter 2 - to design a coordinator-level workload allocation strategy that acts as the second constitutive component of the integrated decision-making model. A framework of how the two components come together to form the complete decision-making model is given, followed by a detailed description of the coordinator strategy. Similar to the previous chapter, the complete model is applied to numerical examples. The performance of the proposed strategy is compared to that of traditional strategies. Analysis is conducted on how various factors affect the benefits provided by the proposed decision-making model.

Chapter 6: Case Examples

This chapter aims to demonstrate the practicality of the proposed decision-making model by applying it to a real-industry setting. The proposed approach is tested on the same fleet of assets but for two different scenarios. This chapter also discusses the conditions under which the proposed methodology is likely to be beneficial.

Chapter 7: Conclusions and Future Research

The last chapter serves as a conclusion of this research. Specifically, a revision of the problem this study aims to address is provided, followed by a summary of the key results and findings. Furthermore, the limitations and applicability of the proposed decision-making model are discussed and directions on how potential future research can be conducted are recommended.

Chapter 2

Literature Review

2.1 Introduction

The previous chapter gave an overview of the research problem, whereas this chapter aims to lay the academic background. There are three goals that this chapter attempts to achieve:

1. to set the scene for this research by giving a brief introduction to asset management, maintenance strategies as part of asset management, and the importance of maintenance strategies aimed at an asset fleet rather than a single asset;

2. to provide a detailed review and in-depth analysis of existing literature on condition-based maintenance and integrated decision making involving maintenance elements in order to define the gap that will be addressed by this research project;

3. to introduce existing methodologies that might potentially give rise to new approaches to answer the research questions proposed in this thesis.

The chapter will start with an introduction to key concepts and elements in asset management, which is the broad category within which the content of this research falls. The importance of maintenance, especially fleet maintenance, is highlighted. After presenting some basic definitions related to asset management, we further proceed to discuss its evolutionary trend along two different dimensions: 1) from reactive to proactive; 2) from maintenance-centred to being more comprehensive. Here, emphasis is placed specifically on how the evolutionary trend is represented in the choice of maintenance strategies. In the later part of the chapter, a detailed review will be given on existing literature related to each of the aforementioned evolution dimensions.

Specifically, for the first dimension, we will show that maintenance strategies have grown to take different shapes through its history of more than half a century, where condition-

based maintenance has gradually become a very important element of modern maintenance strategies. A structured discussion is presented on existing optimisation models for condition-based maintenance decision making. These models are analysed in details at their constitutive components level to help with partially identifying the research gap.

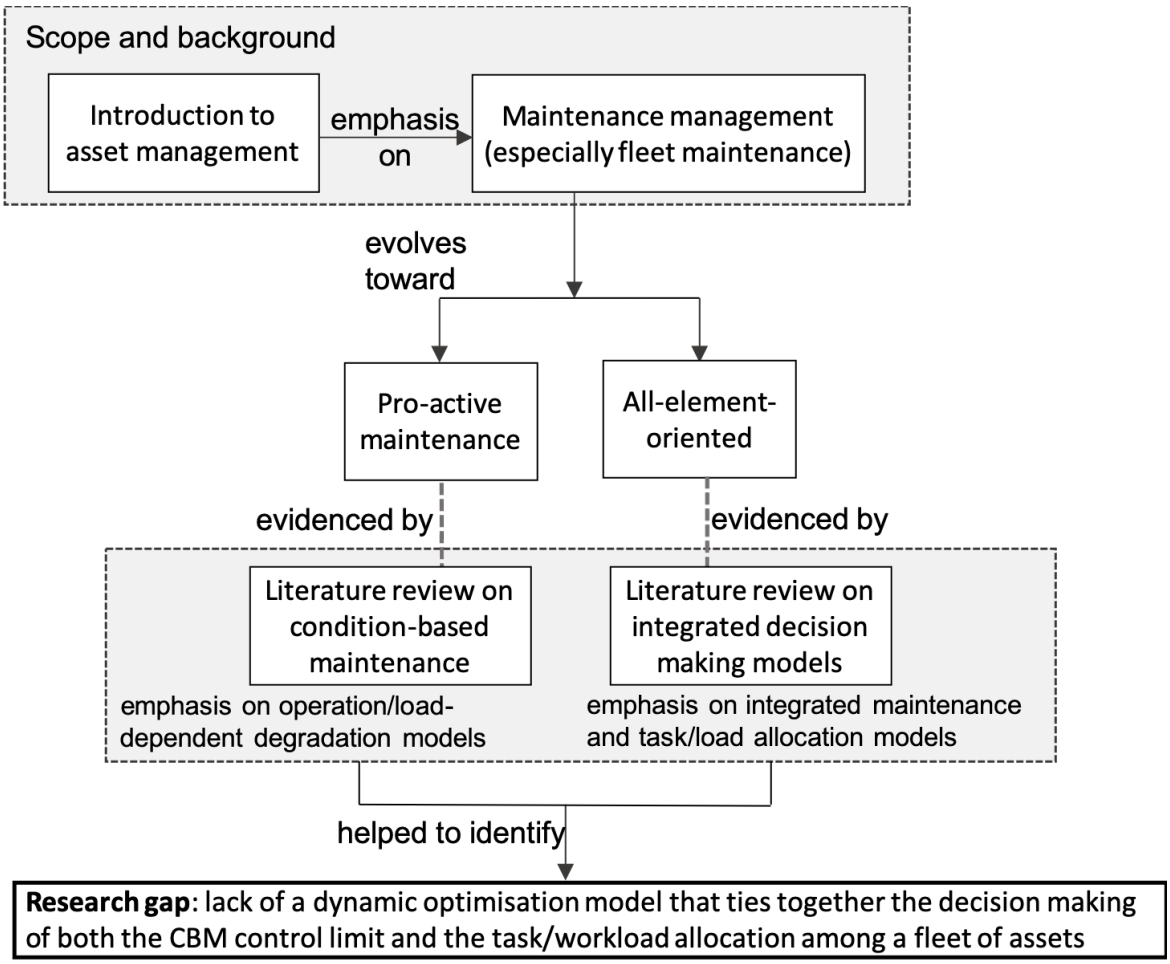


Fig. 2.1 The story-line followed by the literature review chapter

The chapter continues to elaborate on the evolutionary trend of asset management and to further define the research gap. The third evolutionary dimension is mainly about integrating maintenance decision making with other production elements. The technical aspects of joint decision making of condition-based maintenance and workload allocation models are presented, followed by a discussion on the characteristics of industrial settings that these models are invented for. Subsequently, scenarios that may be encountered by asset owners in their day-to-day operation but are not considered by these models have been identified.

A wrap-up of the previous literature review and findings is then presented, which ultimately leads to a discussion of the research gap that will be addressed by this thesis. Then

an introduction is given to a promising methodology to close the research gap. A review of existing literature on models of decentralised decision making in manufacturing systems is provided, with an emphasis on agent-based systems for maintenance optimisation.

The literature review chapter is then closed with a summary. The overall story-line followed by the literature review can be found in Fig 2.1.

2.2 Asset Management Overview

This section provides an introduction to asset management and its role in the domain of manufacturing. It also outlines the historical trend of asset management since its infancy.

2.2.1 Introduction to Asset Management

Effective asset management is key to reducing the total cost of asset ownership while improving machine availability, guaranteeing security, and increasing productivity. The definition for asset management in ISO 55000:20143.3.1 is given as follows:

Definition 2.2.1. Asset management is the coordinated activity of an organisation to realise value from assets.

The note following the definition given by the International Standard attaches the term 'activity' to a broad meaning that includes anything from planning to implementation. For example, design, commissioning, operating, maintaining, replacing, and decommissioning are all considered viable activities for physical and infrastructure asset management.

Amongst all the activities, the maintenance of assets has been intensively studied for one of the following two types of reasons: lessons of the past; and requirements of the present and the future. For the first category, lessons from the past have proved that incorrect planning and scheduling of maintenance activities may lead to unexpected breakdowns and repair works, incurring great costs to the asset owners. This is especially detrimental in companies where production process interruption implies tremendous revenue losses such as those in the oil and gas industry. Maintenance costs may take up from 15% to 70% of total production costs [72], and one minute of downtime in an automotive manufacturing plant could cost the owner as much as $20,000 [33]. When it comes to the second type of reasons, a few recent advances in manufacturing can help to explain the growing interest in maintenance management. Before digging into the details of three such advances, it is worth mentioning that ultimately they all boil down to how the perception of maintenance has changed over time - the concept of maintenance has evolved over time from a conventional perception where

maintenance is considered as a necessary evil that induces excessive costs to the current view where it is seen as a prospective tool for production value generation and business performance improvement. Organisations have progressively started to adopt just-in-time manufacturing and lean manufacturing, which demands agility throughout every step in the production process and leads to substantial attentiveness to asset health condition [24]. Another observable trend in the manufacturing industry is called 'servitisation', referring to the move from selling products towards selling service, where the manufacturers are rewarded based on the availability and performance of the assets they provide [26]. A typical example is the 'Total-Care Package' offered by Rolls-Royce to its customers in the airline business [86]. As the side effects of extensive economic growth have become too severe to be ignored, sustainability is now a frequently mentioned topic on the table. With the capability of increasing asset life and saving energy, maintenance is considered to be of good potential to contribute to sustainability.

This subsection has presented some generic knowledge of asset management and highlighted the importance of its maintenance element. In the subsequent subsection, we narrow down to one specific element of maintenance management - asset fleet maintenance, which sets the scope of this research.

2.2.2 Introduction to Asset Fleet Maintenance

Here the definition of maintenance strategy given by the European Standard on maintenance terminology EN13306 2001 is provided:

Definition 2.2.2. Maintenance strategy is a management method used in order to achieve maintenance objectives or goals.

For the past several decades, it has often been the case that machines in a manufacturing plant are supplied by different original equipment manufacturers (OEMs) that provide maintenance service for their own products. As a result, a significant amount of work has been dedicated to finding the best maintenance approach for a single asset, ranging from models for determining the optimal time interval for time-based maintenance, to data analytics algorithms for abnormality detection of a single component. However, in most real-life situations, certain production goals can only be achieved with satisfactory overall performance of a fleet of assets working jointly. Efforts devoted to reliability engineering research therefore started to shift from the development of single-asset to multi-asset models with more complicated features, which is believed to bring about opportunities to enhance the efficiency of the overall system [77]. For the purpose of clarity and in order to limit the scope of the study, the classification approach proposed in Petchrompo and Parlikad

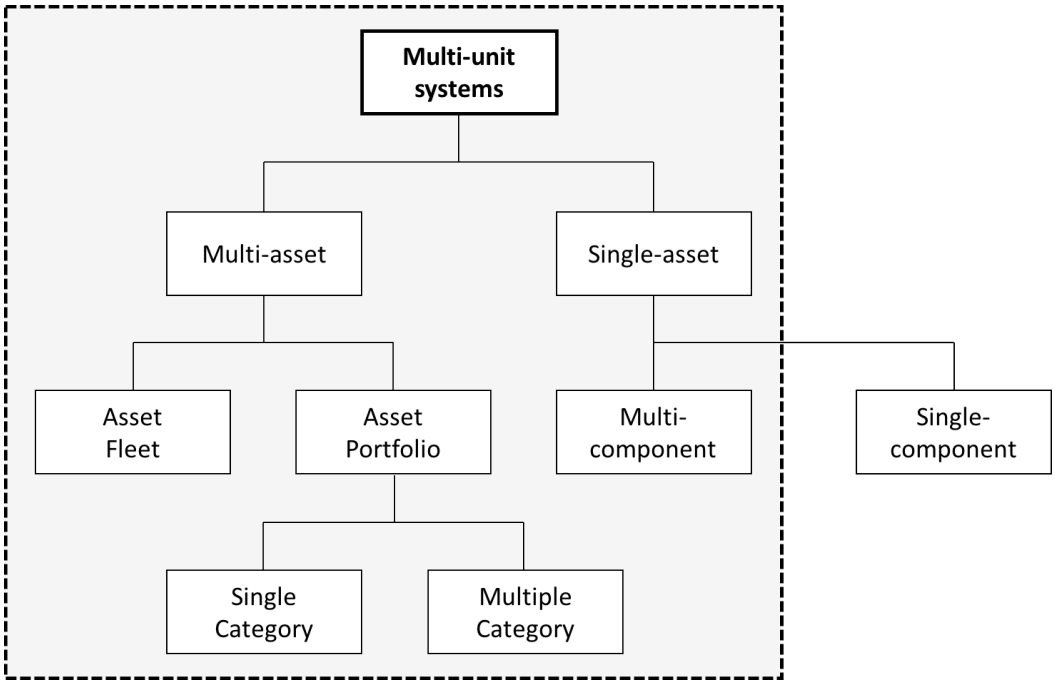


Fig. 2.2 Classification of multi-unit systems [77]

[77] for 'multi-unit systems' is adopted here to differentiate 'multi-asset systems' from 'multi-component systems', as presented in Fig 2.2:

Definition 2.2.3. Multi-asset system is a system composed of multiple assets that share common characteristics or resources under the control of an organisation.

Definition 2.2.4. Multi-component system is a single-asset system composed of multiple components operating together.

Multi-asset systems is then further categorised into 'asset fleet' and 'asset portfolio' according to their fundamentally different characteristics, as shown in Table 2.1. Note that the subject of interest in this research is asset fleet, namely, a group of assets that are identical or slightly different from each other but share similar technical features and missions. A typical example is a fleet of buses that can be assigned service slots to jointly meet passenger logistics demand [109].

Fundamentally, the need for multi-asset maintenance models arises from the inter-dependencies among assets. Three major dependency categories are identified in [77]:

1. performance dependence, which states that the system performance is determined by the performance of assets in the system as well as their configuration as a fleet;

2. stochastic dependence, which states that the effect of deterioration of an asset on other components;

Table 2.1 Summary of fundamental features in two multi-asset categories [77]

Feature	Category	
	Asset fleet	Asset portfolio
No. of assets	Multiple	Multiple
No. of components	Multiple	Multiple
Asset diversity	Homogeneous	Heterogeneous
Asset category	Single	Single/Multiple
Intervention option	Equivalent among assets	Different among assets
Typical asset type	Vehicles	Infrastructure assets

3. resource dependence, which states that a fleet of assets might share common resources in their intervention actions, be it financial, labour, or physical, leading to constraints in maintenance decision making.

Due to the dependencies present in an asset fleet, having every single asset performing optimally does not necessarily lead to optimal system-level performance. For the purpose of clearly showing the need for asset-fleet maintenance models, only performance dependence is discussed for this part, contents related to the other two types of dependencies will be left to Section 2.2.3 where the interrelationship between maintenance and other production elements will be closely examined. To give an example of performance independent model, consider the policy proposed by Van Horenbeek et al. [93]. Each unexpected failure of assets incurs the same downtime cost at the asset level, which implies that the loss at the system-level is exclusively expressed as the sum of losses at the asset level. However, this is only true under very strict assumptions. One counter example is the so-called 'k-out-of-N' systems, where no system-level cost is induced as long as at least k assets in the fleet are operational.

Despite it being apparent in parallel systems, the performance dependence between assets has rarely been specifically discussed in existing research papers. This type of dependence is present where the overall system performance is determined by the performance of individual assets in the fleet as well as their configuration. One contribution that explicitly captures the inter-asset relationship was made by Rasmekomen and Parlikad [80], where they modelled the degradation behaviour of a system consisting of both non-critical machines and a critical machine. Their model inquired into how the individual non-critical unit affects the critical unit and thus the performance of the entire system. Rasmekomen and Parlikad [80] also provided a numerical analysis showing that ignoring the dependence between assets would

lead to a lower profit per unit time. Performance dependence can also be seen in '*k*-out-of-*N*' systems. Albeit similar in configuration, a multi-asset '*k*-out-of-*N*' system is very different from a multi-component '*k*-out-of-*N*' system in the following aspect: the former fails when less than *k* components are in good condition, whereas the latter can still function, but with a deteriorated performance. This concept can be illustrated by the model developed by Liang et al. [59], where they considered the maintenance scheduling problem of multiple fleets each demanding a certain number of machines be functional at all times. Petchrompo and Parlikad [77] developed a maintenance scheduling optimisation model that attempts to ensure the availability of 5 out of 7 vessels at all times. Other studies on '*k*-out-of-*N*' systems can be found in Moudani and Mora-Camino [71], Safaei et al. [83].

In the subsequent subsection, the evolution of asset management is provided, centred around how that is represented in maintenance strategies.

2.2.3 Evolution of Maintenance Strategies

Asset management, maintenance policies in particular, has evolved from its infancy to the current form of sophistication and thoroughness through decades of accumulative improvement, of which the most noticeable changes along two different dimensions, represented by the choice of maintenance strategies, as shown in Fig 2.3 will be discussed later in this section.

From Reactive to Proactive

This subsection first briefly discusses why and how maintenance practice has shifted from reactive to proactive. A description is then given on the three major steps involved in implementing one specific type of proactive maintenance, condition-based maintenance: 1) data acquisition; 2) data processing; 3) maintenance decision making. First, proactive

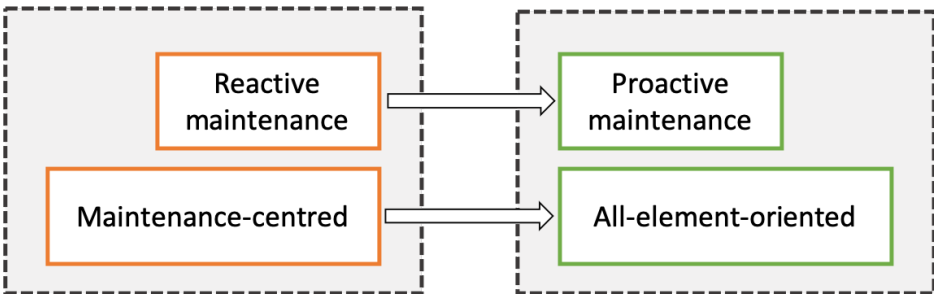


Fig. 2.3 The evolution of asset management along two different dimensions

maintenance has gradually gained popularity over traditional reactive maintenance. The first

and earliest maintenance practice in manufacturing plants is corrective maintenance (CM) [32] where a machine or a component is run to failure before it is repaired or replaced. The historical reason behind this is the light mechanisation in the industry in the 1950ś when unexpected machine breakdowns would not result in consequential losses [70]. As CM does not intervene in the degradation status of assets, no increase in asset reliability is achieved. As the size and complexity of production plants increase over the following two decades, even the failure of one component might cause complete shutdown of the whole system, and complement to pure corrective maintenance is called for. One straight forward solution to this problem is to adopt redundancy in the design of the system, maintenance staff, and spare parts [31]. Despite its ease of implementation, this strategy does not provide guidance on the desirable extent of redundancy and thus tend to incur excessive waste. Consequently, the notion of preventive maintenance (PM) emerged, which is defined by MIL-STD-721B as follows:

Definition 2.2.5. Preventive maintenance refers to all actions performed in an attempt to retain an item in specified condition by providing systematic inspection, detection, and prevention of incipient failures.

Conceptually, PM can be classified into reliability-centred maintenance (RCM), business-centred maintenance, risk-based maintenance, total-productive maintenance (TPM), and etc. [1]. Although this approach of classification clearly represents the operational function of maintenance perceived by a company in its business, for the purpose of this specific research project, another classification method is adopted, which is based on the technique used for solving maintenance problems - time-based maintenance (TBM) and condition-based maintenance (CBM).

After the idea of PM was accepted in the industry, the next question that came immediately to the mind of maintenance experts was when to perform maintenance tasks. At a time with hardly any advanced monitoring technique or systematic documentation of asset degradation status, it was proposed that maintenance decisions could be made using the most readily available information - failure-based or usage-based date of assets, leading to a large number of TBM models. In the simplest form of TBM, a unit is replaced when a pre-set time interval T has passed or upon failure, whichever comes first. The optimal T is usually chosen in order to maximise asset reliability, minimise maintenance cost, maximise availability, or guarantee a certain safety performance [78]. Extensions were later made to the original form of TBM to include other maintenance actions, such as minimal repairs and imperfect preventive maintenance.

Though TBM has to some extent met the expectation of higher machine availability and lower unexpected failure frequencies, it is hardly scheduled at the optimal maintenance

interval, resulting in unnecessary costs caused by excessive maintenance activities [69]. This is mainly due to the fact that the optimisation of maintenance decision often needs to be performed in face of uncertainties. In order to quantify the stochasticity associated with the time to failure or the asset deterioration rate [94], most TBM models adopt one of the following approaches: 1) using a probability distribution to describe asset lifetime, among which Weibull distribution is the most frequently used for its flexibility in modelling lifetime distribution of various characteristics [65]; 2) relying on an age-dependent failure rate or hazard rate function to quantify asset reliability [110, 75]. However, Van Noortwijk [94] pointed out that lifetime distribution models only quantify asset status as being functioning or failed, and the information on the actual status of assets is missing.

The growing concerns about over maintaining equipment, together with the proliferation of sensors equipped on assets generating large amounts of condition monitoring data, has shifted the idea of maintenance from TBM to CBM. Based on the information collected through condition monitoring, CBM is a maintenance scheme that attempts to give recommendations on maintenance actions only when there are implications of abnormal behaviours of a physical asset, thus avoiding unnecessary maintenance tasks while at the same time keeping machine reliability at the required level [53]. Though CBM can bring great benefits such as cost reductions, resource savings, and improved availability of assets [68], it is a complex programme that involves various steps and strategic decisions, which will be reviewed in details Section 2.3.

From Maintenance-centred to All-element-oriented

Manufacturing is a complex process involving multiple aspects and steps starting from raw material procurement all the way to product quality control and customer service. These areas used to be treated independently, yielding separate models for each function. It has been widely acknowledged that these models are likely to provide suboptimal solutions due to the fact that these areas are interrelated [41]. This is how the other two types of dependencies come into play. In short, stochastic dependence is best illustrated in load-sharing systems: when an asset in a parallel system fails, other operational assets are required to cope with the extra workload. This might lead to faster degradation of other assets and ultimately result in changes to the original maintenance plan. Resource dependence is mostly related to assets reliant on the same workforce, inspection tools, maintenance facility, and spare parts. When multiple assets fail simultaneously, it is not always the case that maintenance resource is available for all of them, in which case production may be delayed for an unexpected long period of time.

Realising the stochastic and resource dependence in asset fleets has opened up new grounds for improving the efficiency of production systems. Thus, a growing research interest in this field has been seen in the development of decision models that take into account more than one production-related factors. Later in Section 2.4.1 a review will be given on existing models that consider multiple production elements, with an emphasis on integrated decision making of maintenance and workload allocation.

2.3 Condition-based Maintenance

This section serves to provide a background to the major processes in the implementation of condition-based maintenance. Though CBM is believed to have the potential to increase the asset reliability and reduce unexpected downtime, it requires careful planning, advanced data analytics, complex mathematical modelling, and incurs extra installation costs. Before diving into the technical steps and attaching sensors onto every equipment, it is necessary to identify the type of assets whose importance could justify such investment. Various approaches have been proposed to determine the criticality of assets like Fault Tree Analysis (FTA), Reliability Centred Maintenance (RCM), and Failure Mode and Effect Criticality Analysis (FMECA), just to name a few [90]. Once the critical assets have been identified, three key steps need to be followed for successful CBM practice: data acquisition, data processing, and maintenance decision making [53]. The purposes of the aforementioned three procedures are illustrated in Fig 2.4.

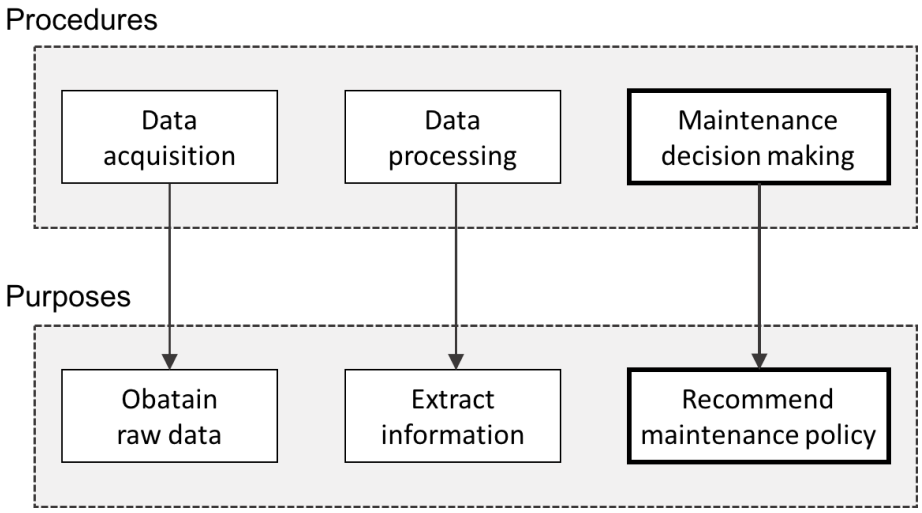


Fig. 2.4 Mapping of the procedures and purposes of condition-based maintenance

2.3.1 Data Acquisition

Data acquisition is the very first step in achieving condition-based maintenance decision-making. It is aimed at obtaining and storing data that can provide insight into the health status of assets. In general, data is collected using various sensors, such as accelerometers, voltage sensors, and vision systems [82], and are classified into two types:

1. event data, which records events that have taken place (e.g. breakdowns, inspections, preventive maintenance actions, etc.). Apart from relying on human efforts, organisations have also put to use manufacturing information management systems to improve efficiency of event data collection and input. Enterprise resource planning(ERP) and computerised maintenance management systems (CMMS) are two examples of such systems;

2. condition monitoring data, which can take various forms depending on the type of assets that are being monitored. For instance, the contamination level of oil is often measured to reveal lubrication problems of machine components. On the other hand, cracks in bearings and shaft misalignment in oscillating machines can be detected via vibration analysis. Other condition data that can be measured include strains, pressure, temperature, acoustic data, etc.. Upon completion of data collection, wireless technologies, such as Bluetooth and Wi-Fi, are used for data transmission.

2.3.2 Data Processing

This is the step right after data acquisition. Data processing involves first cleaning the data obtained, and then analysing it using appropriate tools. Condition monitoring data is diverse in nature and the technique used to extract degradation information needs to be selected accordingly. The condition monitoring data collected mainly falls into one of the following categories: 1) value type, which is collected as a single value at a specific time epoch; 2) waveform type, whose value is in the form of a time series, such as acoustic data, vibration data, etc.; 3) multi-dimension type, such as X-ray images, infrared thermographs, etc..

Value type data is often processed using principle component analysis (PCA) and independent component analysis (ICA). Though seemingly easy compared with waveform data analysis, the analysis can still be complicated due to the possible correlations between multiple variables.

Since signal processing for multi-dimensional data can be seen as a more complicated version of that for waveform data, only the recent advances in waveform signal processing is presented here. Time-domain analysis, frequency-domain analysis, and time-frequency

analysis are the frequently used tools to analyse waveform signals. Time-domain analysis is directly applied to the time waveform itself to extract descriptive time-domain features such as mean values, standard deviations, peak-to-peak intervals, etc.. Examples of typical time-domain analysis approaches are time synchronous average (TSA) [27], the autoregressive (AR) model, the autoregressive moving average (ARMA) model, and the principal component analysis (PCA).

Frequency-domain analysis is done on the frequency-domain waveform transformed from the acquired time-domain signal to isolate components of certain frequencies. Fast Fourier transform (FFT) is normally used to compute the discrete Fourier transform of a sequence, which is further processed for spectrum analysis. Another frequency-domain transform of a signal called a cepstrum is also widely used. A cepstrum is the result of taking the Inverse Fourier transform (IFT) of the logarithm of the estimated spectrum of a signal, which has advantages over FFT in detecting harmonics and sideband patterns [53].

Time-frequency analysis has been developed to make up for the inability of frequency-domain analysis to deal with non-stationary waveform signals. Short-time Fourier transform (STFT) [4] and Wigner-Ville distribution [64] are the most commonly accepted time-frequency distributions that can be found in the literature.

The aforementioned methods are mainly built to handle only condition monitoring data. More often than not, however, effective fault diagnostics and prognostic can only be achieved when event and condition monitoring data are jointly analysed. Time-dependent proportional hazards model (PHM) and Hidden Markov model (HMM) are two widely adopted tools serving this purpose. PHM relates the failure probability of a component to both its age and health indicators through a hazard function [52, 91]. Due to its success in speech processing and high computational efficiency, HMM is extrapolated to deal with vibration data of machines. HMM describes the degradation of a system using a Markov chain with hidden states that account for the actual observations [11].

2.3.3 Maintenance Decision Making

The last stage in implementing CBM is maintenance decision making. This step aims to assist asset owners in achieving better system performance by taking appropriate maintenance actions at the right time. Techniques applied in this step are divided into two categories: diagnostics and prognostics. Diagnostics is backward-looking in the sense that it aims to detect, isolate, and identify faults after their occurrence, whereas prognostics is forward-looking as it mainly concerns the prediction of faults. CBM is traditionally incorporated in prognostics, but due to the focus of this study it is covered here as the third category at the stage of maintenance decision making.

Diagnostics

Diagnostics maps the obtained knowledge and feature after the data processing step to machine faults, which is equivalent to pattern recognition [53]. Some popular approaches taken to perform pattern recognition tasks are statistical approaches, artificial intelligent (AI) approaches, and model-based approaches. The fault diagnosis problem is described as a hypothesis test problem in statistical approaches, where condition monitoring information is processed to construct test statistics. The test result helps to decide whether to accept or reject the hypothesis that a certain fault is present. Statistical process control (SPC), cluster analysis, and hidden Markov model (HMM) all fall into this category. In recent years, AI techniques, such as artificial neural networks (ANNs), fuzzy logic systems, evolutionary algorithms (EAs), etc., have gained popularity over conventional statistical approaches. The lack of effective data, however, has hindered the actual implementation of AI techniques in fault diagnostics. Siddique et al. [85] has given a more detailed review of recent applications of AI in machinery fault diagnostics. Model-based approaches build explicit mathematical models of the equipment being monitored based on mechanics and physics. The models act as a reference of expected operation, and fault diagnostics is based on the differences between model predictions and the monitored system behaviour. Some of the existing studies on model-based machine fault diagnostic include Choi and Choi [22], Jalan and Mohanty [51], Howard et al. [44].

Prognostics

The scale of literature of machinery prognostics is much smaller than that of diagnostics. The two most common types of prediction in prognostics are: (1) the time left before the next failure given the current machine age and condition-remaining useful life (RUL); (2) the probability of a machine operating without failure until the next scheduled inspection. Similar to diagnostics, prognostics are performed following one of the three approaches: statistical approaches, AI approaches, and model-based approaches. Some of the statistical models used in fault prognostics are SPC [39], ARMA time series model [98], and HMM [20]. AI-based approach can be useful when no clear failure definition models are available, as evidenced in [21]. As is the case with diagnostics, model-based approaches to prognostics is also built upon specific knowledge and theory about the structural and mechanical characteristics of the asset. A typical case is presented in [73] where a physical model based on the law of crack growth is applied to predict RUL.

Condition-based Maintenance Models

Once insight related to fault diagnostics and prognostics has been extracted from the data obtained, it is natural to associate this condition information with maintenance decision making. The most common decision in CBM is whether and when to perform inspections, minimal repairs, preventive maintenance, and replacement. These decisions are often made to achieve a balance between maintenance cost, downtime, reliability, availability, and other performance indicators.

A CBM model, as mentioned in Sharma et al. [84], typically consists of the following elements:

1. a description of a technical system and its major functions;
2. a modelling of the way the system deteriorates in time and the possible consequences;
3. a description of the available information regarding the system status;
4. the action pool open to the management
5. objection functions and optimisation techniques to help to find the best solution.

CBM models - Deterioration Modelling

As the first constitutive component of a CBM model is often a qualitative description of the functions and features of the system, the modelling normally starts with the second step where a mathematical representation of the asset degradation process is required. Due to the uncertainties and randomness that exist in the change of asset condition, it is recommended that time-dependent stochastic processes be used to model deterioration [94]. The stochastic degradation is modelled with either a discrete or a continuous process when system condition is directly observable, and with a proportional hazard model (PHM) when it is not [2]. In PHM models, it is assumed that multiple factors, called covariates, together cause assets to degrade. For instance, Liu et al. [62] applied PHM to characterise the joint effect of cumulative damage and ageing on asset failure rate. Discrete stochastic process is best suited for systems where distinctions between states are obvious. The most commonly mentioned technique in the literature to characterise discrete degradation process is the Markov-based models, which assumes that the asset state in the next stage is solely determined by its current state and the transitive probability between them [45]. Chan and Asgarpoor [16] applied a Markov chain to describe the degradation states of a single component. Chen and Trivedi [18] built a semi-Markov decision process (SMDP) for joint optimisation of the inspection interval and CBM policy. The capability of such models to quantify the evolution of asset status via a transition probability matrix makes them suitable to be used to find an optimised

general policy where a finite number of alternative actions are available. There are, of course, other systems subject to continuous deterioration and hence cannot be adequately described by traditional Markov chains. This is especially the case when the accumulated damage to the system is in the form of erosion, corrosion, and cracks. In these cases, continuous stochastic process such as Brownian motion with drift, compound Poisson process, Gamma process, and inverse Gaussian (IG) process are considered to be more promising approaches [94]. These continuous models are similar to each other in that they all conform to the assumption of independent increments, yet differ in their applicability in characterising different deterioration mechanism:

1. Brownian motion with drift was originally used to capture share price and small particle movement whose value tends to move up and down with time. As a result, it is usually chosen to model the resistance of a structure, which tends to increase and decrease alternatively. Modelling the system degradation with a Brownian motion process , Guo et al. [40] obtained the RUL distribution and proposed a CBM policy considering mission constraints;

2. Gamma process is a monotonic jump process with independent and non-negative increments that follow a Gamma distribution. It is ideal for modelling degradation processes that are irreversible and gradual in nature, such as crack growth, corrosion, and creep. The extensive application of Gamma process in CBM for continuously deteriorating systems has been thoroughly reviewed by Van Noortwijk [94];

3. compound Poisson process is in many ways like Gamma process, except for that it has a finite number of increments within a finite time interval, whereas for Gamma process it is an infinite number of jumps. As a result, compound Poisson process is mainly chosen for modelling degradation caused by sporadic shocks. Examples of using compound Poisson process to characterise fatigue can be found in [88, 89];

4. IG process is a limiting compound Poisson process flexible in modelling heterogeneous degradation. It was proposed by Wang and Xu [96] as a complement to Gamma and compound Poisson process to model equipment deterioration for which these two are not suitable, such as the GaAs Laser data. The application of IG process in degradation modelling is relatively new, and there exist only a few papers addressing CBM problems with it [96, 103, 63, 19].

Despite their innovation introduced from various angles, all the aforementioned models have made the assumption that the degradation of equipment is an inevitable self-evolving process. The assumption has indeed simplified the modelling and solution seeking process, it

has however overlooked the fact that the rate of deterioration varies with the type of task and the amount of workload assigned to the production unit [42]. Consequently, the potential benefits of intervening in machine degradation have rarely been explored. For example, controlling the degradation of assets by adjusting its workload would enable asset owners to postpone or bring forward maintenance to a more preferable time. Stakeholders can gain more flexibility and control over their assets from this option, as now they may not have to wait for maintenance requests that can be triggered by any machine at any time.

A few attempts have been made to develop mathematical models that explicitly characterise the relationship between the degradation behaviour of a machine and the tasks or workloads it takes on. For instance, Iakovou et al. [47] used a continuous Markov process to describe the deterioration of machining tools, where the load-dependent aspect is accounted for by the inverse relationship between the useful tool life and the cutting speed. Similarly based on continuous-time Markov chains, Zhou et al. [108] developed a deterioration model for components in a load-sharing system. It is assumed in their work that the sojourn time is dependent on the load level selected for the component. Marseguerra et al. [68] modelled a serial-parallel load-sharing system using Markov chains, where components under higher burdens have amplified transition probabilities towards worse states. Based on accelerated degradation testing models under constant stress, an extension was proposed by Liao et al. [61] to enable reliability inference with varying stress conditions. It was assumed in their model that the asset deterioration follows a Brownian motion with drift process, and the extension was made possible by expressing the model parameters as functions of the stress. Yang et al. [100] focused on parallel machines with two alternative throughput modes and associated the fast mode with a higher failure rate. AlDurgam and Duffuaa [3] studied a system with different production rates using a POMDP framework, and the influence of various operation levels is represented by different transition matrices. Celen and Djurdjanovic [13] and Celen and Djurdjanovic [14] both modelled a flexible manufacturing system using Markov chains, where the operation-dependent property of the system is treated in a similar manner as in the work of AlDurgam and Duffuaa [3]. Jin [55] also developed a deterioration model using Markov chains with operation-dependent transition matrices, but took an analytical approach to study the structural properties of the objective function. Hao et al. [42] employed a Brownian motion with drift model to describe the degradation of machines, where the instantaneous degradation rate is proportional to the workload. A more recent publication that has adopted the same approach as in Hao et al. [42] was by Manupati et al. [66], which is an extension based on a previous conference paper of theirs [17].

It can be observed from the above that for discrete stochastic models, the effect of workload is reflected in the scale of transitive probability between states, whereas for continuous

models it is accounted for by the different choice of distribution parameters.

CBM Models - Deterioration Consequences and Maintenance Modelling

There can be various ways to model the possible consequences incurred by an over-degraded system. The most straight forward approach is to directly link the level of deterioration with breakdown, where an asset is considered to stop functioning once its degradation reaches a certain threshold. Consider the wearing and thinning of car tyres, the tyres have to be replaced once their thickness has dropped to a certain level. An application of this modelling method is demonstrated in Castanier et al. [12] where the increment degradation is described by an exponential distribution. For some other assets, the deterioration itself does not lead to breakdown, but rather increases the likelihood of failure. Transformers is one of the assets that belong to this category. The insulation of a new transformer is able to withstand severe incidents such as lightning strikes, which is hardly the case for an old transformer. A degraded transformer is more vulnerable to fluctuations in the environment and may fail under subtle disturbance. Liu et al. [62] adopted this approach by linking the hazard rate of an asset to both its age and level of degradation. It is argued by AlDurgam and Duffuaa [3] that apart from failures of assets themselves, deterioration may also lead to quality or performance problems related to the products or service produced. Following this argument, they developed a CBM optimisation model that associates a low quality rate with heavily degraded machines. Similar impacts can also be observed in the shortening charge cycle of phone batteries as they age, as well as in the increasing energy consumption rate of HVAC systems as dirt builds up in the tubes [95].

When an undesirable event happens to a system, regardless of its trigger being internal status change or external environment fluctuation, most CBM policies demand that maintenance actions be carried out to improve the status quo. The degree of maintenance, however, varies across different models. Some existing CBM models have been restricted to perfect maintenance actions, acknowledging that the system is brought back to the 'as-good-as-new' status once it has been maintained [30]. While this is acceptable for small machine pieces such as screws and nuts, which are normally replaced periodically, its application is limited when it comes to more complex systems. In order to capture the effects of imperfect maintenance, researchers have developed various models which usually take on one of the following routes:

1. maintenance is done on equipment in the form of minimal repairs that restore the system back to the stage exactly before the maintenance task is called for. Normally the degradation level of assets will not be modified by minimal repairs. In the model

developed by Huynh et al. [46], minimal repairs are adopted when assets experience soft failures caused by external shocks;

2. after maintenance, asset degradation is reduced by a random amount called the rectification factor. Researchers have proposed a couple of approaches to characterise the rectification factor. In the model proposed by Liao et al. [61], the post-maintenance deterioration of assets follows a beta distribution. The mean of the distribution is assumed to monotonically increase with the number of maintenance tasks performed, representing the impact of imperfect maintenance on the residual damage. Similarly but for a system with discrete states, Le and Tan [58] assumes that maintenance brings the system back a better state with probability P;

3. the focus of this approach is on how maintenance changes the degradation behaviour itself. For instance, Zhang et al. [106] developed a model with the assumption that rather than reducing the level of accumulated damage, maintenance will modify the system rate of degradation.

CBM Models - Objective Functions and Optimisation Techniques

Mainly four types of performance measures are adopted by CBM models for the design of objective functions - cost minimisation, reliability or availability maximisation, value maximisation, and other comprehensive metrics. The optimisation technique is often chosen based on the form of the deterioration model and the objective function. A review is given as follows on several widely-used optimisation criteria and optimisation algorithms in maintenance decision making.

Objective functions of CBM models often fall into one of the following classes:

1. **Cost minimisation** - A large number of maintenance optimisation models have been developed using cost-centric objective functions. Cost components in cost-centric models mainly include preventive maintenance cost, replacement cost, corrective maintenance cost, downtime cost, inspection cost, etc.. These models seek to find the optimal expected total cost over an infinite time horizon. Do et al. [34] proposed a maintenance model with both perfect and imperfect PM to minimise the long-run expected maintenance cost per unit time. Chen et al. [19] also chose the long-run expected cost as the objective function, but included an extra parameter r to quantify the time value of money;

2. **Reliability/Availability maximisation** - In cases where the estimation of cost parameters is difficult, system availability/reliability becomes a practical indicator for maintenance efficiency. Availability is also deemed to be an important performance

metric for systems with huge downtime costs. Liao et al. [61] designed a CBM model to maximise the average short-run system availability by setting a PM threshold. Zhu et al. [111] considered a competing risk maintenance situation and developed a model to maximise the achieved system availability. Other examples belonging to this category can be found in Klutke and Yang [57], Biswas et al. [8];

3. **Value maximisation** - Marais and Saleh [67] stated that existing cost-centric models are blind to the value of maintenance, which can lead to sub-optimal maintenance policies, and proposed a framework for capturing and quantifying values generated by maintenance activities. One of the major assumptions behind the argument is that the value of an engineering system is assessed by the market in terms of the service it provides over its life time. By considering the business-market conditions, Godoy et al. [37] has designed a model to optimise replacement intervals of critical components. Part of the emphasis of their work is on the value-adding effect of postponement or acceleration of replacements;

4. **Other comprehensive metrics** - while the optimisation criteria mentioned above each favours a specific dimension of system performance, models using more comprehensive metrics have also been developed. For instance, Overall System Efficiency (OSE), proposed by AlDurgam and Duffuaa [3] as an extension of the popular Overall Equipment Efficiency (OEE), is adopted as the objective function for a Partially Observable Markov Decision Process (POMDP) maintenance model.

Once the problem has been formulated, the last step is to accordingly devise an appropriate algorithm to find the optimal solution. For most cost-centric models for continuously degrading systems, traditional optimisation methods such as gradient descent and stochastic gradient descent are considered sufficient. Though customised variation exist among the actual implementation process, heuristics are popular for models based on a discrete stochastic degradation process. For example, Yang et al. [99] used elitist non-dominated sorting genetic algorithm (NSGA II) to solve a multi-objective maintenance optimisation problem. In order to find the optimal maintenance and production sequence solution, Celen and Djurdjanovic [15] applied a metaheuristic method called Tabu search algorithm based on the results of a discrete-event simulation of the system studied.

2.4 Integrated Maintenance-related Decision Making

This section reviews recent work dedicated to various aspects of integration attempted to save operational costs and improve production system efficiency. The emphasis is placed on

research that attempts to explore alternatives apart from traditional repair and replacement actions. These alternatives are usually in the form of adjusting sequence of operations and reallocating operation tasks in order to combat machine degradation and failure. The review paper of Hadidi et al. [41] categorised literature on this topic according to the approaches taken to account for the integration. Here, we adopt the same categorisation. The first type, referred to as interrelated models, aims to optimise one element while the others are formulated as constraints, whereas the second type, the integrated ones, simultaneously build into the model at least two production-related factors. This section surveys related decision support models in both categories, which for simplicity will both be called integrated models. A substantial proportion of the following content will be focused on the combined maintenance and operation decision-making of machines with operation-dependent degradation process.

2.4.1 Integrated Maintenance, Production, and Quality Models

This part reviews research work dedicated to decision making models that address at least two of the elements among maintenance, production planning, and product quality. Yao et al. [101] proposed a joint maintenance and production policy for an unreliable production-inventory system. The decisions to be made are both how much to produce and whether or not to conduct PM for various system states. It is assumed in their model that asset failure is time-dependent. Rausch et al. [81] designed a model to find the optimal inventory level and CBM threshold to minimise the operating cost while conforming to the constraint of stock-out probability. Ben Ali et al. [7] presented a job shop problem of the simultaneous scheduling of production and preventative maintenance (both time-based and age-based) to optimise two objectives: the total make span and the total maintenance cost. A genetic algorithms is applied to find a good enough solution. Radhoui et al. [79] developed a combined quality control, CBM, and buffer stock optimisation model for a production system that randomly produces both conforming and non-conforming parts. The limiting threshold for CBM is in the form of the rate of non-conforming products. Jafari and Makis [49] applied a continuous-time Markov process to describe a stochastically deteriorating system and presented a model to find the economic manufacturing quantity (EMQ) and the CBM threshold that minimises the long-run average cost. Another piece of research along this line was done by Peng and van Houtum [76] to optimise the same decision variables but for a system whose degradation is characterised by a Gamma process.

2.4.2 Integrated Maintenance and Maintenance Resources models

While the assumption of unlimited maintenance resource largely simplifies the decision-making process, it is hardly the case in reality for two reasons: 1) it is very costly to have all the required maintenance resource in place at all times; 2) even if feasible cost-wise, not all failed machines can be maintained simultaneously due to other limiting factors such as time windows and space constraints. As a result, a number of maintenance models have been formulated that also account for the resource elements, such as personnel, spare parts, and special equipment needed for maintenance tasks. A few examples are given here for the purpose of illustration. Taking into account human resource availability, Bouzidi-Hassini and Benbouzid-Sitayeb [10] proposed an agent-based production and maintenance scheduling system, where a Human Resource Agent is assigned to represent a personnel with necessary skills to intervene in maintenance tasks. Ilgin and Tunali [48] studied the joint optimisation of maintenance and spare part provision of an automotive factory using a metaheuristic algorithm and simulation modelling. Zhou et al. [109] developed a multi-agent system to heuristically solve the maintenance scheduling problem for a bus fleet, with the constraint that for a given time slot, a maintenance bay is only capable of handling one type of repair work on one bus.

2.4.3 Integrated Maintenance and Workload/task Allocation Models

Though there exists abundant work that simultaneously considers maintenance and the above production elements, the decision making of maintenance with workload and task allocation has not been addressed until very recently. Only a few papers consider the varying degradation rates due to different operations performed and attempt to actively intervene machine degradation process to improve system-level performance. These works will be reviewed in details in the rest of the subsection.

To our best knowledge, one of the earliest attempts to supplement traditional repair and replacement activities with operation alternatives is made by Yang et al. [100]. A model was proposed for optimising joint planning of maintenance and throughput changing operations in a manufacturing system consisting N machines in parallel. Part of the machines have two possible throughput modes - one slow and one fast mode. A linear machine life reduction model is assumed in this paper where the fast mode corresponds to a larger life reduction rate, and vice versa. In order to maximise the expected overall production benefit under constant production targets, the solution variables are coded into a matrix chromosome and various versions of Genetic Algorithm (GA) have been used to find the optimal combination of maintenance and throughput setting.

In a more recent paper, an explicit mathematical model to tie maintenance, production rate, and product quality within the Partially Observable Markovian Decision Process (POMDP) was proposed by AlDurgam and Duffuaa [3]. This model also takes the basic assumption that higher operation speed results in higher machine deterioration rate, which is reflected in the different transition matrices. The value iteration algorithm is applied to find the set of maintenance and production rate that maximises the expected total Overall System Effectiveness (OSE) over a finite time horizon, a concept extended from Overall Equipment Effectiveness (OEE). The optimal strategy, in the form of a projection that maps a state occupancy vector in the belief space into a decision on the most appropriate action to be taken, is obtained for a three-state single-machine system with three maintenance options (do nothing, repair, and replace) and two speed options. However, the decision-making model proposed in their work is targeted at single-machine systems, yet in reality it requires the cooperation of multiple machines to achieve production goals. Moreover, no constraint is considered to obtain the optimal solution while constraints such as production demands are likely to exist in reality.

Another piece of work on maintenance decision making with operation-related alternatives is conducted by Zhou et al. [107] where reconfiguration is considered as a means of mitigating production loss caused by machine degradation and failure. A framework is proposed for the integrated decision making of re-configuration and age-based preventive maintenance for a generic two-stage parallel-serial system with re-configurable capabilities to transfer operations between the two stages. At each time epoch, a decision is chosen from conducting preventive maintenance, re-configuring the system, and doing nothing to maximise the system throughput rate according to its current machine age and system reliability. A two-phased heuristic search strategy is adopted in cooperation with discrete event simulation to find the best combination of age-based preventive maintenance threshold for each machine and the reliability threshold for the system. The paper by Zhou et al. [107] gave guidance on how to couple system reconfiguration with maintenance actions. However, the optimal solution is obtained under the constraint imposed by the operation transfer strategy that the system throughput has to be put at its maximum at every decision-making point.

A couple of papers published in the past few years have been motivated by the use of highly integrated and highly flexible production systems in the semiconductor industry. Such systems are often able to perform more than one operations that degrade the systems at various rates. Celen and Djurdjanovic [13] looked into the dynamic interactions between machine degradation, operations, and product quality, and devised a combined operational and maintenance decision-making policy. The policy is tested on a generic cluster tool with multiple production chambers commonly seen in semiconductor manufacturing. The

degradation process of each chamber is assumed to be fully observable and is modelled as a series of Markov Chains where the various effects of different operations are characterised by non-identical transition matrix. One unique feature of this model is that it acknowledges the fact that preventive maintenance triggering states are also operation-dependent and vary from chamber to chamber. The goal is to optimise a customisable objective function while attempting to meet the production demand within a limited time frame T. In the metaheuristic optimisation method used in this paper, a set of candidate solutions is first generated by the local search technique, Tabu Search (TS), and each feasible solution is evaluated using the average value of the objective function obtained from multiple runs of discrete-event simulation, which is then fed back into the TS algorithm. In their later work [15], the model is expanded on the condition that maintenance conducted at 'less busy shifts', such as weekends and evenings, costs less than maintenance executed at normal working shifts. In both their papers, it is assumed that the first task in the incoming queue is assigned to the least degraded machine.

Similarly targeted at operation-dependent deteriorating systems under various operation conditions, Jin [55] took an analytical approach to explore the structural properties of the objective function. The aim of their study is to identify the conditions that can limit the optimal solution to a set of monotone procedures, which has the potential to largely reduce the number of candidate solutions and thus simplifies the computational process.

Hao et al. [42] has come up with a load allocation and maintenance strategy. A decision support model is developed that strategically adjusts the workload assigned to each unit dynamically. The motivation that underpins the model is that by actively controlling the residual life of parallel units, two objectives could be achieved: 1) the prevention of overlapping of unit failures; 2) the prevention of a too high average degradation rate among all units. Specifically, higher workloads are assigned to less healthy units so that the more degraded units can fail even sooner. Unit degradation is modelled with a linear stochastic differential equation (SDE) and the degradation rate is assumed to be a random variable following a continuously updated distribution. To account for the fact that higher loads lead to faster degradation, it is assumed that the instantaneous degradation rate is proportional to the workload multiplied by a degradation coefficient. A simulation-based numerical example of dynamic workload adjustment among five identical units is presented and compared against two benchmark strategies. However, in their work no explanation has been given on what the decision-making interval is and how it is determined, which has a large influence on the performance of the model. Moreover, a potential upside of system performance has been omitted in their model as the threshold for preventive maintenance is given rather than obtained as part of the solution.

A more recent publication by Manupati et al. [66], which is an extension on one of their conference papers [17], has adopted the same approach as in Hao et al. [42]. The innovation of their work lies in that they have extended the dynamic workload assignment policy to account for not only pure parallel systems, but also hybrid systems. The performance of the proposed strategy is also compared with that of uniform and random allocation.

2.5 Research Gap

It can be observed from the above discussion that there exists a clear trend towards a more comprehensive practice of maintenance decision making. To be specific, the literature review indicates that substantial benefits could be realised from the integration of CBM decision making and other related production factors. Indeed, a growing number of optimisation models have been developed to exploit the intrinsic connection between maintenance, production lot-sizing, inventory planning, and human resource management. However, very little research effort has been dedicated to the joint optimisation of task/workload allocation and CBM decision making, despite the obvious impacts of the two on each other.

Besides, most of the existing work attempting to tackle this problem exhibit at least one of the following three limitations: 1) as the optimal solution is obtained based on the assumption of long-term steady state, it lacks the flexibility of providing a timely and 'good-enough' solution to cope with volatile circumstances such as a sudden demand change; 2) the model is built for a single-asset system and lacks generalisation for an asset fleet; 3) it is either targeted at time-based and age-based maintenance, or treats the CBM threshold as given instead of a decision variable. Namely, the joint decision making of CBM and task/workload allocation is not fully realised. A detailed presentation of the aspects that each of the models fails to address can be found in Table 2.2.

Research gap: it follows from the above analysis that there is still a lack of a dynamic optimisation model that ties together the decision making of both the CBM threshold and the task/workload allocation among a fleet of parallel assets. In order to address the research gap identified in this chapter, the key research questions that need to be answered are listed below:

1. what are the constitutive components for a dynamic optimisation model for joint decision making of CBM threshold and task/workload allocation for an asset fleet? Later discussions in Chapter 4 will reveal the two important components of such a model: a machine-level workload-dependent CBM optimisation model and a system-level task/workload allocation strategy.

Table 2.2 Lack of considerations in studies that involve the integrated decision making of maintenance and task/workload allocation

Literature	Features to be considered		
	Flexibility	Aimed at an asset fleet	Integration of task/load allocation and CBM
Yang et al. [100]	✗	✓	✗
AlDurgam and Duffuaa [3]	✗	✗	✓
Zhou et al. [107]	✗	✓	✗
Celen and Djurdjanovic [13]	✗	✗	✓
Celen and Djurdjanovic [15]	✗	✗	✓
Jin [55]	✗	✗	✓
Hao et al. [42]	✓	✓	✗
Manupati et al. [66]	✓	✓	✗

2. How can we quantify the impact of workload on the degradation behaviour of individual assets, and how can such knowledge be used to set the most appropriate CBM threshold?

3. How can we quantify the impact of a specific workload allocation on maintenance and production at the system-level, and what type of information is needed in order to do this?

2.6 Chapter Summary

The literature review chapter has established the theoretical background for this research topic. It also served to identify the research gap that will be addressed later in the thesis.

In this chapter, key concepts and elements in asset management have been introduced, with an emphasis on the important role of maintenance decision making. The scope of this thesis is then further narrowed down to maintenance decision making for a fleet of parallel assets. A clear evolution trend of maintenance strategies can be observed - research focus is shifting towards proactive and more comprehensive maintenance decision making. A closer analysis of the trend has revealed that, among all the interrelationships between various production elements, that between condition-based maintenance and task/workload allocation among a system of assets is considered as one of the least sufficiently studied pairs. After

digging into the optimisation models on integrated maintenance and operation, we found that there is currently a lack of a decision making strategy to improve system-level performance by dynamically generating combinations of CBM thresholds and task/workload allocation. We have also proposed the three key research questions that need to be answered in order to address the research gap. The first one is to identify components of such a strategy; the second one calls for a task/load-dependent degradation and CBM optimisation model for a single asst; the last one concerns the approach to reach a system-level solution of both maintenance and task/workload allocation as well as obtaining the right information needed to achieve this.

In order to supplement the literature findings with a practical perspective, the next chapter presents a couple of exploratory case studies on maintenance and production decision making in the industry. The insights obtained from the case studies are also a critical step in the problem formulation process of this research.

Chapter 3

Industrial Rationale

3.1 Introduction

The previous chapter has set the academic background for this research by providing a detailed review of papers related to the topic studied in this thesis. The discussion around existing literature has identified the need for a decision making strategy to improve system-level performance by generating combinations of CBM thresholds and task/workload allocation. Apart from the findings from a pure theoretical point of view, it is also essential to look at this problem from a practical angle. The aim of this chapter, therefore, is to provide an industrial understanding on the necessity of such a decision making strategy. The three objectives of this chapter are as follows:

1. to provide insights into the specific industries and asset configurations with the most potential benefits from integrated decision making of task/workload allocation and condition-based maintenance.

2. to give an overview of the current practice that companies adopt for maintenance and operation decision making in order to highlight the lack of a systematic approach.

3. to back up the academic gap in knowledge with an industrial rationale

The outline of this chapter is shown in Fig 3.1. The chapter will start with a description of case studies in the academic literature, which reviews industrial cases and problems encountered by other researchers that are of relevance to the strong dynamics between maintenance and task/workload allocation. The discussion will then move to the exploratory case studies, presenting the methodology adopted and the findings about the challenges faced by companies that has the need to integrate operation and maintenance decision making.

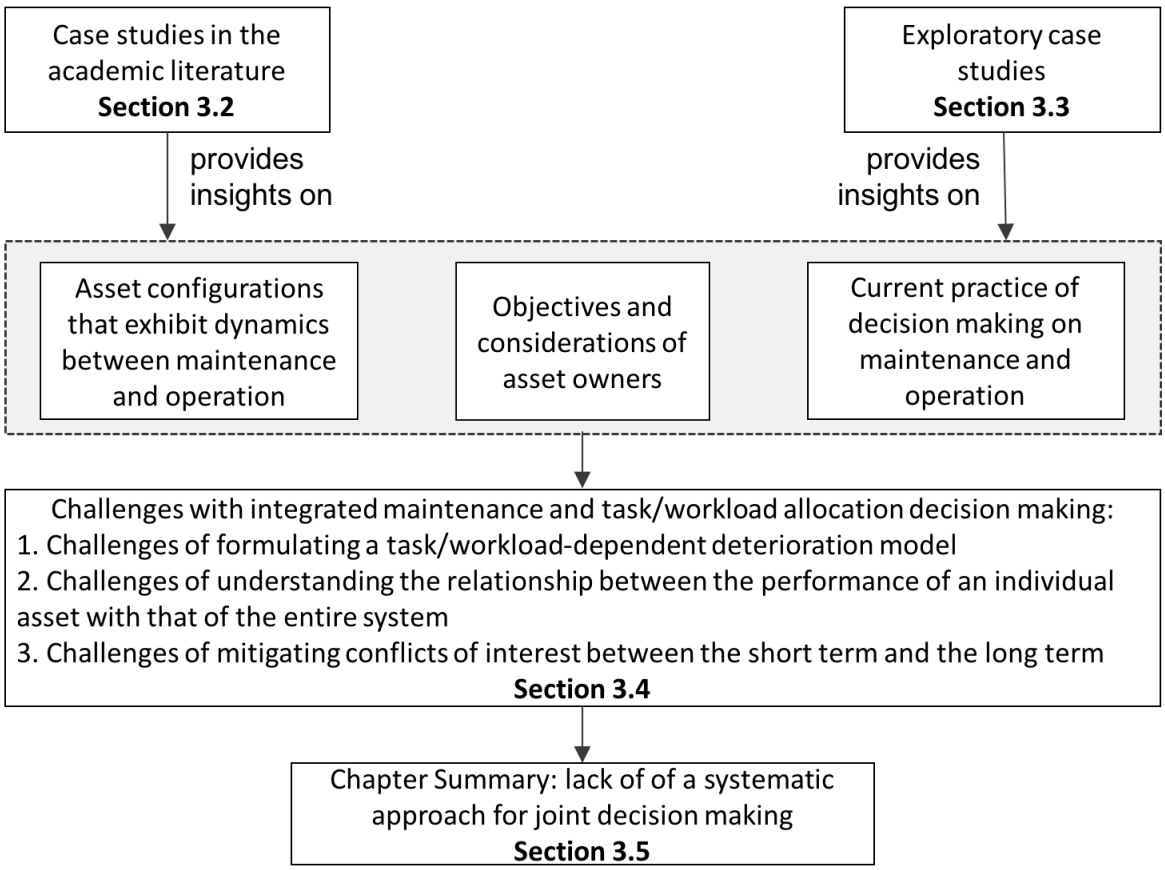


Fig. 3.1 The story-line followed by the industrial rationale chapter

3.2 Industrial Case Studies in the Academic Literature

This section aims to introduce the typical industrial settings that exhibit close interactions between task/workload allocation and maintenance planning by referring to examples mentioned in the existing literature. The type of systems identified in the academic literature normally belongs to one of the following two categories: 1) load-sharing systems, which can normally be found in process industries, power plants, transportation and logistics systems, computer processors, and pump facilities; 2) flexible manufacturing cells [15]. An overview of the cases found in the academic literature is given in Table 3.1.

Ye et al. [102] described a practical problem faced by a water utility company which is in need of a decision making strategy to manage its assets across the network. The focus of their study is on Rapid Gravity Filter (RGF), an asset very key to the water treatment business. The filtration process starts with having raw water flow into RGF, and the water is cleaned and filtered down the media by gravity. RGF degrades with use and the level of degradation can be estimated by direct observation of the media, or by inference from the quality of

Table 3.1 List of case studies in the academic literature

Literature	Industry	Asset studied	Relationship among assets in a fleet
Ye and Chen [103]	Water utility	Rapid gravity filters	Parallel
Yu and Chan [105]	HVAC systems	Chillers	Parallel
Keizer et al. [56]	Gas utility	Pumps	Parallel
Celen and Djurdjanovic [15]	Semiconductor	Flexible wafer manufacturing cells	Hybrid

the output water. In most cases multiple RGFs exist in a water treatment system, and the number of RGFs is normally determined by the amount of water that is to be filtered at peak times. At off-peak periods, not all RGFs are required to be activated and a choice needs to be made on which ones are to be put into cold standby. The current practice, according to their conversation with the company, is that a human operator makes the decision on how many and which RGFs should be up and running based on his own estimation of the deterioration status of the RGFs. The purpose here is to balance the level of degradation among the filters and to ensure the approximately simultaneous failure of all RGFs. Over-simplified the decision rationale may seem, it indicates the intention of the company to realise potential benefits by actively managing its asset degradation using workload adjustment. The authors of the paper proposed a CBM model where the threshold is based on the cumulative usage of RGFs. However, the way that the operator decides on which RGFs to use at each time epoch is taken as given. Consequently, the potential benefits that can be realised by simultaneously optimising workload allocation and CBM thresholds have been overlooked.

Yu and Chan [105] considered the need for a load sharing strategy in the case of multiple chillers in HVAC systems. The chiller fleet operates in parallel in order to meet the cooling requirements of an air-conditioned building. The efficiency of chiller systems is measured by an index called the aggregate coefficient of performance (COP), which is dependent upon the working environment, the compressor efficiency, the part load ratio at which each chiller is operating at, etc.. As the part load ratio is mainly a function of the volume of chilled water flowing through and the change in water temperature, it is usually the same for all chillers in conventional pumping systems with a uniform temperature rise and equal water flows. With the growing support for installing variable-speed pumps for HVAC systems due to its energy saving effect [43, 5], the need naturally rises for a load allocation policy to optimise the aggregated COP. Though the case presented in Yu and Chan [105] has barely

touched the role that maintenance plays in this issue, the link between maintenance decision making and load sharing strategy has been implicitly established in another paper by Wang and Hong [95]. Wang and Hong [95] studied the impacts of various types of maintenance practice on the energy performance of the HVAC system of a large office building in Chicago. The results demonstrated the substantial role maintenance plays in improving the efficiency of HVAC systems. From the above discussion, we can conclude that the performance of chillers that operate in parallel in a HVAC system can potentially be improved in practice by joint decision making of maintenance and load sharing for the following reasons: 1) there is proved relationship between part load ratio, chiller efficiency, and maintenance activities; 2) there exists a tendency to install variable-speed pumps for chillers, which enables load control in such systems. Furthermore, it can be inferred that one important piece of jigsaw is missing - an approach to quantify the impact of workload on the degradation behaviour (potentially characterised by the level of efficiency) of chillers.

Keizer et al. [56] mentioned a real-life problem they encountered while interacting with a gas company that runs the business of pumping gas and distributing it to its customers. As gas storage is not a viable practice, the company has to continuously produce gas for both residential and industrial needs. It is therefore very important to have high pump availability at all times. The benefits of having redundant pumps, as mentioned in their case study, include the following: 1) redundancy ensures that the demand can be met even in case of pump failures; 2) with more pumps than needed, the load that each component undertakes is reduced, which mitigates the failure rate. The above benefits, of course, support the view that a failed pump should be repaired or replaced immediately upon failure. It contradicts, however, the fact that clustering the maintenance of multiple components can lead to lower set-up costs. Again, similar to the case presented by Ye et al. [102], the way load is allocated is taken as given, where the operational pumps share an equal share of the total demand. The link between workload and degradation is established by having the failure rate to be a function of the number of units in operation. Such 'passive' load sharing scheme is far from optimal, as the asset owner has no control over when failures will be triggered for each pump. It follows that the company might benefit economically from making maintenance decisions while consciously taking into consideration how much load each pump should bear.

The last case to be presented here from the academic literature is related to flexible manufacturing cells, such as cluster tools in the semiconductor factory mentioned in Celen and Djurdjanovic [15]. A cluster tool usually consists of multiple chambers that are capable of handling more than one type of operation. A separate system feeds materials into and takes out finished products (in this case wafers) from these chambers. Different wafers require different operations and often these operations have varying impacts on the degradation of

the chambers. With the chambers being multi-functional, there exists substantial flexibility in assigning operations to chambers. If properly exploited, this flexibility can prove to be a favourable trait for maintenance decision making. For instance, when one of the chambers is heavily degraded to the point that it cannot sustain another demanding task without having to be maintained the next day, a less abrasive task can be assigned to it in order to postpone its maintenance to a more preferable time. The inherent interaction between maintenance decisions and the rules of product sequencing and dispatching thus implies that these two elements need to be considered simultaneously for system performance optimisation. In addition, due to the nature of flexible manufacturing cells, such decisions are usually readily executable. Celen and Djurdjanovic [15] modelled the task-dependent degradation behaviour using a Markov process where the transition matrix is a function of the task assigned to each chamber. Their approach is also constrained by a pre-defined rule which demands tasks to be assigned first to newer chambers.

The above discussions around case studies found in the academic literature indicate that having a fleet of assets, including redundant ones, working in parallel is a common practice for companies in various industries, such as those in water utility, oil and gas, and HVAC service. Flexible manufacturing cells can be seen as a more complex and hybrid entity that also exhibit features of parallel systems. Furthermore, it is obvious that asset owners have control, or at least there is a tendency to enable them to take control, over the type of tasks or the amount of load to be allocated to individual assets. Besides, it is evident that the intrinsic coupling effect between maintenance and operation has been widely acknowledged. In some cases, as mentioned in Ye and Chen [103], the company even has a clear understanding of the benefits of actively intervening the degradation process of its assets and has already been doing so for a while, although not in a very systematic manner. The review of academic literature case studies has pointed out the industries and specific system configurations where maintenance and operation are closely interrelated. It also shed some light on the approach currently being used to handle such interactions. However, the common practice of workload allocation is still quite passive or follows a rule of thumb. Furthermore, it is not clear yet what factors are considered for joint decision making of maintenance and task/workload allocation from an industrial perspective. There is also a lack of understanding of the challenges faced by manufacturing processes involving parallel assets. Exploratory case studies have been conducted to address these questions.

3.3 Exploratory Case Studies

The aim of this section is to provide insights into the status quo of integrated decision making of operation and maintenance from a practical point of view. The methodology applied in the case study is introduced first, followed by a description of major findings from exploratory case studies throughout the research project.

3.3.1 Case Study Methodology

An appropriate approach needs to be designed in order to realise the two objectives of conducting exploratory case studies: 1) to understand the challenges faced by and the practical needs of organisations in terms of exploiting the impact of task/load allocation on maintenance decision making; 2) to narrow down companies that can potentially be turned into real case examples of the model developed in this research project. The steps taken in order to achieve the above mentioned objectives are listed as follows:

1. **Selecting case companies**: as the target of this study is asset fleets, the most important criteria for the selection of case companies is on the existence of assets operating in parallel. Organisations that typically have redundancy in their manufacturing facilities include utilities, oil and gas, transportation, and consumer packaged goods companies. Three cases are selected for this study and according to Yin [104] this number is sufficient since similar results have been obtained from all of them.

2. **Data collection**: four steps are followed in this process in order to obtain accurate and sufficient information.

    - **Preparation of questionnaires and data chart** - the first step involves establishing a preliminary understanding of the manufacturing process, nailing down the company departments that might possess useful information related to the research topic, drafting questionnaires to be used over the interviews, and preparing data charts.
    - **Semi-structured interviews** - this step involves attending presentations given by and conducting interviews with personnel working with relevant divisions within the company, which in most cases are the maintenance/reliability, operation, and finance departments.
    - **In-person observation** - this is done along with the interviews over plant visits and involves directly observing the physical equipment as well as the operation and maintenance practice associated with it.

- **Follow-up telephone conference calls** - this is carried out to collect missing pieces of data if not all of them have been obtained from the previous interviews and on-site visits.

3.3.2 Case Description and Findings

Some basic information regarding the case companies and the type of assets being studied can be found in Table 3.2. During the process, both qualitative and quantitative data have been collected. Qualitative data has been used to understand the deterioration mechanism of assets as well as to define the problem in a more descriptive way, whereas quantitative data is collected for model formulation and assessment.

Table 3.2 List of exploratory case study companies

Company	Location	Industry	Maintenance Regime	Asset Studied
Company A	Southampton, UK	Oil and gas	TPM and CBM	Secondary effluent treatment vessels
Company A	Southampton, UK	Oil and gas	TPM and CBM	Demineralisation vessels
Company B	Cambridge, UK	Packaged foods	TPM and CBM	Conveyor belts

Company A - Asset Type I

Company A is a refinery plant of a multinational oil and gas corporation. The major function of this plant is to process crude oil transported to the refinery by sea and convert it into a wide range of products, such as petrol, marine fuels, heating oil, and diesel. The plant handles the entire process from the provision of utilities such as water, steam, and air, to the separation of crude oil by distillation, to the treatment of waste water before it is discharged into the river. The company follows a reliability-based maintenance policy and categorises its assets into five risk levels labelled from A to E, which correspond to different failure rates. An asset is categorised into a certain risk level based solely on its age, regardless of how heavily the asset is used over its life-cycle. Most of the systems involved in the process have layers of redundancy within them, but it has been less obvious to the company how to manage its assets in more complex cases than a traditional ‘one in service, one spare’ system. This has led to under-maintenance of equipment, and consequently more expensive reactive work when problems arise.

For the purpose of this study, a particular type of assets is considered - the secondary effluent treatment (SET) plant. The SET plant consists of seven treatment vessels that contain layers of anthracite, sand, and gravel to remove oil and suspended solids from the effluent. It is considered to be very critical by the management due to the huge downside potential of failing to meet the environmental regulation. The internal walls of the vessels are lined with an epoxy glass flake material called Plasite 4550S for corrosion protection. Degradation of the internal lining is a gradual process which is caused by the following three factors:

1. **Ageing**: this is the natural ageing process of the lining material.

2. **Erosion**: when the effluent travels through the refinery to the SET plant, it is likely to pick up leaks with erosive chemicals on its way. As the effluent flushes through the vessel walls, the erosive chemicals are likely to cause damage to the internal lining;

3. **Scratches**: the materials in the filter media, especially hard solids such as sand and gravel, are flushed up and down the vessel walls over back-washes, and are likely to leave scratches on the internal lining.

If the internal lining deteriorates to the point where it is not providing protection to the pressure vessels, significant corrosion of the underlying vessel walls will take place. It suffices to say that the status of the lining is a very important indicator of the health of the vessels. Another observation is that the load assigned to a vessel has both direct and indirect impact on its rate of degradation. The direct impact comes from the chemicals in the effluent eroding the lining as it flows through the vessel. The indirect impact is implied by the fact that if a vessel is used more often than others, it will go through more frequent back-washes to regenerate its layers of filter media, which leads to a higher probability of scratches on the lining. Though the management acknowledges the fact that the deterioration of vessels is a workload-dependent process, it is not yet considered in the current maintenance practice.

As the SET plant has multiple redundant vessels, it can maintain its desired capacity with two filters off line for back-washes or maintenance. Due to this considerable flexibility, in devising a maintenance policy the team has assumed that a ‘run-to-failure’ strategy is acceptable. The strategy has worked well until recently when all vessels have deteriorated to a very poor condition and the plant is at serious risk from multiple vessel break-downs. There is currently a major programme underway of re-lining all the vessels, and an effective strategy is thus needed to address the maintenance issues of a fleet of vessels to avoid getting into a similar situation in future. Another interesting observation during the plant visit concerns the operation of the vessels. The company has a team of operators that each day choose the vessels that should be put into service. The purpose here is to ensure enough capacity

for effluent treatment while pursuing least operating cost. However, for now the decision is made on an ad hoc basis. For instance, the operators tend not to use two of the seven vessels as their back-washing procedure is not completely automatic. Another vessel is also less frequently used due to the problem with one of its valves. Despite its high-cost nature, maintenance has rarely been considered while the company is determining how to operate the vessels on a day-to-day basis.

It can be found from this case study that the practitioners understand that degradation is highly correlated with workload, yet they lack the tools to quantify such relationship. Moreover, as there has been no guidance on which vessels should be used on a daily basis, the choice is made without a comprehensive understanding of how these short-term workload decisions will affect the long-term performance of the systems.

Company A - Asset Type II

The target company here is still the same as the previous case, but the focus is on another type of assets it owns - the demineralisation (Demin) Plant. The purpose of the Demin Plant is to provide a secure and reliable supply of fully demineralised water for use in main boiler systems on the site. Fully demineralised water is water that has had all dissolved salts removed and will not cause any scale deposition in boilers operating at high pressures.

The plant consists of four types of vessels - D-1, D-2, D-3, and D-4. The plant has 6 parallel units of each type D-1, D-2, and D-3 apart from the polishing units D-4 of which there are 5. All vessels of the same type are between a common inlet and outlet manifold system that allows any vessel to be shut down or taken off line at any time without affecting the plant throughput, subject to the minimum number of vessels being in service. Each vessel has its own controller that equalises the flow between all vessels. The controller measures the volume of water treated by the vessel and puts the vessel into the regeneration sequence when the predetermined volume has been treated. Similar to the SET plant, the internal lining of the vessels is considered critical since it prevents them from corrosion caused by strong acid in the water being treated.

With multiple spares, a run-to-failure strategy results in an accumulation of defects and reliability vulnerabilities in the system, which calls for a comprehensive maintenance strategy that can guarantee high reliability at the fleet level. Moreover, there exists an extra layer of complexity apart from the interaction between workload and vessel deterioration, as mentioned by the reliability manager of the company, that comes from the impact of workload on the regeneration sequence. As the ion beds in the vessels need to be replaced after a predefined number of regenerations, their life span is also highly correlated with the accumulated workload that a vessel has undertaken. While one vessel is down, extra

workload goes to other vessels, which interrupts the regeneration sequence of the ion beds. The company has noticed the above intrinsic connections and is in need of a maintenance decision making strategy that can incorporate operation considerations as well as its influence on the regeneration sequence.

Clearly, the operation-maintenance puzzle here is another call for an approach that explicitly characterises the relationship between workload and asset deterioration. Furthermore, the discussion above regarding the regeneration of the ion beds also indicates that the performance of the system can be largely shaped by that of an individual asset in the system. If a conscious decision aiming at improving system-level performance is to be made, such influence needs to be incorporated into the decision-making model adopted.

Company B

Company B is one of the UK's leading food companies with multiple manufacturing sites across the UK. Its parent company owns over 30 sites world-wide. Company A manufactures both its own brands and private label foods ranging from juice, jams, and ready-to-eat soups, all out of natural ingredients. The factory selected for this case study consists of a number of highly automated production lines that are dedicated to the production of jams, honey, jelly, as well as juice, and one independent Business to Business (B2B) production cell that processes ingredients for downstream enterprises. The production lines normally run 24 hours a day, seven days a week, whereas the independent cell will mostly operate on demand. Though relatively independent, the three production lines in the factory possess certain reconfiguration flexibility. For instance, the line that is struggling to meet a high demand may temporarily expropriate the conveyor belt or packaging facility of another. The question for the operations planner is now whether he should reconfigure the process every time in order to meet a rising demand. Multiple factors need to be considered in this scenario, such as the setup cost, cost for loss of production, cost of extra deterioration to the conveyor belt or packaging facility, etc.. This requires a clear understanding of the impact of the status of one asset on system performance both in the short term and long term.

In terms of maintenance policies, time-based preventive maintenance is performed on most assets in the plant. A few pilot projects of condition-based maintenance, however, have been conducted on critical assets, which the reliability manager believes may lead to a 30% reduction in maintenance costs. One interesting part of the pilot projects is a newly purchased filling machine manufactured by an OEM in Italy. The machine is equipped with data communication device that enables its OEM to monitor its health condition and make changes to the control system embedded within the machine accordingly. This implies that

from the perspective of an OEM, potential benefits might arise from actively managing the operating condition of the assets based on its health status.

3.4 Challenges in Integrated Decision Making

From the analysis of cases in the academic literature and exploratory case studies, we can summarise the challenges faced by the industry to be the follows:

1. Lack of a mathematical model to quantify the impact of task/workload on the path of asset degradation.
2. Lack of a clear understanding of how the performance or condition of individual assets affects the overall performance of an asset fleet.
3. Lack of a mechanism to reach a balanced trade-off between short-term goals and long-term goals.

Together with findings from the literature review in Chapter 2, the challenges mentioned above are very useful for this research project in multiple ways: 1) these findings provide insights into the industrial settings where the dynamics between operation and maintenance needs to be considered, which helps with identifying the research question; 2) they also revealed the key factors that asset owners care about while making operation or maintenance decisions, which gives hints on the formulation of mathematical models; 3) they provide some assurance that the model developed in this thesis can potentially be applicable to certain asset configurations in the industry. How these challenges have been used for formulating and addressing the research questions identified in Chapter 2 can be found in Table 3.3. A more detailed discussion about each of the challenges is presented in the following section.

Table 3.3 Challenges used for formulating and addressing research gaps

Challenges found from case studies	Research Question 1	Research Question 2	Research Question 3
Impact of task/workload on asset deterioration		✓	
Impact of individual performance on system performance	✓		✓
Impact of short-term performance on long-term performance	✓		✓

3.4.1 Impact of Task/Workload on Asset Deterioration

It can be observed from both the academic literature cases and the exploratory cases that the operators understand intuitively that a more frequently used asset degrades faster than an asset often in standby. The correlation between task/workload and asset deterioration, however, is at most in an over-simplified format, such as the deterministic usage-based estimation in the case referred by Ye and Chen [103], if not entirely qualitative - such as gut feelings of the operator. To represent the stochastic nature of deterioration and the unnecessarily linear relationship between workload and the rate of asset degradation, a more generic mathematical model is needed.

3.4.2 Impact of Individual Performance on System Performance

All three cases mentioned in this chapter require multiple assets to work in coordination to achieve a common goal. It is therefore crucial to base operation and maintenance decisions on the fleet-level performance. Due to the various forms of dependence between assets in a fleet, the relationship between the performance at the asset level and that at the fleet level can be very complex. In cases where the operator is aware of such a relationship, certain measure is taken to exploit it for lower maintenance costs, as illustrated by Ye and Chen [103] where the operator intentionally balances the deterioration among assets for group replacement. Opposite results can be observed if such relationship is neglected. For instance, Company A has underestimated the impact caused by one asset loss in a degraded fleet, and thus ended up in a situation where the system is at very high risk of failure. This observation indicates that a more structured understanding of the impact of individual asset condition on system-level performance is likely to generate additional benefits to asset owners. It also provides guidance on the formulation of objective functions in the decision making model to be presented in Chapter 4 and 5.

3.4.3 Impact of Short-term Performance on Long-term Performance

One intrinsic conflicts between operation and maintenance resides in the time they span. While operation is a daily or even hourly practice, maintenance is conducted at relatively sparse time intervals. As a consequence, decisions related to task/workload allocation are sometimes made on a rather ad hoc basis without realising how such decisions will affect future maintenance actions. Such practice can be observed in the case of Company A where the operator allocates workload based on how convenient back-washes are done on the vessels, as well as in Company B where conveyors are reconfigured to meet a sharp increase

in the demand of a certain type of product. While the consequence of occasional decision making based solely on short-term operation needs can be insignificant, it cannot be neglected in terms of cumulative effects in the long run or in situations where the asset is reaching its limiting condition for preventive maintenance [15]. In order to resolve the dilemma faced by decision makers, the model to be proposed must explicitly incorporate a mechanism to: 1) quantify the instant consequence of a short-term decision; 2) relate the long-term effect to this short-term decision; 3) attempt to refine such a decision as asset condition changes.

3.5 Chapter Summary

This chapter has provided a description of cases found in the literature related to integrated operation and maintenance decision making as well as findings from exploratory case studies, which together indicate that there is a need for a systematic strategy for joint decision making of task/workload allocation and condition-based maintenance. Specifically, it is first noted that the industrial settings most likely to benefit from such an integrated approach are those with parallel assets or flexible manufacturing cells. Furthermore, the studies indicate that some of the companies have noticed the dynamics between operation and maintenance and made preliminary attempts to exploit such relationship. Later in the discussion, the three main challenges faced by the industry have been revealed: 1) the need for a load-dependent degradation model; 2) a lack of understanding of how individual asset performance impacts system-level performance; 3) the dilemma between short-term and long-term objectives. These findings have been used to help identify, formulate, and resolve the research problem.

The next chapter (Chapter 4) will first discuss the constitutive components of such a decision-making model and move on to develop the first component - a load-dependent condition-based maintenance model for individual assets. The individual model will be further incorporated into the workload allocation strategy to be presented in Chapter 5.

Chapter 4

Condition-based Maintenance Optimisation for Individual Assets

4.1 Introduction

This chapter serves to answer the first and second research questions defined in Chapter 2 by: 1) identifying the constitutive components of a joint workload allocation and condition-based maintenance decision making model; and 2) formulating a mathematical model of workload-dependent deterioration and condition-based maintenance optimisation for individual assets. Specifically, the model consists of a description of the relationship between workload and deterioration, mechanisms that lead the asset to be out-of-service, and effects of various maintenance measures. As an important constitutive component of the integrated decision-making system, this model will be used to generate insights necessary for optimising workload allocation among an asset fleet - such as the relationship between workload and maintenance costs at the individual asset level.

The outline of this chapter is presented in Fig 4.1. The chapter is organised as follows: it starts with a discussion on what should be the constitutive components of the joint decision making strategy in Section 4.2. Then it moves to actually develop the individual asset condition-based maintenance model: first the essential traits of the individual asset model are identified in Section 4.3; existing maintenance models are evaluated in Section 4.4 against the identified elements, where emphasis is placed on model characteristics that have the potential to meet the specifications demanded by this research; then the model developed for this research project is presented in Section 4.5. An optimisation algorithm is proposed and mathematically verified in Section 4.6. To demonstrate the capability of the

model of capturing the impacts of various factors, some numerical examples are provided in Section 4.7, followed by a chapter summary given in Section 4.8.

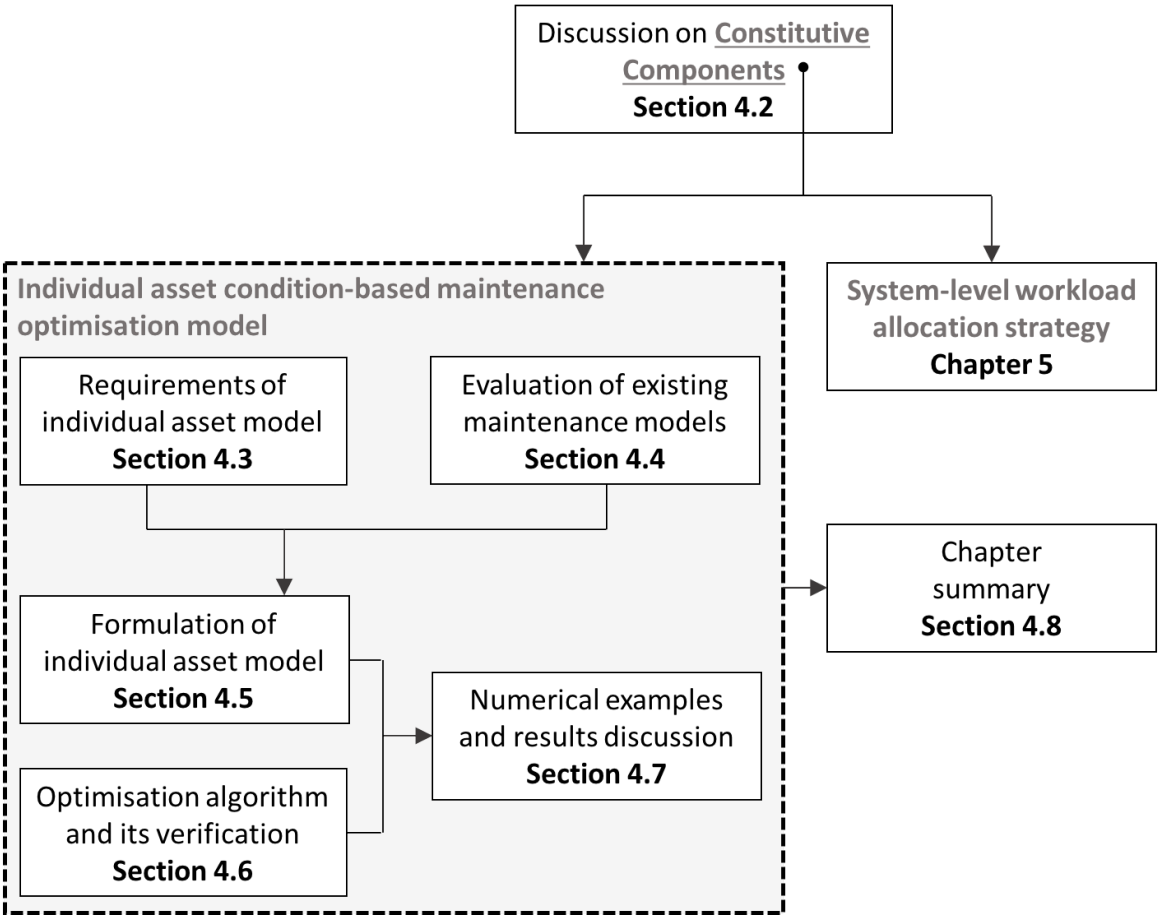


Fig. 4.1 The outline followed by the individual model chapter

4.2 Constitutive Components for Joint Load Allocation and Maintenance Strategy

The purpose of this section is to identify the constitutive components for an integrated task/workload allocation and maintenance decision making strategy. This is achieved by a joint consideration of the following:

1. the need to fill the research gaps identified in Chapter 2;
2. the need to address the challenges faced by the industry while attempting to apply such a decision making strategy, as presented in Chapter 3;

3. inspiration from existing approaches which have been designed to solve similar problems.

Before diving into the discussion, a brief description of the problem setting is given here so as to build a qualitative understanding of the context. Mathematical details will be left to later sections in this chapter as well as in Chapter 5. In plain language, the question we are trying to answer here is: knowing the demand for a certain product and the health condition of a fleet of machines capable of manufacturing such product, how should the workload be allocated amongst these machines and how should maintenance be planned. Namely, decisions need to be made on the amount of workload assigned to as well as the threshold for condition-based maintenance for each asset.

One of the findings from the review in Chapter 2 is that most of the approaches proposed in the literature are intended for a long-term static solution using probabilistic models. Such approaches thus will miss the potential benefits that could be realised by dynamically adjusting the workload assigned to a fleet of assets. Furthermore, they lack the flexibility of coming up with a timely and 'good enough' solution in face of volatile circumstances. It can also be noticed that for the existing models, the optimal solutions are obtained with a centralised algorithm in the sense that the computational tasks are performed and the value of decision variables is picked by a single entity. While it is still feasible to take this approach for active control of asset degradation by task/workload adjustment, it is restricted to cases with very few parallel machines or when there is no time constraint. Here we argue that it is more reasonable to adopt a distributed approach due to its demonstrated advantage of flexibility and high efficiency brought by parallel computing.

Distributed problem solving, according to Decker [28], as a subset of distributed artificial intelligence, deals with the interactions of groups of agents that attempt to solve problems cooperatively. Distributed artificial intelligence can be further classified as fine-grained (statement-level) and coarse-grained (task-level) based on the granularity of problem decomposition, where distributed problem solving belongs to the latter, making it well-suited for the purpose of this research. One of the motivations of constructing solutions in a distributed way is that for problems that can be divided into independent parts, a solution can be found more quickly as the parts can be solved in parallel [28]. One of the first steps in applying a such an approach to solving the proposed integrated workload allocation and CBM threshold optimisation problem is to determine the level of distributed control in the solution-seeking process. Specifically, the following three dimensions are considered [28]:

1. **Cooperation:** fully cooperative systems often suffer from high communication costs and are especially useful when there is no clear *a prior* what each node should be doing or what information is to be exchanged [9]. This is, however, not the type of problem

this thesis attempts to address. As the objective of optimising workload allocation and CBM threshold is rather clear, cooperation of modest to low level will be sufficient.

2. **Coherence and Organisation:** the possible organisations of distributed problem solving range from totally free teams to master/slave relationships. Totally free teams consider a problem solved when one team member comes up with an answer, while in hierarchical organisations some nodes are responsible for the construction of intermediate solutions (as in Upasani et al. [92]) or providing global directions (as in Giordani et al. [36]). Taking into account that one of the challenges faced by the industry, as presented in Chapter 3 concerns the impact of individual asset performance on system-level performance, a supervisory node that has relatively comprehensive knowledge of the system is needed.

3. **Dynamics:** given a cooperating system, the organisation can either be statically specified when the system is being designed or dynamically created at run time. Given the clear objective and relatively simple formulation of the problem this thesis aims to address, the static organisation will be adopted.

Since the form of the decision-making strategy has been decided, the task that follows immediately is to identify the constitutive components (sub-problems) of a joint workload allocation and maintenance strategy. Machines require maintenance when their health status falls below a certain threshold. However, if maintenance is conducted whenever such requests are raised, production is very likely to be interrupted as a single machine has limited knowledge of the system. It is therefore important that an entity is present that has a comprehensive understanding of the system as a whole. From the discussion above, the constitutive components are found to be the following:

1. a condition-based maintenance optimisation model for individual assets;

2. a system-level task/workload allocation model that utilises data obtained from the individual models.

Multi-agent system (MAS) is a commonly adopted approach of implementing distributed decision making, and these two constitutive components will be hosted in a multi-agent system (MAS). This thesis has designed a two-layer structure. To be specific, the first layer consists of machine agents that host the first component and analyse the condition of individual assets, whereas the second layer hosts a coordinator agent that collects necessary information from machine agents to establish a comprehensive understanding of fleet performance and hosts the second component. The responsibility of setting the right amount of workload for individual machines naturally falls on the shoulders of the coordinator agent.

However, for the purpose of this research, we give the term 'agent' a different meaning from its traditional and classical definition. According to [97], the definition of an agent is given as follows: an agent is a computer system that is situated in some environment, and that is capable of autonomous action in this environment in order to meet its design objectives. Further to that, it is emphasised that the action of an agent is directed towards achieving its own goals and that there is often a repertoire of actions for the agent to choose from. An agent should be able to reach agreement through negotiation on matters of common interests and dynamically coordinate its activities with agents whose goals and motives are unknown. From the previous discussion we can see that the level of distributed control in the proposed solution-seeking process is at the lower end in terms of communication intensity and dynamic relationship-forming. Though the 'agents' involved in the problem-solving process has certain level of awareness of its environment, the interaction between them are rather straight-forward and there is no complex reasoning in terms of what actions to take in the next step. Therefore, the following definition is adopted in this thesis for the purpose of avoiding confusion. We would also like to clarify that this definition is proposed specifically for this thesis and for the purpose of ease of description of the proposed model:

Definition 4.2.1. Agent is a specialised computational entity that is tasked with solving a sub-problem of a system-level problem and has basic communication capabilities.

The multi-agent structure to be used in the proposed decision-making model is presented in Figure 4.2. Specifically, each machine is assigned a machine agent that monitors its health condition. These machine agents are also responsible for generating and sending relevant information to a coordinator agent who has more visibility over the entire system. Workload is then allocated accordingly by the coordinator agent to the units.

The rest of the chapter deals with the formulation of the first constitutive component. A detailed description of how these two parts come together as well as the steps involved in the decision-making process is provided in Chapter 5. The development of the second component is also left to Chapter 5.

4.3 Requirements and Rationale for Individual Asset Model

The role of the individual asset maintenance model within the entire decision-making algorithm is to provide information regarding the impact of load allocation on various aspects of asset performance, such as degradation behaviour, expected maintenance cost, expected time to maintenance, etc., which assists the coordinator in arriving at an integrated system-level optimal decision for both production and maintenance.

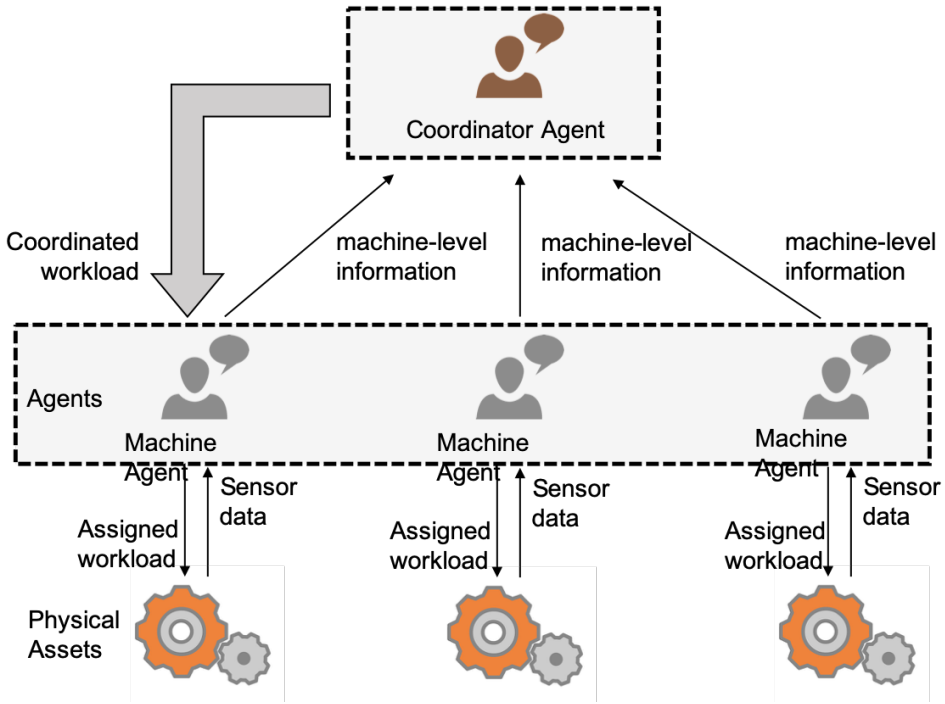


Fig. 4.2 The multi-agent structure to be used in the proposed decision-making model

The requirements for such an individual asset maintenance model are thus twofold: 1) as an individual asset model, it is expected to capture the load-dependent deterioration of assets and come up with an optimised maintenance strategy accordingly; 2) as part of the decision-making system, it will need to satisfy the requirements imposed by the system-level objective and dynamic nature of the decision-making process. These requirements will be discussed in details in the rest of the section.

4.3.1 Requirements as an Individual Asset Model

According to Sharma et al. [84], a typical maintenance optimisation model consists of the following elements:

1. a description of a technical system and its major functions;
2. a modelling of the way the system deteriorates in time and the possible consequences;
3. a description of the available information regarding the system status;
4. the action pool open to the management
5. objective functions and optimisation techniques to help to find the best solution.

Following the general requirements, here we have identified certain traits the individual asset condition-based maintenance model needs to possess for the purpose of this research project. Specifically, it should be able to:

1. *Model the load-dependent degradation behaviour of the asset*
   A fundamental component of the model is the description of the load-dependent behaviour of the asset. It will be used to quantify and predict the change of asset status under different load conditions for a given time duration. The model should also be able to give the degree of uncertainty associated with such quantification. In reality, the level of workload assigned to an asset can be either continuous (such as the volume of water flowing through a filter) or discrete (such as the number of products to produce). In order for the model to be more generic, continuous value of load ratio (which is defined as the assigned workload divided by the maximum machine capacity) will be considered in this study.

2. *Model the impacts of maintenance and lack of maintenance*
   Maintenance, in its nature, can be seen as an action that improves the asset condition by consuming at least one of the following: time and money. Therefore, the model should incorporate the rectification effects of different types of maintenance, on both the current degradation status and future status evolution of the asset. The corresponding costs of various maintenance actions, either in the form of time or monetary terms, also need to be built into the model. The impact of lack of maintenance can be included in the model explicitly as the asset deterioration being non-decreasing without maintenance, or implicitly accounted for by an increased risk of failure.

4.3.2 Requirements as Part of the System

As part of the system-level optimisation, the asset maintenance optimisation model needs to act as a care-taker of the asset itself as well as an information provider to assist in the process of joint workload allocation and maintenance decision-making. Specifically, the following requirements should be satisfied:

1. *Providing decision-useful information on workload-maintenance relationship*
   The model is required to provide information sufficient to quantify the relationship between production and maintenance using metrics that are of interest to decision makers. Some potential metrics include expected average maintenance cost, expected unplanned loss of production, and expected time to the next preventive maintenance. These pieces of information constitute an important part of the input for the coordinator agent to finalise a joint plan of workload allocation and condition-based maintenance.

2. *Updating maintenance decisions for the remaining life-cycle*
   The model should be able to update its maintenance decisions dynamically based on its current status and the given demand for production. The reason for imposing this requirement is that as demands and load allocation vary in time, the initial maintenance decision optimised at the onset of the asset life-cycle may very likely be rendered non-optimal later. The individual model should thus be able to refine its maintenance decisions for the remaining life-cycle in order to optimise its full life-cycle performance. Namely, whenever a decision is to be made, the model must be able to update its previous optimal solutions based on its current status and expectation for the future.

4.4 Evaluation of Existing Maintenance Models

This chapter begins by reviewing and evaluating some existing maintenance models, which aims to lay the foundation and provide inspiration for the optimisation model proposed in this thesis.

From the literature review in Chapter 2 we can see that the deterioration of assets in CBM models is often described by a stochastic process. In general, there are two ways to incorporate the load-dependent behaviour into the degradation model: 1) for discrete stochastic models, the effect of workload is reflected in the scale of transitive probability between states, as is the case in AlDurgam and Duffuaa [3]; 2) for continuous models it is accounted for by the different choice of distribution parameters, as the approach adopted in Hao et al. [42]. Since a more generic model is preferred in this research project as mentioned in the previous chapter, the latter path will be taken.

There can be both direct and indirect consequences associated with lack of maintenance. The direct consequence is often incorporated in the model as machine breakdowns due to degradation reaching a failure limit, whereas the indirect consequences can take the form of lower production efficiency, increased vulnerability to shocks [62], or higher probability of producing faulty parts [3]. Maintenance actions can be described as imperfect [46, 61], or perfect [23], depending on their degree of rectification effect on asset condition. Since most often than not, the consequence of an asset being over-degraded is rather complex and that uncertainty exists regarding the actual effect of maintenance practice, this research will consider both direct and indirect implications of asset degradation and treat the post-maintenance asset status as a random variable.

Note that most existing load-dependent maintenance models are able to meet the requirements as an individual asset model with insubstantial modifications. It is not the case, however, when it comes to the requirements imposed by the system-level optimisation strat-

egy. This is mainly caused by the inconsistency between the proposed decision-making strategy being dynamic in nature and these maintenance models seeking an optimal solution for the long-run steady state [34, 19]. As objective functions designed for an infinite time horizon are not suited for the purpose of this research, it is therefore necessary to come up with a new objective function that can be used by individual assets to generate updated solutions in accordance with their ever-changing internal health condition as well as fluctuating external environment. One instance of maintenance models that choose not to optimise long-run performance indicators can be found in Liao et al. [61], where the objective function is set to be the achieved availability of one replacement cycle. Inspired by their work, we will adopt a similar approach in this study and propose an individual asset model that aims to optimise a cost-based objective function for one life cycle.

4.5 Individual Asset Maintenance Optimisation model

This section starts with the general assumptions of the condition-based maintenance optimisation model, after which a detailed formulation of the model components is presented. The maintenance model proposed in this section is based on the framework developed by Liao et al. [61]. Modifications and extensions have been made where appropriate to meet the specific requirements of this research. Details of these modifications and extensions are given later in the rest of the section. An algorithmic approach to solve the optimisation problem is proposed and verified, followed by numerical examples, analysis, and discussions of modelling results.

4.5.1 General Assumptions

General assumptions underlying the formulation of the individual asset model are as follows:

1. Though there is often upfront investment associated with attaching sensors and adding data communication capabilities to assets, a detailed cost analysis is out of the scope of this research. It is assumed that sensors on the machines provide accurate information regarding their condition with no additional costs. Hence, the degradation level of asset is known at all times. Under this assumption, preventive maintenance will always be performed before disastrous machine breakdowns and random failures can be detected instantly.

2. The degradation of assets is load-dependent. Specifically, the rate of degradation is positively correlated with the workload allocated to the asset. When asset degradation

reaches a pre-defined threshold, imperfect preventive maintenance needs to be carried out to bring the asset back to a better state.

3. Random shocks can happen at any time, which will lead to asset breakdown, such as an operator error or a power surge. As assets become more vulnerable to shocks with accumulated damage, the rate of shocks is assumed to be positively correlated with the degradation level of assets. When an asset experiences a shock failure, it will undergo minimal repairs, which bring it back to the 'as-bad-as-old' status.

4. Replacement of assets is triggered after a certain number of imperfect preventive maintenance has been performed, restoring the asset to 'as-good-as-new' status.

5. At the asset level, the time taken to carry out minimal repairs, imperfect preventive maintenance, and replacement is all assumed to be negligible. The rationale for this assumption is that for a single asset, the duration of maintenance is rather insignificant compared with its entire life cycle. Furthermore, the time spent on maintenance is often associated with production loss. For this research study, loss of production is defined as the penalty cost resulted from unmet demand. However, as the subject of this study is a fleet of parallel assets working in a coordinated way to meet a certain demand, it would not be possible for any individual asset to know how much production losses are incurred. It is thus neither sensible nor necessary to consider loss of production for a single asset. Production losses will be accounted for by the second constitutive component introduced in Chapter 5.

4.5.2 State of an Asset

The state of an asset is fully described by three variables: 1) its health condition x_0, which represents the level of degradation of the asset at the current point in time; 2) the time that has already elapsed in its life cycle, denoted as t_0; 3) the number of preventive maintenance that has been completed on the asset, N_0. The value of all three variables will be reset when the asset life cycle is renewed.

4.5.3 Modelling Load-dependent Degradation

An asset is subject to an inevitable degradation process as long as it is in operation. Here, a single continuous-state stochastic process, $X(t)$, is used to describe the degradation state of assets, where $X(t) = 0$ denotes an asset that is in perfect condition. Multiple stochastic processes exist, such as Wiener process, Brownian motion process, and Gamma process,

that are applicable to the modelling of continuous degradation behaviours. In view of the irreversible nature of the degradation mechanisms considered in this thesis, such as corrosion, cracks, erosion, etc., Gamma process has been generally preferred to model monotonic process, as compared with Wiener or Brownian motion process whose increments could be both positive and negative [94]. Moreover, Gamma process is rather flexible and simple in mathematical terms. Thus, following the choice of Liao et al. [61], Gamma process is adopted here and extended to incorporate the workload-dependent degradation behaviour of assets to form the basis of this model.

Suppose that the degradation increment of an asset from time t to $t+\Delta t$, denoted as $\Delta X(t) = X(t+\Delta t) - X(t)$ follows a Gamma process with shape parameter $\kappa = \kappa_0 \Delta t$, where κ_0 is the shape parameter that corresponds to the unit-time Gamma distribution, and scale parameter θ_t. In view of the finding in Chapter 2 that for a continuous process, the impact of workload on asset degradation process can be reflected in the values of distribution parameters, here we set θ_t to be a function of the workload u_t in the proposed model as follows,

$$\Delta X(t) = X(t+\Delta t) - X(t) \sim g(x;\, \kappa_0 \Delta t, \theta_t) \tag{4.1}$$

$$\theta_t = \theta_0 \left(\exp(r_t)\right)^s \tag{4.2}$$

$$r_t = \frac{u_t}{W} \tag{4.3}$$

where $g(\cdot)$ represents the PDF of Gamma distribution, W is the capacity of the asset, r_t is defined as the load ratio at time t, θ_0 denotes the value taken by the scale parameter when no workload is assigned, s is the load sensitivity coefficient, which describes the sensitivity of asset degradation behaviour to the load ratio. For instance, if the maximum volume of water that a filter can process in one hour is 1000 litres, and it is assigned 500 litres of water for the next hour, then $W = 1000$, $u_t = 500$, and $r_t = 500/1000 = 0.5$. It can be noticed that $s = 0$ simply corresponds to a load-independent degradation process, and that a larger s implies a more significant influence of workload on degradation. As shown in Figure 4.3, as s increases from 0.5 to 2, the rate of degradation as the load ratio increases becomes increasingly significant. The sensitivity of asset degradation to the load ratio can be affected by various factors, such as the nature of the task and the age of the asset. Note that when no workload is assigned to the asset, θ_t simply takes the value of θ_0, leading to a positive mean

degradation rate $\kappa_0\theta_0$. This property of the model captures the ageing behaviour of materials that deteriorate even if not in use, such as the hardening of natural rubber in storage [35].

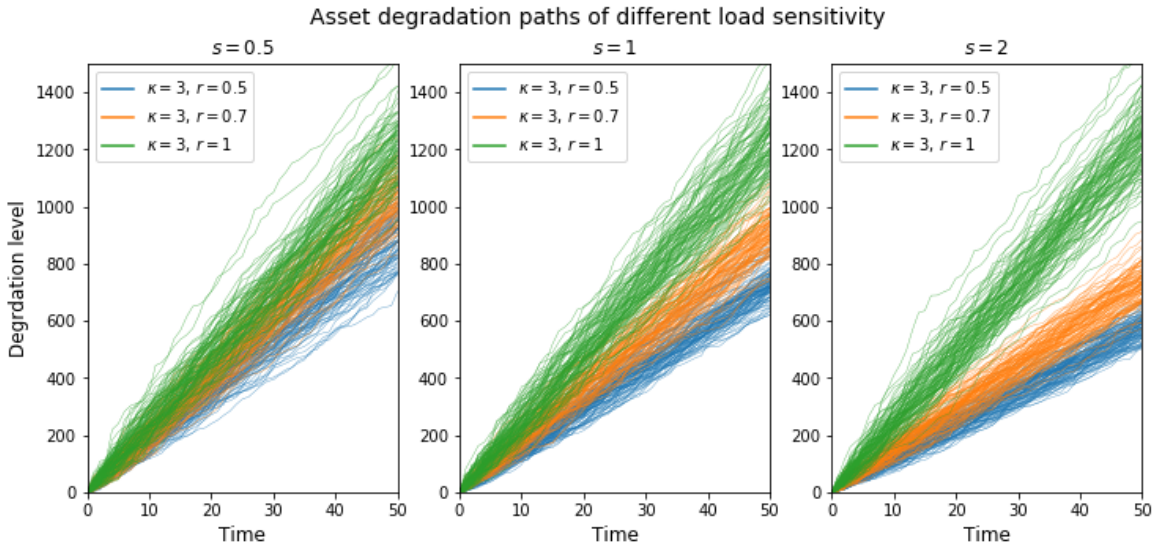


Fig. 4.3 Comparison of asset degradation paths with different load sensitivity

Under the current modelling assumptions, the variance of Gamma distribution, $\kappa\theta^2$, also increases with the assigned workload, as can be seen in Figure 4.4. This is consistent with experiment findings that asset increment degradation tends to be more volatile and less predictable under higher loads, as demonstrated by Liao and Elsayed [60].

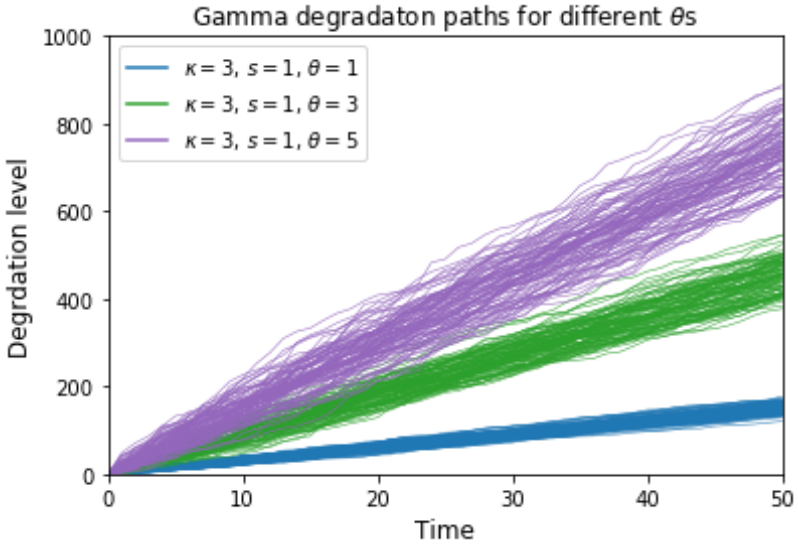


Fig. 4.4 Comparison of Gamma degradation paths with different scale parameters

4.5.4 Degradation-dependent Random Shocks

Apart from continuous internal degradation process, assets are also subject to external random shocks that cause undesired consequences, be it abrupt jumps in the degradation level, complete asset failures, or temporary breakdowns. Recall that one of the requirements identified in Section 4.3 for the individual model is that it needs to model the impact of lack of maintenance, either directly by having the degradation level monotonically increase without maintenance (as adopted by Liao et al. [61]), or indirectly by exacerbated vulnerability to breakdowns. Here we further assume here that shocks might lead to machine breakdowns which can be fixed with minimal repairs. As assets grow more vulnerable to sudden changes in the working environment with increasing age as well as accumulated usage history, the actual condition of the asset, $X(t)$, is a more comprehensive measure of asset sensitivity to shocks. Following the approach adopted by Jaturonnatee et al. [54], it is further assumed that shocks that will lead to machine breakdown occur following a non-homogeneous Poisson process with intensity $\lambda(t)$. To take into account the actual condition of the asset, $X(t)$ is incorporated into the expression $\lambda(t)$ so that it captures more than the age factor,

$$\lambda(t) = \lambda_0 \exp(\lambda_1 \frac{X(t)}{F}) \tag{4.4}$$

where F represents the failure threshold of asset degradation. In other words, an asset is considered completely failed and has to be replaced if its degradation level reaches F. λ_0 represents the rate of shock failures happening to an asset in perfect condition ($X(t) = 0$), and λ_1 represents the significance of influence of asset degradation on vulnerability to shocks. Note that when $\lambda_1 = 0$, the arrival rate of random shocks becomes a constant λ_0, which corresponds to the extreme case scenario where the sensitivity of assets to external shocks is independent of the degree of deterioration.

4.5.5 Maintenance and Replacement

As mentioned in Section 4.4, the model proposed in this chapter will consider various forms of maintenance practice: imperfect preventive maintenance, minimal repairs, and replacement.

Imperfect Preventive Maintenance

Ideally, maintenance actions are expected to restore an asset to its 'as-good-as-new' state, as is the case addressed by Jain et al. [50], Cui et al. [25], Liu et al. [62]. However, such perfect rectification effect can hardly be achieved in reality, attributable to either the nature of the

degradation mechanism such as corrosion, rusting, and wearing, or certain undesirable side effects of the maintenance action itself. An example of the latter is patch repairs of internal linings of large vessels in a petrochemical plant. This type of repair requires the set-up and tear-down of scaffolds, which tends to damage the linings and result in areas of accelerated corrosion.

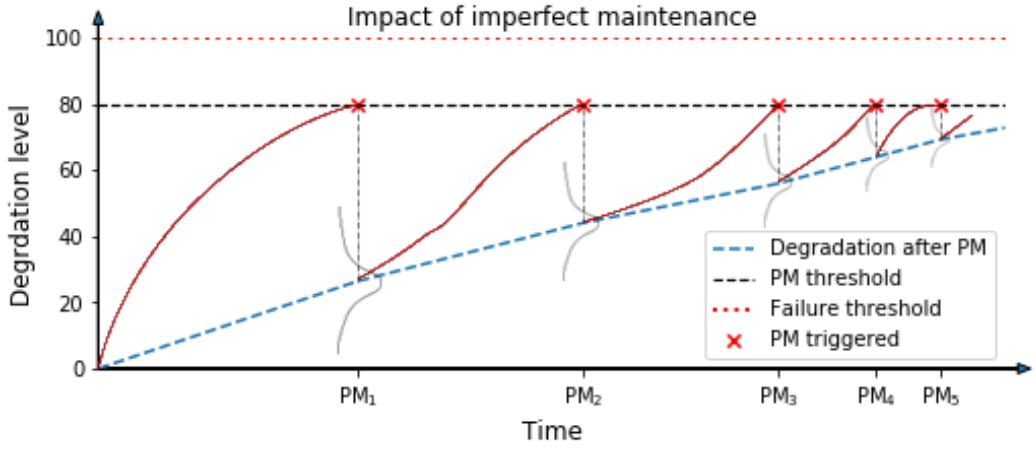


Fig. 4.5 Impact of imperfect maintenance on asset degradation status

The effectiveness of imperfect preventive maintenance can take various forms in different models. Following the approach taken in Liao et al. [61], it is assumed for this research that the degradation of the asset immediately after the i^{th} preventive maintenance, denoted as $X(R_i^+)$, takes a random value within the interval $[0, X(R_i^-)]$, where $X(R_i^-)$ is the degradation of the asset at the instance when the i^{th} preventive maintenance is triggered. In view of the ageing mechanism and possible accumulated damage caused by maintenance actions, it is reasonable to have the rectification effect of preventive maintenance gradually diminish as i gets larger. Assuming the level of degradation ranges from 0 to $F = 100$, the impact of imperfect preventive maintenance is illustrated in Figure 4.5. Specifically, the variance and mean value of $\frac{X(R_i^+)}{X(R_i^-)}$ are given as follows,

$$\mathrm{E}[\frac{X(R_i^+)}{X(R_i^-)}] = \mu_i \tag{4.5}$$

$$\mathrm{Var}[\frac{X(R_i^+)}{X(R_i^-)}] = \sigma_i^2 \tag{4.6}$$

where μ_i and σ_i are the mean value and variance of the degradation distribution after the i^{th} preventive maintenance, which are independent of $X(R_i^-)$. It can be noticed that perfect

maintenance is a special case of the proposed model when $\mu_i = 0$ and $\sigma_i^2 = 0$, and imperfect maintenance with deterministic improvement effect can be depicted by setting $\mu_i > 0$ and $\sigma_i^2 = 0$. Variables that follow a Beta distribution are naturally bound by 0 and 1, and thus Beta distribution is employed to describe $\frac{X(R_i^+)}{X(R_i^-)}$. A series of PDFs, $f_{X(R_i^+)}(x), i \in \{\mathbb{N} : i \geq 1\}$ that satisfy the above assumptions, as proposed by Liao et al. [61], are presented as follows,

$$f_{X(R_i^+)}(x) = \frac{1}{X(R_i^-)} \frac{\Gamma(\alpha_i + \beta i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} (\frac{x}{X(R_i^-)})^{\alpha_i - 1} (1 - \frac{x}{X(R_i^-)})^{\beta_i - 1} \tag{4.7}$$

where $\Gamma(\cdot)$ is the Gamma function given by $\Gamma(y) = \int_0^\infty \tau^{y-1} e^{-\tau} d\tau$, and the distribution parameters $\alpha_i > 0$ and $\beta i > 0$ are assumed to take the following values,

$$\alpha_i = \alpha_0 = const.$$

$$\beta_i = \beta_0 \exp(-ib) \tag{4.8}$$

Thus, the mean and variance of $\frac{X(R_i^+)}{X(R_i^-)}$ can be obtained by

$$\mu_i = \mathrm{E}[\frac{X(R_i^+)}{X(R_i^-)}] = \frac{\alpha_i}{\alpha_i + \beta_i} = \frac{\alpha_0}{\alpha_0 + \beta_0 \exp(-ib)} \tag{4.9}$$

$$\begin{aligned} \sigma_i^2 = \mathrm{Var}[\frac{X(R_i^+)}{X(R_i^-)}] &= \frac{\alpha_i \beta_i}{(\alpha_i + \beta_i)^2 (\alpha_i + \beta_i + 1)} \\ &= \frac{\alpha_0 \beta_0 \exp(-ib)}{(\alpha_0 + \beta_0 \exp(-ib))^2 (\alpha_0 + \beta_0 \exp(-ib) + 1)} \end{aligned} \tag{4.10}$$

An illustration of how the distribution of asset degradation status immediately after preventive maintenance changes as the number of preventive maintenance tasks increases is given in Fig 4.6. The PDF of the distribution exhibits a noticeable shift to the right with a larger i.

Minimal Repairs

As proper maintenance is more time consuming and requires more preparation, minimal repairs are carried out when random shocks lead to machine breakdowns in the middle of the production process. Since minimal repairs are designed to deal with emergency, it is often the case that they are assumed to only bring the asset back to the 'as-bad-as-old' state. Namely, the asset will be operational again after minimal repairs, but its health condition will not be improved. It is also assumed that shocks and minimal repairs have no influence on the shock

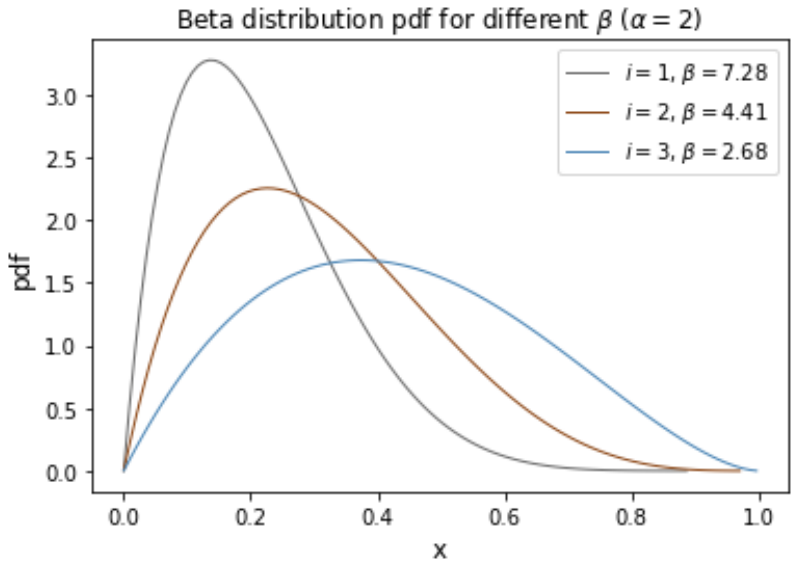


Fig. 4.6 Distribution of asset degradation after PMs

intensity. In mathematical terms, after experiencing a shock followed by a minimal repair, $X(S_i^+) = X(S_i^-)$, and $\lambda(S_i^+) = \lambda(S_i^-)$, where $X(S_i^-)$, $X(S_i^+)$, $\lambda(S_i^-)$, and $\lambda(S_i^+)$ represent the degradation level and shock intensity of the asset before and after the minimal repair, respectively.

Replacement

An asset will be replaced after the completion of a certain number of preventive maintenance. If an asset is left to degrade without preventive maintenance, it will eventually reach a degradation level of F and fails and a replacement needs to be carried out. A replacement restores the asset to the 'as-good-as-new' state with no accumulated degradation.

Costs of Maintenance and Replacement

As loss of production is not considered at the individual asset level, the costs resulted from maintenance and replacement include only those associated with the resources needed for such tasks, such as spare parts, maintenance tools, and labour. The costs for a single preventive maintenance, minimal repair, and replacement task are all taken to be constant, and denoted as c_{pm}, c_{mr}, and c_{rp}, respectively. Replacement is done with an overhaul of the production system consuming the most maintenance resource and thus has the highest cost among the three, whereas minimal repairs are often associated with much less costs. There-

fore the following relationship holds for the costs of replacement, preventive maintenance, and minimal repairs: $c_{rp} > c_{pm} > c_{mr}$.

4.5.6 Objective Function

Most maintenance model objectives centre around one of the following classes: 1) cost metrics, such as long-run average cost, life-cycle cost, and long-run average profit; 2) operational metrics, such as reliability and availability; 3) value metrics which considers the value-adding effect of maintenance activities; 4) more comprehensive metrics, such as Overall Equipment Effectiveness (OEE) and Overall System Effectiveness (OSE). In the formulation of the proposed model, the cost-based approach has been taken to design the objective function. The most widely accepted cost-based objective for maintenance decision making is long-run average cost, whose calculation is largely based on renewal theory [6]. However, objective functions that aim to optimise performance metrics calculated for an infinite time horizon are not suited for this study for two reasons. First, renewal theory assumes that at the end of a full cycle, the asset is replaced with one with the same failure time distribution, whereas this is hardly the case where asset version update is frequent or where a full cycle runs several decades and the original material is no longer available. Furthermore, the strategy proposed in this study requires maintenance and workload allocation decisions to be updated at each time epoch. As long-run strategies are not designed to react dynamically to changes in asset condition and production environment, here we develop an objective function to minimise the expected remaining maintenance costs averaged over the expected life cycle of the asset, conditioned on its actual state at the moment. The rationale behind using the remaining maintenance costs instead of the total maintenance costs is that, costs already incurred by this moment are sunk costs, and should not be considered while making decisions for the future.

A few clarifications need to be made before we proceed with the formal description of the optimisation problem. Recall that one of the requirements identified in Section 4.3 imposed on the individual maintenance model as part of the decision-making system is that it needs to be able to update its decisions based on the current state of the asset and expectation for the future. The original objective function as proposed by Liao et al. [61], which characterises the achieved availability of the asset over its life-cycle, is not suited for this purpose. A refined objective function is then proposed in this section to meet such specific requirement. The objective function to be presented in this section concerns two important concepts - life cycle of an asset, and the state of an asset, $[x_0, t_0, N_0]$ as defined in Section 4.5.2. Here, we define the life cycle of an asset to be the period of time between the initiation of the asset and the next replacement. Note that the initiation of the asset does not have to be a

replacement. Since maintenance strategies do not always start to be implemented on entirely new machines, the relaxation of this constraint makes the proposed model a better reflection of the reality.

The objective is to find a pair of values: the preventive maintenance threshold H, and the number of preventive maintenance tasks N after which a replacement is carried out, that minimise the expected remaining maintenance cost averaged over the expected life cycle of the asset, $Q(H, N\,|x_0, t_0, N_0)$, given the actual state of the asset. Note that PM is triggered if the degradation of an asset exceeds H, so the actual degradation $X(R_i^-)$, is not necessarily the same as H. In mathematical terms, we need:

$$[H^*, N^*] = \arg\min\{Q(H, N) : x_0 < H \leq F, N \geq N_0 | x_0, t_0, N_0\} \tag{4.11}$$

The mathematical expression for $Q(H, N\,|x_0, t_0, N_0)$ is given by

$$Q(H, N\,|x_0, t_0, N_0) = \frac{\mathrm{E}(C_{mr,t_0,N_0+1}|x_0, t_0, N_0) + \sum_{N_0+1}^{N} \mathrm{E}(C_{mr,i}) + c_{rp} + c_{pm}(N - N_0)}{t_0 + \mathrm{E}(T_{t_0,N_0+1}|x_0, t_0, N_0) + \sum_{N_0+1}^{N} \mathrm{E}(T_i)} \tag{4.12}$$

where the numerator denotes the expected maintenance costs to be incurred over the remaining life cycle of the asset and the denominator denotes the expected length of the asset life cycle, both conditioned on the current state of the asset.

Specifically, in the numerator, C_{mr,t_0,N_0+1} is the minimal repair costs incurred between t_0 and the $(N_0+1)^{th}$ preventive maintenance, $C_{mr,i}$ is the minimal repair costs incurred between the i^{th} and $(i+1)^{th}$ preventive maintenance, c_{rp} is the replacement cost, and $c_{pm}(N - N_0)$ is the cost of the remaining preventive maintenance tasks before the replacement. The expected number of minimal repairs carried out between t_0 and the $(N_0+1)^{th}$ preventive maintenance, denoted as $\mathrm{E}(n_{mr,t_0,N_0+1}|x_0, t_0, N_0)$, is calculated using the following expression:

$$\mathrm{E}(n_{mr,t_0,N_0+1}|x_0, t_0, N_0) = \int_0^{\infty} G(H - x_0;\, \kappa_0 t,\, \theta_0 (\exp(r_{t_0}))^s) \lambda(t) dt \tag{4.13}$$

Then the expected value of C_{mr,t_0,N_0+1} is given by multiplying equation 4.13 with c_{mr}:

$$\mathrm{E}(C_{mr,t_0,N_0+1}|x_0, t_0, N_0) = \int_0^{\infty} c_{mr} G(H - x_0;\, \kappa_0 t,\, \theta_0 (\exp(r_{t_0}))^s) \lambda(t) dt \tag{4.14}$$

where $G(\cdot)$ represents the CDF of the Gamma distribution. The calculation of the expected minimal repair costs for time periods other than between t_0 and the $(N_0+1)^{th}$ preventive maintenance is slightly different as it requires one more factor to be taken into account - the

distribution of the degradation level of the asset after imperfect maintenance. The expression for $\mathrm{E}(C_{mr,i})$ can be easily deduced using equations 4.14 and 4.7, which is written as

$$\mathrm{E}(C_{mr,i}) = \int_0^H \int_0^\infty c_{mr} G(H-x;\, \kappa_0 t, \theta_0 (\exp(r_{t_0}))^s) f_{X(R_{i-1}^+)}(x) \lambda(x,t) dt dx \tag{4.15}$$

In the denominator, the first item t_0 is simply the operational age of the asst, T_{t_0,N_0+1} denotes the first operating time period of the asset from t_0 and before the $(N_0+1)^{th}$ maintenance, and $T_i, i \in [N_0+1, N]$ denotes the operating time between the i^{th} and $(i+1)^{th}$ preventive maintenance, or the replacement in the case where $i = N$. It follows that the expectation of T_{t_0,N_0+1}, can be obtained via the following equation:

$$\mathrm{E}(T_{t_0,N_0+1} | x_0, t_0, N_0) = \int_0^\infty G(H-x_0;\, \kappa_0 t,\, \theta_0 (\exp(r_{t_0}))^s) dt \tag{4.16}$$

The difference in mathematical expression between T_i and T_{t_0,N_0+1} is very similar to that between C_{mr,t_0,N_0+1} and $C_{mr,i}$, and the expression for $\mathrm{E}(T_i)$ can be deduced using equations 4.16 and 4.7, which is given as follows,

$$\mathrm{E}(T_i) = \int_0^H \int_0^\infty G(H-x;\, \kappa_0 t, \theta_0 (\exp(r_{t_0}))^s) f_{X(R_{i-1}^+)}(x) dt dx \tag{4.17}$$

As $X(t)$ is a random variable, attempting to deduce a formula using the distribution of $X(t)$ brings substantial computational difficulty to this problem, here the expected value of $X(t)$, calculated as

$$\overline{X}(t) = x_0 + \kappa_0 \theta_0 (\exp(r_{t_0}))^s t \tag{4.18}$$

is used as a point estimate for $X(t)$ in the calculation. Consequently, the estimate for $\lambda(t)$ is calculated as

$$\overline{\lambda}(t) = \lambda_0 \exp\big(\lambda_1 \frac{x_0 + \kappa_0 \theta_0 (\exp(r_{t_0}))^s t}{F}\big) \tag{4.19}$$

Thus, equation 4.14 is rewritten as,

$$\begin{aligned} \mathrm{E}(C_{mr,t_0,N_0+1} | x_0, t_0, N_0) = \int_0^\infty & c_{mr} G(H-x_0;\, \kappa_0 t, \theta_0 (\exp(r_{t_0}))^s) \\ & \lambda_0 \exp\big(\lambda_1 \frac{x_0 + \kappa_0 \theta_0 (\exp(r_{t_0}))^s t}{F}\big) dt \end{aligned} \tag{4.20}$$

Following the same logic where a point estimate for $X(t)$ is used while calculating the value of $\mathrm{E}(C_{mr,N_0+1}|x_0,t_0,N_0)$, equation 4.15 is rewritten as:

$$\mathrm{E}(C_{mr,i}) = \int_0^H \int_0^\infty c_{mr} G(H-x;\, \kappa_0 t, \theta_0 (\exp(r_{t_0}))^s) f_{X(R_{i-1}^+)}(x) \\ \lambda_0 \exp\big(\lambda_1 \frac{x+\kappa_0 \theta_0 (\exp(r_{t_0}))^s t}{F}\big) dt dx \tag{4.21}$$

4.6 Optimisation Algorithm

In this section, we develop an optimisation algorithm that will be employed to obtain the pair of values $[H,N]$ that will minimise equation 4.12. First of all, the steps involved in the optimisation algorithm are presented. It is then proved that the proposed approach guarantees the solution obtained is the optimal pair $[H^*,N^*]$, where $H^* \in \{H \in \mathbb{R} : x_0 \leq H \leq F\}$, and $N^* \in \{n \in \mathbb{N} : n \geq N_0\}$.

4.6.1 Optimisation Algorithm

The algorithm used by the individual asset maintenance optimisation model consists of the following four steps:

1. Start with $N = N_0$, find the optimal H_N^* that minimises the objective function using gradient descent, and record the tuple $(N, H_N^*, Q_N^* = Q(H_N^*, N|x_0, t_0, N_0))$.

2. Increase N by 1, $N' = N+1$, and find the optimal $H_{N'}^*$ for this new N' following the same method in step 1, and record the tuple $(N', H_{N'}^*, Q_{N'}^*)$.

3. Compare Q_N^* and $Q_{N'}^*$. If $Q_N^* < Q_{N'}^*$, $[H_N^*, N]$ is the optimal solution pair. Otherwise go back to step 2.

4. Repeat step 2 and 3 until the optimal solution pair is found.

4.6.2 Validation of the Optimisation Algorithm

In order to ensure that the proposed algorithm leads to the global optimal solution, the following statements need to hold:

- **Statement 1:** For a given N, $Q(H,N|x_0,t_0,N_0)$ is a convex function w.r.t $H, H \in [x_0, F]$. This guarantees the uniqueness and optimality of the intermediate results obtained in step 1 and 2;

- **Statement 2:** For a given $H, H \in [x, F]$, there exists a unique finite $n, n \in \{\mathbb{N} : n < +\infty\}$, s.t.

$$Q(H, k|x_0, t_0, N_0) > Q(H, k+1|x_0, t_0, N_0), k < n;$$
$$Q(H, k|x_0, t_0, N_0) \leq Q(H, k+1|x_0, t_0, N_0), k \geq n$$

It can be tested with numerical examples that for a wide range of parameters, Statement 1 will hold. As this simply requires brute-force technique, the proof of Statement 1 is skipped. Here we will focus on the proof of the Statement 2. We begin the proof by providing the following Lemma.

Lemma 1. *Given non-negative fractions* $\frac{A}{B} \geq 0$ *and* $\frac{C}{D} \geq 0$, $\frac{A+C}{B+D} \geq \frac{A}{B} \iff \frac{C}{D} \geq \frac{A}{B}$.

Proof. With some simple manipulation, it can be shown that,

$$\frac{A+C}{B+D} = \frac{A+\frac{A}{B}D+C-\frac{A}{B}D}{B+D} = \frac{A}{B} + \frac{D}{B+D}(\frac{C}{D} - \frac{A}{B})$$

□

Corollary 1.1. *Given a non-negative fraction* $\frac{A}{B} \geq 0$*, and a non-negative monotonically increasing sequence* $(\frac{C_i}{D_i})_{i\in\mathbb{N}} \geq 0$, $\frac{C_0}{D_0} = 0$, $\frac{C_{i+1}}{D_{i+1}} > \frac{C_i}{D_i}$, $\lim_{i\to\infty} \frac{C_i}{D_i} \to +\infty$, $\exists n \in \{\mathbb{N} : n < +\infty\}$,*s.t.*

$$\frac{A+\sum_{i=0}^{k} C_i}{B+\sum_{i=0}^{k} D_i} > \frac{A+\sum_{i=0}^{k+1} C_i}{B+\sum_{i=0}^{k+1} D_i}, k < n; \text{ and } \frac{A+\sum_{i=0}^{k} C_i}{B+\sum_{i=0}^{k} D_i} \leq \frac{A+\sum_{i=0}^{k+1} C_i}{B+\sum_{i=0}^{k+1} D_i}, k \geq n$$

Proof. In the case where $\frac{A}{B} \leq \frac{C_0}{D_0}$, following Lemma 1 it is obvious that having $n = 0$ will satisfy the condition. When $\frac{A}{B} > \frac{C_0}{D_0}$, it can be easily deduced from Lemma 1 that $\frac{A+\sum_{i=0}^{k} C_i}{B+\sum_{i=0}^{k} D_i}$ is monotonically decreasing until $\frac{C_i}{D_i}$ keeps increasing to the point where $\frac{C_{p+1}}{D_{p+1}} \geq \frac{A+\sum_{i=0}^{p} C_i}{B+\sum_{i=0}^{p} D_i}$. In the latter case, having $n = p$ will satisfy the condition. □

Lemma 2. *Given sequences* $(w_i)_{i\in\{\mathbb{N}:i\leq n\}}$, $w_i > 0$, $(w'_i)_{i\in\{\mathbb{N}:i\leq n\}}$, $w'_i > 0$, *and* $(\frac{A_i}{B_i})_{i\in\{\mathbb{N}:i\leq n\}}$, $0 < \frac{A_i}{B_i} < \frac{A_{i+1}}{B_{i+1}}$*: if* $\frac{w_i}{w'_i} > \frac{w_{i+1}}{w'_{i+1}}$, *then* $\frac{\sum_{i=0}^{n} w_i A_i}{\sum_{i=0}^{n} w_i B_i} < \frac{\sum_{i=0}^{n} w'_i A_i}{\sum_{i=0}^{n} w'_i B_i}$.

Proof. We will start the proof with the simplest case $n = 1$ for the purpose of demonstration and then provide the proof for any integer n. When $n = 1$,

$$\frac{\sum_{i=0}^{n} w'_i A_i}{\sum_{i=0}^{n} w'_i B_i} - \frac{\sum_{i=0}^{n} w_i A_i}{\sum_{i=0}^{n} w_i B_i} = \frac{w'_0 A_0 + w'_1 A_1}{w'_0 B_0 + w'_1 B_1} - \frac{w_0 A_0 + w_1 A_1}{w_0 B_0 + w_1 B_1}$$

$$= \frac{(w_0 w_1' - w_0' w_1)(A_1 B_0 - A_0 B_1)}{(w_0' B_0 + w_1' B_1)(w_0 B_0 + w_1 B_1)}$$

Since $\frac{w_0}{w_0'} > \frac{w_1}{w_1'}$, it easily follows that $w_0 w_1' - w_0' w_1 > 0$. Moreover, $\frac{A_1}{B_1} > \frac{A_0}{B_0}$ leads to $A_1 B_0 - A_0 B_1 > 0$. Thus $(w_0 w_1' - w_0' w_1)(A_1 B_0 - A_0 B_1) > 0$, which completes the proof of Lemma 2 in the case $n = 1$.

Next it is proved that Lemma 2 will also hold in a general sense. When $n \geq 2$, with some simplification and manipulation of the formulas, it can be seen that,

$$\frac{\sum_{i=0}^{n} w_i' A_i}{\sum_{i=0}^{n} w_i' B_i} - \frac{\sum_{i=0}^{n} w_i A_i}{\sum_{i=0}^{n} w_i B_i} = \frac{\sum_{i=0}^{n-1} \sum_{j=i}^{n} (w_i w_j' - w_i' w_j)(A_i B_j - A_j B_i)}{(\sum_{i=0}^{n} w_i' B_i)(\sum_{i=0}^{n} w_i B_i)}$$

From $\frac{w_i}{w_i'} > \frac{w_{i+1}}{w_{i+1}'}$, it always holds that $w_i w_j' - w_i' w_j > 0$ since $j > i$; from $0 < \frac{A_i}{B_i} < \frac{A_{i+1}}{B_{i+1}}$, we always have $A_i B_j - A_j B_i > 0$. □

Now we will use the proposed Corollary and Lemmas to prove that Statement 2 mentioned previously holds. If we take a closer look at the proposed objective function equation 4.12, we can rewrite it as

$$Q(H, N \,|x_0, t_0, N_0) = \frac{\mathrm{E}(C_{mr,t_0,N_0+1}|x_0, t_0, N_0) + \sum_{N_0+1}^{N} \left(\mathrm{E}(C_{mr,i}) + c_{pm}\right) + c_{rp}}{t_0 + \mathrm{E}(T_{t_0,N_0+1}|x_0, t_0, N_0) + \sum_{N_0+1}^{N} \mathrm{E}(T_i)} \tag{4.22}$$

For a given H, $\mathrm{E}(C_{mr,t_0,N_0+1}|x_0, t_0, N_0)$ and $\mathrm{E}(T_{t_0,N_0+1}|x_0, t_0, N_0)$ both become constants. The objective function thus can be further simplified as

$$Q(H, N \,|x_0, t_0, N_0) = \frac{\sum_{N_0+1}^{N} \left(\mathrm{E}(C_{mr,i}) + c_{pm}\right) + \Upsilon}{\sum_{N_0+1}^{N} \mathrm{E}(T_i) + \Psi} \tag{4.23}$$

where $\Upsilon = \mathrm{E}(C_{mr,t_0,N_0+1}|x_0, t_0, N_0) + c_{rp}$, and $\Phi = \mathrm{E}(T_{t_0,N_0+1}|x_0, t_0, N_0) + t_0$. It is clear to see that, following Corollary 1.1, if we can prove $\frac{\mathrm{E}(C_{mr,i})+c_{pm}}{\mathrm{E}(T_i)}, i \in \mathbb{N}$ is a non-negative monotonically increasing sequence that approaches positive infinity, logically Statement 2 will be true. Now we will prove that for the objective function proposed in this study, $\frac{\mathrm{E}(C_{mr,i})+c_{pm}}{\mathrm{E}(T_i)}, i \in \mathbb{N}$ indeed has such property.

Proposition 1. *$\frac{E(C_{mr,i})+c_{pm}}{E(T_i)}, i \in \mathbb{N}$ is a non-negative monotonically increasing sequence that approaches positive infinity.*

Proof. It is easy to notice that $E(T_i)$ monotonically decreases with i and $\lim_{i\to+\infty} E(T_i) = 0$. Thus $\lim_{i\to+\infty} \frac{c_{pm}}{E(T_i)} \to +\infty$. Now it suffices to say $\frac{\mathrm{E}(C_{mr,i})+c_{pm}}{\mathrm{E}(T_i)}, i \in \mathbb{N}$ is a non-negative monotonically increasing sequence that approaches positive infinity if $\frac{\mathrm{E}(C_{mr,i})}{\mathrm{E}(T_i)}$ is also monotonically

increasing with i. With some manipulation and $h(x,t) = \exp\big(\lambda_1 \frac{x+\kappa_0\theta_0(\exp(r_{t_0}))^s t}{F}\big)$, $\frac{\mathrm{E}(C_{mr,i})}{E(T_i)}$ can be written as,

$$\frac{\mathrm{E}(C_{mr,i})}{E(T_i)} = c_{mr}\lambda_0 \frac{\int_0^H f_{X(R_{i-1}^+)}(x)\int_0^\infty G(H-x;\,\kappa_0 t,\theta_{t_0})h(x,t)dtdx}{\int_0^H f_{X(R_{i-1}^+)}(x)\int_0^\infty G(H-x;\,\kappa_0 t,\theta_{t_0})dtdx} \tag{4.24}$$

which by the definition of integrals can be written as,

$$\frac{\mathrm{E}(C_{mr,i})}{E(T_i)} = c_{mr}\lambda_0 \frac{\lim_{n\to\infty}\sum_{k=0}^n \frac{H}{n}\Big(f_{X(R_{i-1}^+)}(\frac{kH}{n})\int_0^\infty G(H-\frac{kH}{n};\,\kappa_0 t,\theta_{t_0})h(\frac{kH}{n},t)dt\Big)}{\lim_{n\to\infty}\sum_{k=0}^n \frac{H}{n}\Big(f_{X(R_{i-1}^+)}(\frac{kH}{n})\int_0^\infty G(H-\frac{kH}{n};\,\kappa_0 t,\theta_{t_0})dt\Big)} \tag{4.25}$$

and by taking off $\frac{H}{n}$ from both the numerator and denominator, we will have,

$$\frac{\mathrm{E}(C_{mr,i})}{E(T_i)} = c_{mr}\lambda_0 \frac{\lim_{n\to\infty}\sum_{k=0}^n \Big(f_{X(R_{i-1}^+)}(\frac{kH}{n})\int_0^\infty G(H-\frac{kH}{n};\,\kappa_0 t,\theta_{t_0})h(\frac{kH}{n},t)dt\Big)}{\lim_{n\to\infty}\sum_{k=0}^n \Big(f_{X(R_{i-1}^+)}(\frac{kH}{n})\int_0^\infty G(H-\frac{kH}{n};\,\kappa_0 t,\theta_{t_0})dt\Big)} \tag{4.26}$$

The numerator and denominator can be thought of as the sum of $\int_0^\infty G(H-x,t)h(x,t)dt$ and $\int_0^\infty G(H-x,t)dt$, weighted by $f_{X(R_{i-1}^+)}(x)$, respectively. Referring back to Lemma 2, we can see that if 1) $\frac{f_{X(R_{i-1}^+)}(x)}{f_{X(R_i^+)}(x)}$ monotonically declines with x, and 2) $\frac{\int_0^\infty G(H-x;\kappa_0 t,\theta_{t_0})h(x,t)dt}{\int_0^\infty G(H-x;\kappa_0 t,\theta_{t_0})dt}$ monotonically increases with x, $\frac{\mathrm{E}(C_{mr,i})}{E(T_i)}$ will definitely be increasing with i monotonically. We will now prove that both the above statements are true.

Using equation 4.7, we have the mathematical expression of $\frac{f_{X(R_{i-1}^+)}(x)}{f_{X(R_i^+)}(x)}$ as follows

$$\frac{f_{X(R_{i-1}^+)}(x)}{f_{X(R_i^+)}(x)} = \frac{\Gamma(\alpha_{i-1}+\beta_{i-1})}{\Gamma(\alpha_i+\beta_i)}\frac{\Gamma(\alpha_i)\Gamma(\beta_i)}{\Gamma(\alpha_{i-1})\Gamma(\beta_{i-1})}(\frac{x}{H})^{\alpha_{i-1}-\alpha_i}(1-\frac{x}{H})^{\beta_{i-1}-\beta i} \tag{4.27}$$

Taking derivative of the above function over x, we have

$$\frac{d(\frac{f_{X(R_{i-1}^+)}(x)}{f_{X(R_i^+)}(x)})}{dx} = \frac{\Gamma(\alpha_{i-1}+\beta_{i-1})}{\Gamma(\alpha_i+\beta_i)}\frac{\Gamma(\alpha_i)\Gamma(\beta_i)}{\Gamma(\alpha_{i-1})\Gamma(\beta_{i-1})}(\beta_{i-1}-\beta_i)(-\frac{1}{H}) \tag{4.28}$$

where, based on our previous definition, we have $\alpha_{i-1} = \alpha_i$ and $\beta_{i-1} > \beta_i$. It is thus plain to see that the value of equation 4.28 will always be negative. Namely, $\frac{f_{X(R_{i-1}^+)}(x)}{f_{X(R_i^+)}(x)}$ always monotonically decreases with x.

Furthermore, we can easily prove that $\frac{\int_0^\infty G(H-x;\kappa_0 t,\theta_0(\exp(r_{t_0}))^s)h(x,t)dt}{\int_0^\infty G(H-x;\kappa_0 t,\theta_0(\exp(r_{t_0}))^s dt}$ monotonically increases with x by taking its partial derivative over x, which gives us

$$\frac{\partial(\frac{\int_0^\infty G(x,t)h(x,t)dt}{\int_0^\infty G(x,t)dt})}{\partial x} = \frac{\int_0^\infty \frac{\partial G(x,t)}{\partial x}h(x,t)dt \int_0^\infty G(x,t)dt}{(\int_0^\infty G(x,t)dt)^2} + \frac{\int_0^\infty G(x,t)\frac{\partial h(x,t)}{\partial x}dt \int_0^\infty G(x,t)dt - \int_0^\infty \frac{\partial G(x,t)}{\partial x}dt \int_0^\infty G(x,t)h(x,t)dt}{(\int_0^\infty G(x,t)dt)^2} \tag{4.29}$$

$\frac{\partial G(x,t)}{\partial x}$ and $\frac{\partial h(x,t)}{\partial x}$ in the above equation can be replaced with $-\frac{x^{\kappa-1}e^{-\frac{x}{\theta}}}{\theta^\kappa\gamma(\kappa,\frac{x}{\theta})}G(x,t)$ and $\frac{\lambda_1}{F}h(x,t)$, respectively. Then equation 4.29 is rewritten as,

$$\frac{\partial(\frac{\int_0^\infty G(x,t)h(x,t)dt}{\int_0^\infty G(x,t)dt})}{\partial x} = \frac{\frac{\lambda_1}{F}\int_0^\infty G(x,t)h(x,t)dt \int_0^\infty G(x,t)dt}{(\int_0^\infty G(x,t)dt)^2} \tag{4.30}$$

Since equation 4.30 will always take positive values, $\frac{\int_0^\infty G(H-x;\kappa_0 t,\theta_0(\exp(r_{t_0}))^s)h(x,t)dt}{\int_0^\infty G(H-x;\kappa_0 t,\theta_0(\exp(r_{t_0}))^s dt}$ increases with x. Following Lemma 2, we can thus conclude that $\frac{E(C_{mr,i})}{E(T_i)}$ increases with i. □

It is thus proved that Statement 2 holds and that the proposed algorithm will lead to the global optimal solution for the objective function at the individual asset level.

4.7 Numerical Examples and Discussion

In this section, numerical analysis of the proposed individual asset model is presented for the following purposes: 1) to facilitate a general understanding of the characteristics of the model itself; 2) to provide verification of the model using extreme case scenarios; 3) to provide a thorough picture of how sensitive maintenance decisions are to various factors.

4.7.1 Numerical Example

A numerical example is given in this subsection, which illustrates the relationship between the objective function - the expected remaining maintenance costs averaged over the expected life cycle of an asset conditioned on its actual state at the moment, and different values of the decision variables $[H, N]$. The model parameters used in the example are given in Table 4.1, and the range of H is between X_0 and F ([0, 100]). The value that the objective function takes for different combinations of $[H, N]$ is shown in Figure 4.7. Notice how the minimal value of the objective function of different Ns keeps dropping until $N = 2$ and then starts going up, which aligns with our proof in the previous section. The optimal pair of decision

variable is found to be $[67, 2]$ which gives an average expected remaining maintenance cost of 95.63.

Table 4.1 Individual asset model parameters used for illustration

κ_0	θ_0	s	λ_0	λ_1	X_0	N_0	r
1.2	0.4	0.5	0.001	6.907	0	0	1
c_{mr}	c_{pm}	c_{rp}	α_0	β_0	b	W	F
600	2000	10000	2	12	1	100	100

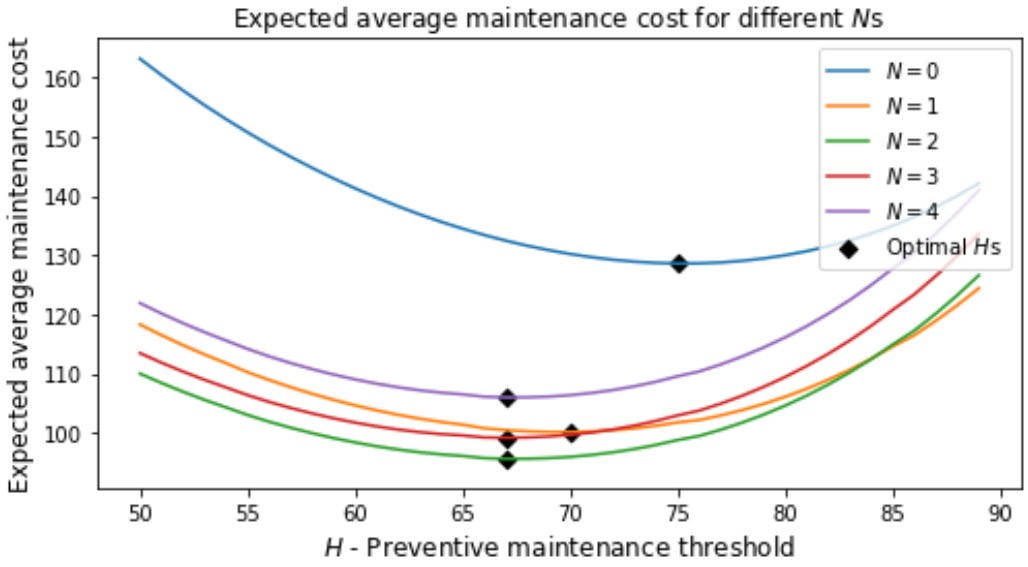


Fig. 4.7 Expected average maintenance costs for different $[H, N]$

4.7.2 Verification Using Extreme Scenarios

In certain extreme scenarios, it is quite straightforward to intuitively come up with the optimal solution. Such cases can be utilised to partially verify the correctness of the model and the optimisation algorithm. Here we test the model under two special case scenarios.

Case Scenario 1

In this case, the same parameters are used as those shown in Table 4.1 except for the preventive maintenance cost c_{pm}, which is set to be the same as the replacement cost $c_{pm} = c_{rp} = 10000$. Since replacement completely restores an asset to 'as-good-as-new' state whereas preventive maintenance only partially rectifies the accumulate damage of the

asset, it is obvious that no preventive maintenance should be performed. Indeed the model produced $[H^*, N^*] = [75.31, 0]$ as the solution. The asset will therefore be replaced the first time its degradation level reaches H^*.

Case Scenario 2

In this case, the only parameter that is set to be different from those in Table 4.1 is the minimal repair cost c_{mr}. This example attempts to demonstrate how the model handles situations where no costs are associated with shock failures. Theoretically $c_{mr} = 0$ should be used. However, this may render the denominator of the objective function to be 0 due to the nature of the model. c_{mr} is thus given a very small value to avoid the trouble. In this scenario, one would expect the optimal value of H to be very close to the failure threshold $F = 100$. The H^* obtained by the algorithm is indeed 99.99.

4.7.3 Model Characteristics Analysis

This subsection presents a thorough analysis of the proposed model based on numerical examples, the purpose of which is to demonstrate the ability of the model to capture the influence of various factors on the optimal maintenance decision. The set of parameters used in the analysis is the same as those outlined in Table 4.1 except for what is specified to be different in each scenario. The rest of the subsection is based on the optimal solutions output by the proposed model with the designed parameters, and for a series of load ratios $[0.0, 0.2, 0.4, 0.6, 0.8, 1.0]$.

Influence of Load Sensitivity Coefficient

The load sensitivity coefficient s describes how strongly the degradation behaviour of an asset reacts to increasing load ratios. $s = 0$ denotes load-independent degradation characteristics, as can be seen from Fig 4.8 that shows identical H^* and minimal expected average cost regardless of the load ratios. An increase in s does not only give rise to higher minimal expected average cost in all load levels, but also leads to larger change in the cost for the same amount of difference in load ratios. An increase in the optimal threshold H^* is observed for a higher load ratio r as well as for a larger s. As asset degradation rate is positively correlated with both r and s, a reasonable speculation is that the model attempted to obtain longer operating time by pushing preventive maintenance further away. However, the extent to which the threshold can be lifted is constrained by the increased vulnerability to external random shocks, which will lead to higher minimal repair costs. Another interesting observation from Table 4.2 is that N^* is always equal to 2 for any s and r, which implies

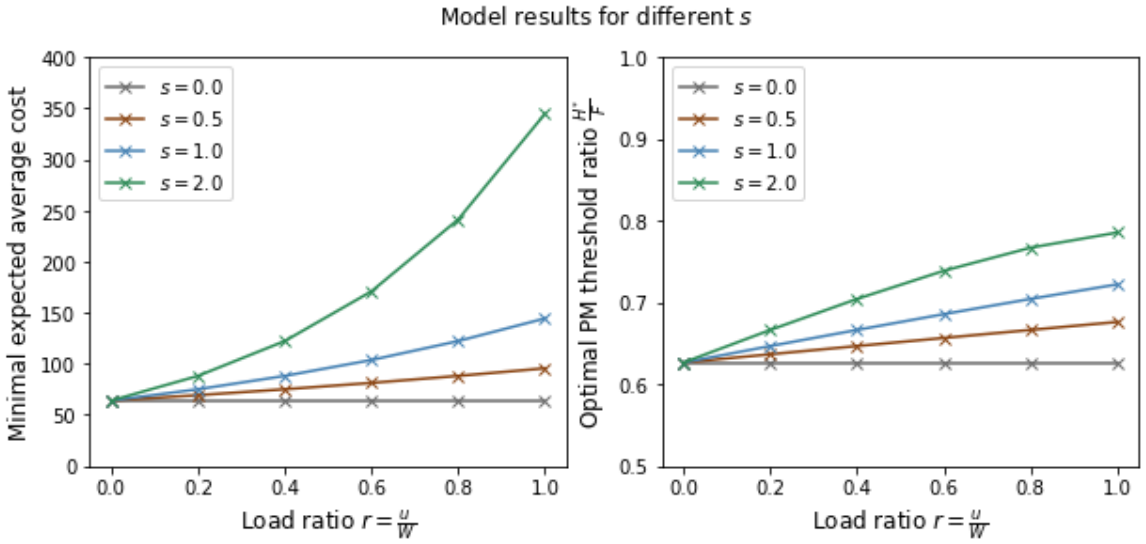


Fig. 4.8 Expected optimal results obtained for assets of different load sensitivity

in this case the trade-off between elongated inter-maintenance duration and increasingly frequent shock-rendered machine breakdowns is more important than that between the cost-effectiveness of replacement and preventive maintenance.

Table 4.2 N^* for assets of different load sensitivity s

Load sensitivity s	Load ratio r					
	0.0	0.2	0.4	0.6	0.8	1.0
0.0	2	2	2	2	2	2
0.5	2	2	2	2	2	2
1.0	2	2	2	2	2	2
2.0	2	2	2	2	2	2

Influence of Rectification Effect

The rectification effect of imperfect preventive maintenance is quantified by two coefficients, β_0 and b as implied by the equation $\beta = \beta_0 \exp(-ib)$. It can be noticed from the equation that when β_0 decreases or b increases, the rectification effect weakens, and for a larger b it decrease faster with every preventive maintenance done. The numerical results exhibited in Figure 4.9 also shows an increasing trend of the minimum value of the objective function with declining β_0 and increasing b. This is a natural outcome as with the same amount of money spent on preventive maintenance, there is less likelihood of a significant improvement

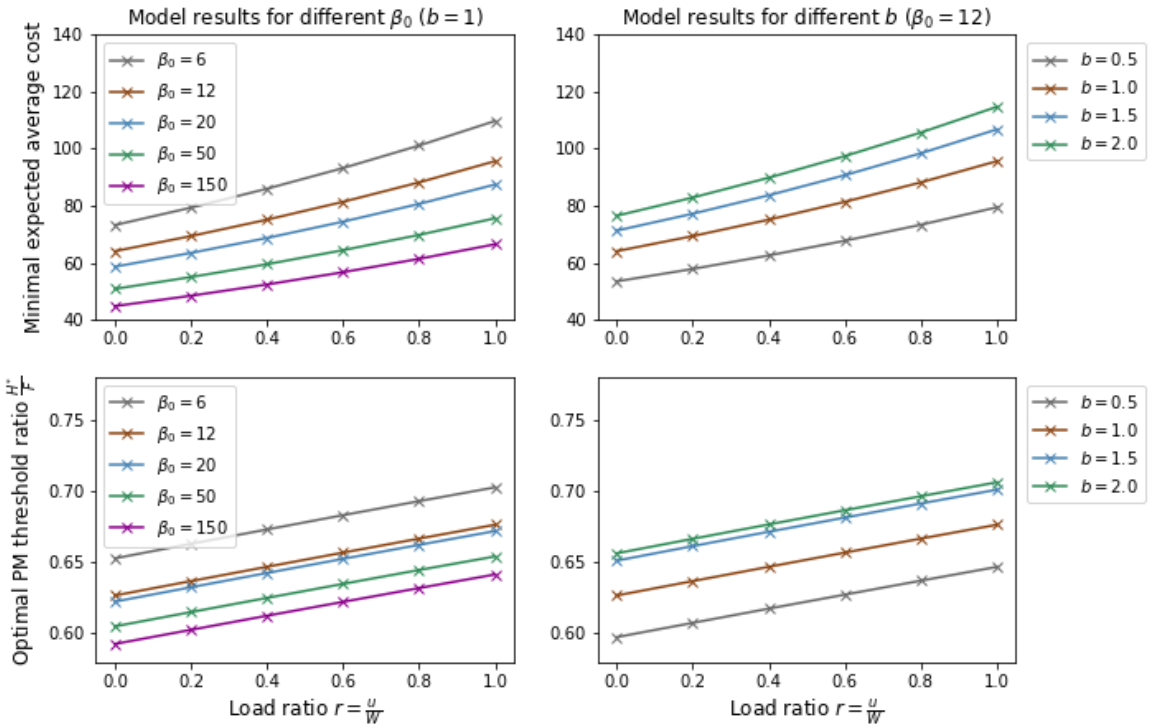


Fig. 4.9 Expected optimal results obtained for assets of different rectification factors

in asset degradation status. Besides, we can observe that, similar to the case of increasing load sensitivity, there is a trend of pushing the preventive maintenance threshold higher as a means to extend asset life. Furthermore, Table 4.3 shows that the model tends to allow more preventive maintenance before a complete replacement with larger β_0 and smaller b, both of which are indicators of more cost-effective preventive maintenance. Another interesting observation is that b has a more dominant role in shaping the model behaviour than β_0. This can potentially be explained by the following two reasons: 1) in equation 4.8, b exists in exponential rather than linear form which is the case with β_0; 2) b controls the diminishing rate of rectification effect every time one additional preventive maintenance is carried out.

Influence of Maintenance Costs

Similar to most cost-based maintenance models, the essence of the proposed model is to find the best balance between replacement cost, cost incurred by random machine breakdowns, and preventive maintenance cost. Thus, it is more reasonable to focus on the relative value of the three components of the total cost, rather than their absolute figures. Here, the results are obtained for a wide range of c_{pm} while the replacement cost and minimal repair cost are kept constant - $c_{rp} = 10000$, $c_{mr} = 600$. Again a rising trend of preventive maintenance

Table 4.3 N^* for assets of different rectification factors

Load ratio r	β_0					b			
	6	12	20	50	150	0.5	1.0	1.5	2.0
0.0	1	2	2	3	4	4	2	1	1
0.2	1	2	2	3	4	4	2	1	1
0.4	1	2	2	3	4	4	2	1	1
0.6	1	2	2	3	4	4	2	1	1
0.8	1	2	2	3	4	4	2	1	1
1.0	1	2	2	3	4	4	2	1	1

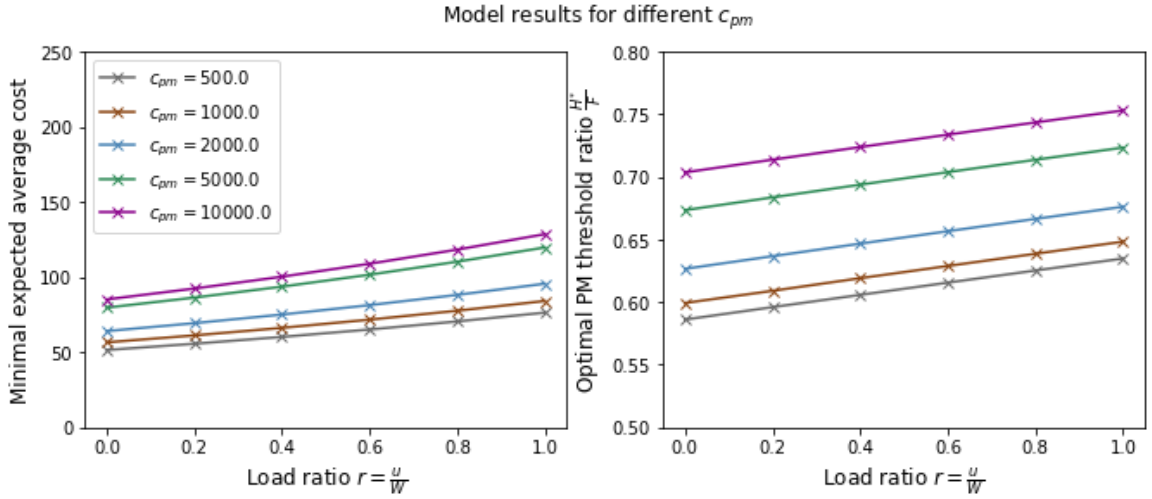


Fig. 4.10 Expected optimal results obtained for assets of different preventive maintenance cost

threshold can be observed when the cost goes up, as shown in Figure 4.10. Since loss of production is not explicitly included in the individual asset model, lower c_{pm} would imply higher cost-effectiveness associated with preventive maintenance tasks. It is desirable to have a lower threshold and thus more frequent preventive maintenance, so that the asset can be kept in a relatively healthy state. As preventive maintenance gets more expensive, such actions become less cost-effective. The data in Table 4.4 also provides evidence in support of this argument as N^* is clearly negatively correlated with c_{pm}, dropping from 3 when $c_{pm} = 500$ to 1 when $c_{pm} = 5000$. In the extreme scenario where $c_{pm} = c_{rp}$, the optimal solution given by the model is to replace the asset without performing any imperfect preventive maintenance.

Table 4.4 N^* for assets of different PM costs c_{pm}

PM cost c_{pm}	Load ratio r					
	0.0	0.2	0.4	0.6	0.8	1.0
500	3	3	3	3	3	3
1000	3	3	3	3	3	3
2000	2	2	2	2	2	2
5000	1	1	1	1	1	1
10000	0	0	0	0	0	0

Influence of Degradation Parameters

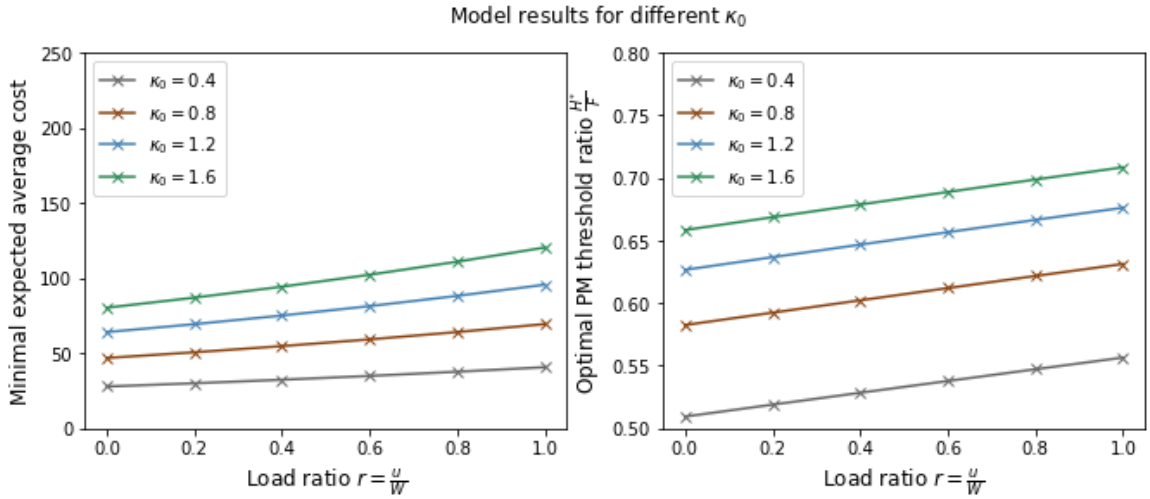


Fig. 4.11 Expected optimal results obtained for assets of different Gamma shape parameters

Another factor that affects the maintenance decision-making on the individual asset level is the degradation characteristic. The Gamma degradation process of the asset is governed by its shape and scale parameters. Figure 4.11 presents the results given by the model for different values of κ_0. It is not surprising to spot an upward trend of the value of the objective function as κ_0 goes up. One more thing to notice is that the difference of the remaining average maintenance cost between a larger and a smaller κ_0 is more significant with higher load ratios. One explanation for this is that the model is trying to reach the best balance between the expected length of its life cycle and the total costs incurred. When the life cycle is extended by the same length, more shocks are likely to take place at higher load ratios, which is further exacerbated by a larger κ_0. Table 4.5 outlines N^* calculated by the model - constant for all combinations of κ_0 and r. As the choice of N^* is more of a sign of the

cost-effectiveness of preventive maintenance, it is less affected by the degradation process of the asset itself under the assumptions of the model.

Table 4.5 N^* for assets of different Gamma shape parameters κ_0

Gamma parameters κ_0	Load ratio r					
	0.0	0.2	0.4	0.6	0.8	1.0
0.4	2	2	2	2	2	2
0.8	2	2	2	2	2	2
1.2	2	2	2	2	2	2
1.6	2	2	2	2	2	2

Influence of Shock Intensity

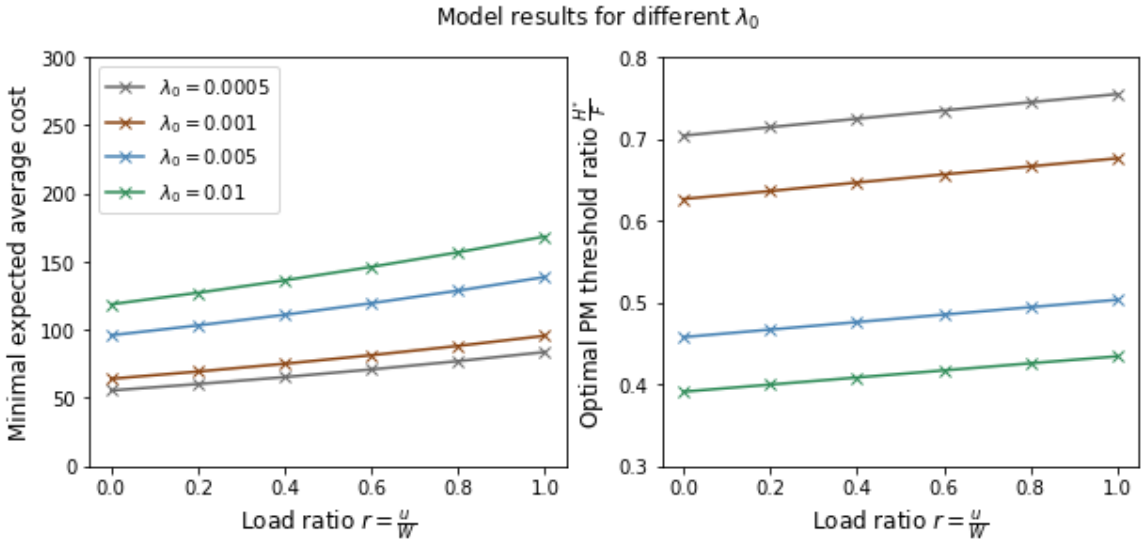


Fig. 4.12 Expected optimal results obtained for assets of different shock intensity

Shock intensity λ_0 can be understood as a determinant of the incoming rate of external random shocks. For the same degradation level, a larger λ_0 leads to higher frequency of shock occurrence. In order to offset threat caused by this effect, assets are inclined to avoid being in a state of high degradation, as evidenced by a declining optimal preventive maintenance threshold in Figure 4.12. Again, we can see that N^* is unaffected by the value of λ_0 within a reasonable range.

Table 4.6 N^* for assets of different shock intensity λ_0

Shock intensity λ_0	Load ratio r					
	0.0	0.2	0.4	0.6	0.8	1.0
0.0005	2	2	2	2	2	2
0.0010	2	2	2	2	2	2
0.0050	2	2	2	2	2	2
0.0100	2	2	2	2	2	2

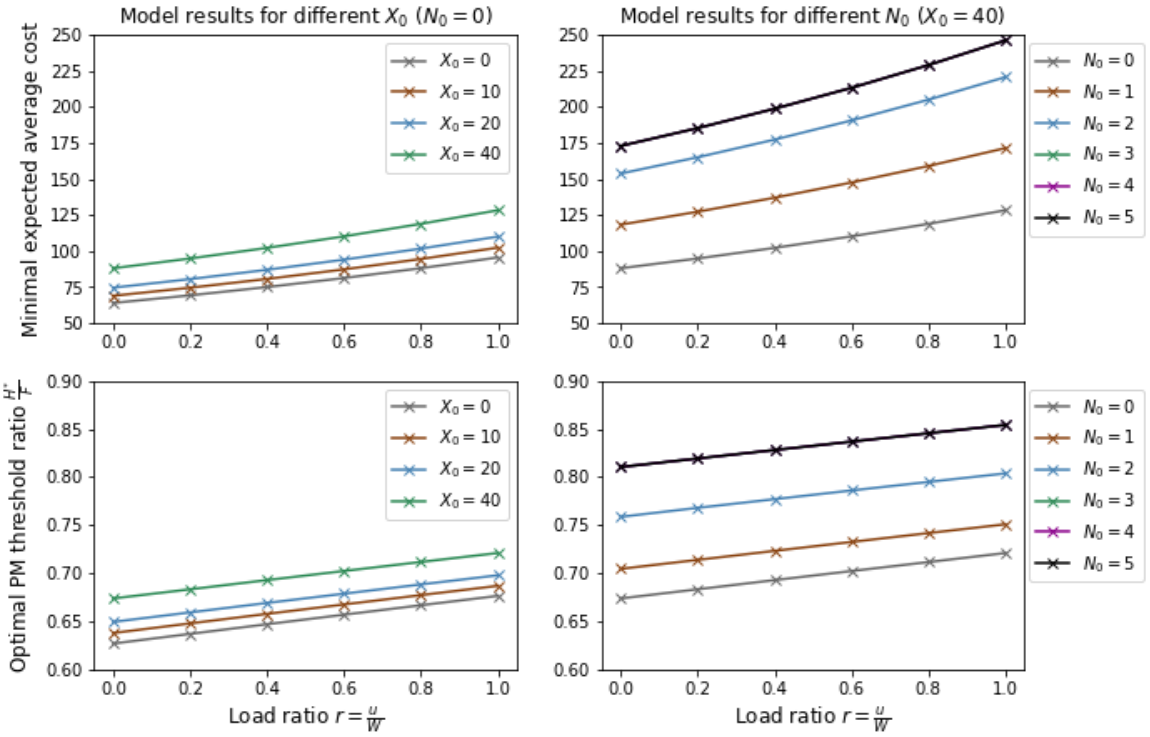


Fig. 4.13 Expected optimal results obtained for assets of different initial conditions

Influence of Asset Initial Condition

Apart from all the aforementioned factors, the initial condition of assets is also an influential element to consider while making maintenance decisions. In reality, we often need to implement maintenance strategies on assets with various degradation behaviour and in different stages of their useful life. For instance, assets might have defects before they are put to use due to deviation and errors in the manufacturing process, insufficient quality control, or wrong modes of transport from the OEM to the user. It can also be the case that the asset is a second-hand purchase and has already undergone several rounds of preventive maintenance, making it inappropriate to assume the asset to be 'as-good-as-new'. It is thus very important

that the maintenance model has the ability to handle such heterogeneity. As can be expected, in terms of minimal expected maintenance cost, the case where $X_0 = 0, N_0 = 0$ outperforms all other combination of X_0s and N_0s. It can be observed in Figure 4.13 that when X_0 is increased, the H^* output by the model also goes up. However, the increase in H^* is much less significant than the initial difference between X_0, which can be explained by the fact that the nonlinearly increasing shock intensity at the higher end of the degradation is quickly offsetting the benefits brought by an elongated asset life cycle. Table 4.7 provides data on the distinct behaviours N^* exhibits towards changes in X_0 and N_0. We can see that N^* is generally unaffected by X_0, but rather sensitive to the value of N_0. In this case, $N_0 = 3$ is the break-even point that marks the complete disappearance of the advantage that preventive maintenance has over replacement, as N^* is always equal to N_0 thereafter. It may first seem counterintuitive that in Figure 4.13 the plots for $N_0 = 3, 4, 5$ are overlapped, which is a consequence of assuming the same initial degradation $X_0 = 40$ for all 3 cases. However, it is less likely that this will happen in reality as the distribution of asset condition is noticeably different immediately after the $3^{rd}, 4^{th}$, and 5^{th} preventive maintenance.

Table 4.7 N^* for assets of different initial conditions

Load ratio r	X_0				N_0					
	0	10	20	40	0	1	2	3	4	5
0.0	2	2	2	2	2	3	3	3	4	5
0.2	2	2	2	2	2	3	3	3	4	5
0.4	2	2	2	2	2	3	3	3	4	5
0.6	2	2	2	2	2	3	3	3	4	5
0.8	2	2	2	2	2	3	3	3	4	5
1.0	2	2	2	2	2	3	3	3	4	5

4.7.4 Discussion and Remarks on the Individual Asset Model

Through the numerical analysis, a couple of general observations can be made from the figures in Section 4.7.3:

1. The expected remaining maintenance costs averaged over the expected life cycle increases as the load ratio goes up, which is a natural consequence as higher loads shorten the lifespan of assets.

2. While comparing the average costs for different values of a certain parameter, the increase in cost is more significant at the higher end of the load ratio. A closer

examination of the objective function as given by equation 4.12 reveals that the individual asset will try to reach the perfect balance between the expected length of its life cycle and the total costs incurred. When the life cycle is extended by the same length, more shocks are likely to take place at higher load ratios, leading to more significant increases in minimal repair costs than the case with lower load ratios.

The results presented in in Section 4.7.3 also indicate that among the two decision variables $[H, N]$, H is more prone to changes in certain factors while N to the others. Specifically, the value of H^* is largely determined by factors that influence the trade-off between a longer inter-maintenance time interval and a higher frequency of breakdowns cause by shocks. For instance, load sensitivity s, shock intensity λ_0, and the Gamma process shape parameter κ_0 all belong to this group of factors. The optimal value of the other decision variable N, however, is more of the result of the choice between replacement and preventive maintenance actions, which eventually goes down to the cost-effectiveness of preventive maintenance. In the proposed model, the cost-effectiveness of preventive maintenance action is shaped by both the relative magnitude relation between c_{pm} and c_{rm} characterised by $\frac{c_{pm}}{c_{rm}}$ and the distribution of post-maintenance asset degradation status quantified by β_0 and b. The relationship between the cost of minimal repairs and preventive maintenance is less of a concern for determining the optimal N in the current formulation of the model. This is mainly due to the shock intensity being independent of preventive maintenance actions. More sophisticated models can be developed by introducing other inter-dependencies between the existing model elements.

Comparing the basic characteristics of the model illustrated in the numerical examples to the desired traits identified in Section 4.3, we can see that the model proposed here is able to meet the requirements for the individual asset maintenance model. To be specific, as a stand-alone maintenance optimisation model, it is capable of capturing the impact of different load levels on the rate of asset degradation as well as reacting properly to varying cost-effectiveness of maintenance practice. As a constitutive component of the integrated task/workload allocation and maintenance decision-making strategy, it is able to provide knowledge on how the expected average remaining maintenance cost and time to the next preventive maintenance are affected by various factors, especially the amount of workload assigned to an asset. The proposed model is also capable of updating maintenance decisions dynamically if asset status changes, as demonstrated by the numerical examples of how initial asset condition influences the model behaviour.

4.8 Chapter Summary

The chapter addressed the first two research questions proposed in Chapter 2. The constitutive components of a joint task/workload allocation and condition-based maintenance decision making system are first identified. Then the first identified component - a load-dependent individual asset maintenance model is presented, whose outputs can be used as the inputs for subsequent decision-making processes. The model proposed in this chapter has been designed to meet the requirements identified in Chapter 2 as well as to partially address the challenges faced by the industry as presented in Chapter 3. After evaluating existing models against requirements imposed for purpose of this study, a model following a preventive maintenance number counting policy is designed in order to minimise the expected remaining maintenance costs averaged over the expected life cycle of an asset, conditioned on its actual current state. The major elements that constitute this model include the following: 1) a load-dependent stochastic process; 2) a degradation-dependent random shock process; 3) imperfect preventive maintenance. The approach taken to obtain the optimal solution is then elaborated and a detailed proof is provided to validate the uniqueness and optimality of the solution obtained by the proposed approach. Results of numerical examples are also presented to partially verify the correctness and elicit a thorough understanding of the model. The model has exhibited the capability to capture the impact of loads on maintenance decisions and to perform appropriate trade-off between various cost components considering the actual state of the asset. In the next chapter, a complete picture of the joint task/workload and maintenance decision making strategy will be presented, where the proposed model is used as the basis for quantifying the impact of workload allocation strategies on maintenance costs.

Chapter 5

Coordinated Workload Allocation Strategy for Parallel Assets

5.1 Introduction

Following the individual model, this chapter aims to contribute to the third research question proposed in Chapter 2. Recall that the previous chapter is mainly focused on the first constitutive component of the joint maintenance threshold and workload allocation decision-making model - the load-dependent individual asset maintenance optimisation model, this chapter will be used to give a detailed description of the second constitutive component - the coordinated workload allocation strategy, as well as a step-by-step framework of how these two components come together to form the proposed decision-making model. Furthermore, we will demonstrate through numerical examples and qualitative analysis how certain traits of the proposed methodology can help address the research gaps and real-world challenges faced by the industry, such as those related to the dynamic update of maintenance decisions and the need to understand the impact of short-term system performance on that of the long term. In order to highlight the cost-saving benefits of adopting the proposed approach, we will compare its performance against traditional practices of workload allocation (e.g., uniform allocation, random allocation, etc.).

Sensitivity analysis on some of the model parameters are also conducted in this chapter to: 1) gain deeper insights into the behaviour and working mechanism of the proposed decision-making model; and 2) advance the understanding of the applicability and limitations of the model.

In short, this chapter attempts to: 1) introduce an innovative methodology for integrated optimisation of workload allocation and condition-based maintenance; 2) assess the effec-

tiveness of the model by its computational efficiency and cost-saving potential; 3) discuss the limitations of the proposed methodology.

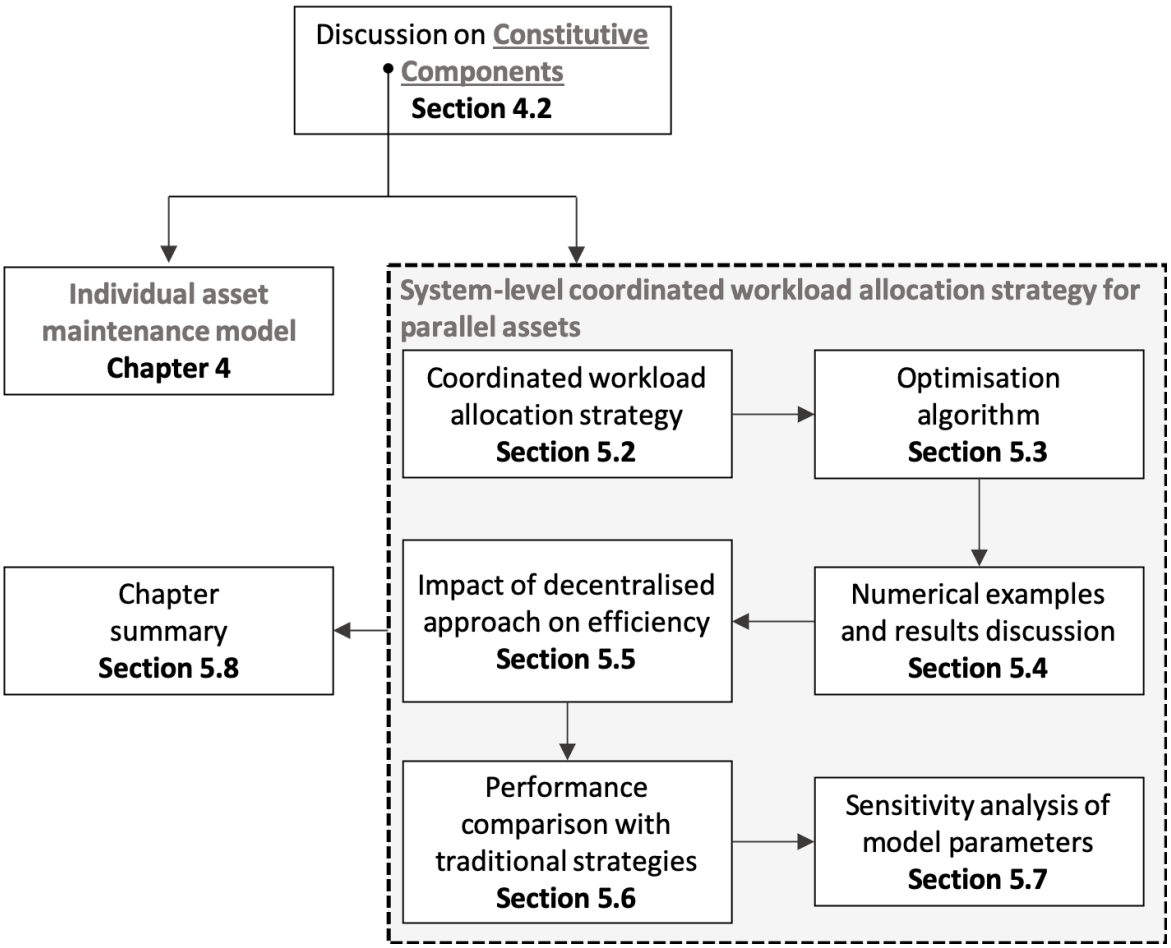


Fig. 5.1 The outline followed by the coordinated strategy chapter

The outline for this chapter is presented in Figure 5.1, which complements the outline for Chapter 4. Section 5.2 first gives a brief description of the problem setting and then introduces the agent-based structure and formulation of the integrated decision-making strategy. The optimisation algorithm adopted for the system-level decision-making model is then presented in Section 5.3, followed by some numerical examples and discussion of the model performance in Section 5.4. Section 5.5 is mainly focused on analysing the impact of the decentralised approach on computational efficiency as well as how various factors affect the intensity of such impact. Section 5.6 provides a comparison between the performance of the proposed model and that of traditional strategies, and Section 5.7 is devoted to conducting sensitivity analysis of some of the model parameters. The chapter is closed with a brief summary in Section 5.8.

5.2 Coordinated Workload Allocation Strategy

This section aims to develop the coordinated workload allocation strategy for parallel assets and present the integrated decision-making model in its complete form.

5.2.1 Problem Description

We consider a system that consists of M parallel production units/machines. Function-wise, all of the machines in the system are identical. To be specific, they are designed for the purpose of producing the same type of product or performing the same type of task. Note that though here we assumed that the machines have the same capacity, the model proposed in this thesis is equally applicable to machines with various capacities. At each evenly-spaced time epoch ε_k, the demand for such products, denoted as D_k, is released into the system and is due to be fulfilled by the next time epoch ε_{k+1}. It is assumed here that the finished products cannot be stored, and that backlogs are not allowed. The machines all have limited production capacity and are subject to an inevitable degradation process. The degradation behaviour of the machines is assumed to be task/workload-dependent as introduced in Chapter 4. Other characteristics related to each machine, such as shock failures, minimal repairs, imperfect preventive maintenance, and replacement have also been defined and elaborated in Chapter 4.

Note that previously, the time spent on maintenance was assumed to be negligible since it is not necessary to calculate loss of production at the machine level. The only costs considered in Chapter 4, therefore, were those associated with the resources consumed by maintenance actions. It is however not the case at the system level. For instance, consider a three-machine system with each machine capable of producing 100 components per day. Assume that the system has a constant daily demand for 200 components, and two machines are taken off-line concurrently for a two-day maintenance task. At the machine level, neither of the non-operational machines has the ability to estimate the production loss as they have no visibility of the condition of each other. Only to some entity that has a complete view of the system will it become obvious that a total production loss of 100 components will be generated. In view of the situation described above, it is clear that such considerations are only necessary at the coordinated workload allocation level. We denote the time needed for each minimal repair, imperfect preventive maintenance, and replacement for machine m as $t_{mr}^m, t_{pm}^m, t_{rp}^m$, respectively. We also denote the total amount of unmet demand for the time period between ε_k and ε_{k+1} as l_k and the resultant total monetary penalty for this period as q_k. It is also assumed that q_k is a function of the demand D_k and total production $U_k = \sum_{m=1}^{M} u_k^m$ between decision epoch ε_k and ε_{k+1}, $q_k = y(U_k, D_k)$. q_k can be seen as a generic measure of both the lost profit and other negative consequences. For instance, failure to meet production

demand is likely to lead to a drop in customer satisfaction, which may then lead to fewer future orders. Though it may be tempting to assume $y(\cdot)$ to be a linear function, it is not always the case in reality. Consider an oil refinery with vessels for treating effluent water, if due to multiple vessel failures the water is discharged into the river without meeting environmental requirements, the refinery has to bear legal responsibilities for such deeds and may even be forced to shut down. The relationship between q_k and ε_k is clearly beyond linear in such a situation. It follows from the above discussion that for a manufacturing system consisting of multiple parallel assets, both maintenance costs and penalty costs resulted from unfulfilled orders need to be taken into account for the purpose of performance control. In view that the degradation behaviour of an asset is closely tied to the workload assigned to it, in the next section we will propose a strategy to dynamically adjust the workload and condition-based maintenance threshold in order to lower the total system-level cost.

5.2.2 Joint Load Allocation and Maintenance Decision-making Model

Recall that in Chapter 4, the two constitutive components of a joint task/workload allocation and maintenance decision-making model are 1) a load-dependent individual asset condition-based maintenance optimisation model; and 2) a coordinated workload allocation strategy. The discussion in Chapter 4 has also led to the conclusion that a multi-agent structure would be suitable for formulating such a model. The multi-agent structure to be used in the proposed methodology is presented in Figure 5.2.

Each machine is assigned a machine agent that monitors its health condition. These machine agents are also responsible for generating and sending relevant information to a coordinator agent who has more visibility over the entire system. Workload is then allocated accordingly by the coordinator agent to the machines. Specifically, we argue that machine agents need to provide the coordinator with the following two types of information for a conscious decision to be made on workload allocation.

1. **Information that facilitates understanding of how the value of the machine-level objective function varies with workloads**
   Given a workload and using the individual asset model, the machine agent will attempt to find the pair of $[H^*, N^*]$ that minimises the expected remaining maintenance costs averaged over the expected asset life cycle, as elaborated in Section 4.5. Recall that H denotes the condition-based maintenance threshold and N denotes the maximum number of preventive maintenance an asset can go through before it is replaced. The optimal value of the machine-level objective function represents the lowest expected increase in daily maintenance cost a machine can expect if it is assigned a certain

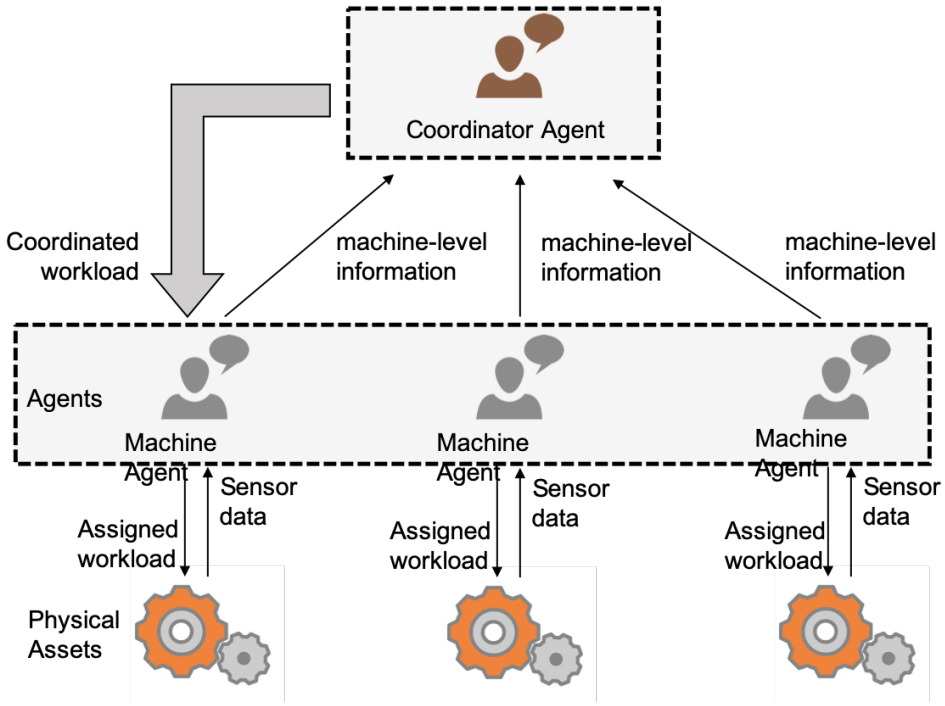


Fig. 5.2 The multi-agent structure to be used in the proposed decision-making model

workload for its remaining life cycle. It can also roughly be seen as the additional maintenance cost incurred for the period of time between two decision epochs for a machine under a specific workload. This information is deemed useful as it enables the model to address the second challenge identified in Chapter 3. The possession of such knowledge allows the coordinator to relate individual performance to system performance by establishing a link between marginal maintenance cost at the machine level, the share of production demand each machine contributes, and the total production loss if any. It is also worth mentioning that since the individual assets have picked their preferred $[H^*, N^*]$ based on the total cost rate related to their entire life cycle, part of the long-term considerations have already been built into the information to be sent to the coordinator agent. This long-term effect, however, has only captured maintenance-related costs. Other arrangements need to be made for the long-term penalty costs to be included, which is accounted for by the second type of information to be elaborated in the next paragraph.

2. **Information that generates insights into how the expected time till the next preventive maintenance or replacement of a machine varies with workloads**
   For a machine undergoing a given workload, the optimal solution $[H^*, N^*]$ can be obtained through the individual model. One by-product generated over the calculation

process is the expected time till the next preventive maintenance, or the replacement if the machine is approaching the last stage of its life cycle. Since the asset degradation is a stochastic process, the by-product is just an estimate of when an asset will actually be maintained or replaced. Even so, it is of great importance to resolve conflicts between short-term benefits and long-term performance.

To elaborate further on this statement, we will take a hypothetical two-machine system as an example and assume that the two machines have different initial deterioration. Despite the fact that the parallel configuration guarantees some level of redundancy and robustness, the system may still suffer from production losses if both machines are taken down simultaneously for maintenance. Without any long-term view, an intuitive strategy would be to load the newer machine to its full capacity and then leave the remaining workload to the older one, implying a higher rate of degradation for the newer machine and vice versa. With everything else being equal, the two machines will eventually reach the same level of deterioration and get exposed to identical risks of failure, leading to a higher probability of overlapping maintenance. What seems to be a good strategy in the short term may turn out to be the opposite in the long term. We can even find a counter-intuitive solution proposed by Hao et al. [42], where it is preferable to allocate the most workload to the most degraded machine to ensure a satisfactory overall long-term performance. It can be elicited from the discussion above that having the coordinator gain partial visibility over the future time slots for maintenance tasks will to some extent help avoiding myopic decisions. The second piece of information here also provides what the first one is short of - the consideration for long-term penalty costs.

Next we will present the decision-making model in its entirety. A detailed flowchart of the proposed methodology can be found in Figure 5.3. The decision-making process is as follows:

1. **First-round machine-level calculation and data preparation**
   It would be ideal if a unit in operation, indexed by m, could obtain analytical expressions of the optimal objective function value $(Q^*(H, N|x_0, t_0, N_0))^m$, and the expected time to the next preventive maintenance or replacement $\mathrm{E}(T_{t_0,N_0+1}|x_0, t_0, N_0)^m$, as functions of the load ratio r and send these functions to the coordinator. This is, however, not feasible in reality due to mathematical complexity of the model formulation. Here we adopt a regressed quadratic function and a piece-wise linear function to approximate the actual expressions of $(Q^*(H, N|x_0, t_0, N_0))^m$ and $\mathrm{E}(T_{t_0,N_0+1}|x_0, t_0, N_0)^m$, respectively.

First, at the beginning of a decision epoch ε_k, for machine m, its machine agent will check the unit status quantified by the set of variables $[x_0^m, t_0^m, N_0^m]$, and calculate a series of $[(H^*)_n^m, (N^*)_n^m]$ and $(Q^*(H, N|x_0, t_0, N_0))_n^m$ for a selection of n load ratios $[r_1, r_2, \cdots, r_n]$. Once the calculation for the required $[(H^*)_n^m, (N^*)_n^m]$ has been completed for all the selected load ratios, the machine agent will continue to calculate the expected time to the next preventive maintenance or replacement $\mathrm{E}(T_{t_0,N_0+1}|x_0, t_0, N_0)_n^m$. Regressing the obtained $(Q^*(H, N|x_0, t_0, N_0))_n^m$ and $\mathrm{E}(T_{t_0,N_0+1}|x_0, t_0, N_0)_n^m$ against $[r_1, r_2, \cdots, r_n]$, we will have the following functions:

$$(Q^*(H, N|x_0, t_0, N_0))^m(r) = v^m(r) \tag{5.1}$$

$$\mathrm{E}(T_{t_0,N_0+1}|x_0, t_0, N_0)^m(r) = z^m(r) \tag{5.2}$$

It is apparent that the larger the value of n, the closer the regressed functions will be to the actual relationship between $\mathrm{E}(T_{t_0,N_0+1}|x_0, t_0, N_0)$, $(Q^*(H, N|x_0, t_0, N_0))$, and r. The choice of n, however, will be restricted to the computational power available and may be determined through trial and error. Furthermore, operational parameters might also restrict these options as some machines are designed to work only at full and half load ratios. As this is out of the scope of this research study, for now we will leave this issue to future research and will not elaborate further. At the end of the first-round machine-level calculation, the machine agents of all operational machines will send their own $v^m(r)$ and $z^m(r)$ to the coordinator agent. Note that $v^m(r)$ and $z^{m(r)}$ respectively correspond to the first and second piece of information identified as needed by the coordinator.

2. **Coordinator-level workload allocation**
   The aim of the coordinator agent is to find the best workload allocation based on its knowledge of the production demand and the relevant information sent in by the machine agents. Since some kind of long-term view is expected from the coordinator in the decision-making process, it is essential to first determine how far into the future the model should examine. Although at the machine level, the machine agent is able to calculate the expected time duration between any consecutive PM actions, the accuracy of such predictions declines after the immediate next PM. This is mainly because the prediction for the occurrence time of the second next PM is dependent upon the asset condition after the immediate next PM, which is a random variable itself. The extra layer of randomness erodes the confidence of machine agents on

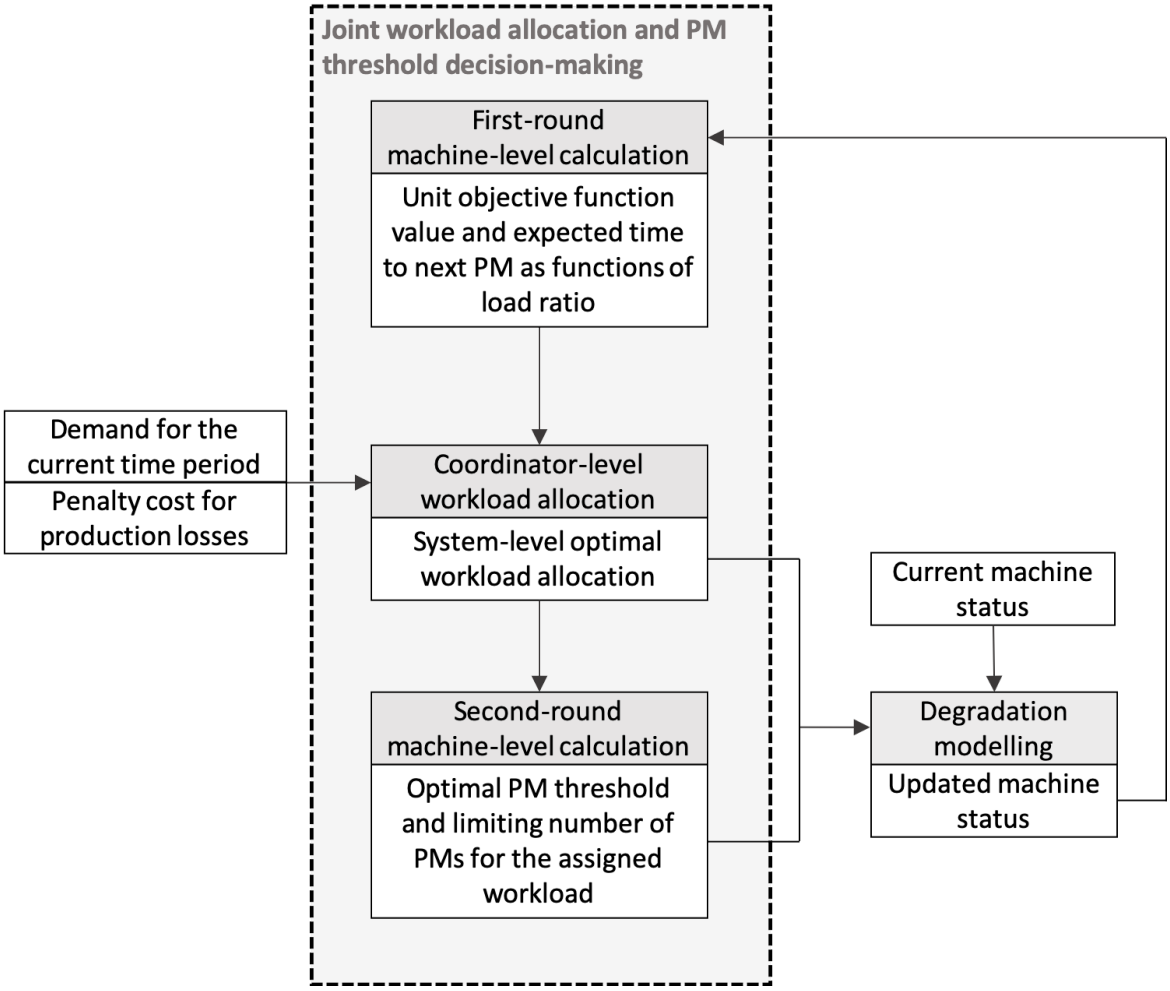


Fig. 5.3 Flowchart of the joint workload allocation and maintenance decision-making model

the predicted results. It is therefore more cost-effective to only look as far as the immediate next PM (or replacement if no more PM should be performed) for the machines. One more thing worth mentioning is that machine agents are very likely to report different $\mathrm{E}(T_{t_0,N_0+1}|x_0,t_0,N_0)^m$ since machines could be at different states within their life cycle. In the proposed methodology, the model is designed to have a relatively long-term view which ends at the latest $\mathrm{E}(T_{t_0,N_0+1}|x_0,t_0,N_0)^m$ among all machines. Since the probability of one machine having undergone two PMs while another has been operating continuously without any PM is rather small, it is assumed here that from the current decision-making epoch till the latest upcoming PM, only one PM will be conducted on each machine.

The coordinator-level objective function which will be proposed later is based on the rationale that, any decision made at the moment will have an instant effect as well

as long-term consequences. If there exists an approach to map such consequences to an equivalent or correlated short-term impact, it is possible to improve long-term performance by adopting a seemingly short-term strategy. In the context of this study, for a specific decision on workload allocation,

- The penalty costs of failing to meet the demand of the current time period is an instant effect caused by this decision.
- The sum of additional maintenance costs incurred for the current time period on all assets can be seen as an short-term reflection of the long-term consequences of maintenance-related costs.
- The potential loss of production due to possible excessive overlapping maintenance in the future is another type of long-term consequences that needs to be considered.

Therefore, the optimisation problem for the coordinator agent is formulated as follows: at each decision epoch ε_k, the coordinator is tasked with finding a workload allocation plan $[u_k^1, u_k^2, \cdots, u_k^M]$ that minimises the following objective function. Here for simplicity in writing, we temporarily omit the subscript k:

$$\min_{[u^1,u^2,\cdots,u^M]} \quad \sum_{m=1}^{M} v^m(r^m)\delta^m + y(\sum_{m=1}^{M} u^m, D) + O(\sum_{m=1}^{M} u^m, D) + \frac{\sum_{l=1}^{L}\sum_{S_l\in \mathbf{S}_l} P(S_l)\cdot y(\bar{U}(S_l), D)}{L} \tag{5.3}$$

$$\text{subject to} \qquad 0 \le u^m \le W^m, \tag{5.4}$$

$$u^m = 0, \text{ if } \delta^m = 0, \tag{5.5}$$

where the first item in the objective function equation 5.3 represents the sum of marginal maintenance costs incurred on all machines for the current time period (this item will be referred to as the maintenance-related component from this point onwards), and the load ratio r^m is calculated as $r^m = \frac{u^m}{W^m}$. The value of δ^m determined by equation 5.6 represents whether a machine is available at the decision epoch, and implies that machines that are off-line are not considered for the calculation of marginal maintenance costs.

$$\delta^m = \begin{cases} 1 & \text{if machine } m \text{ is available at this decision epoch} \\ 0 & \text{if machine } m \text{ is not available at this decision epoch} \end{cases}. \tag{5.6}$$

The second item is the penalty cost for production losses over the current time period, whereas the role of the third item is to ensure that the system does not over produce since no inventory or backlog is allowed, and the expression for $O(\sum_{m=1}^{M} u^m, D)$ is given by

$$O(\sum_{m=1}^{M} u^m, D) = \begin{cases} 0 & \sum_{m=1}^{M} u^m \leq D \\ \text{A very large number} & \sum_{m=1}^{M} u^m > D \end{cases}. \tag{5.7}$$

The last item is the expected average potential production losses resulted from possible excessive overlapping maintenance tasks over the considered time horizon in the near future (this item will be referred to as the long-term penalty-related component from this point onwards).

Before giving a detailed explanation of how the long-term penalty-related component is calculated, we first introduce the symbol $(\delta')_l^m$ to denote whether a machine is 'considered' operational at a certain point in time in the near future:

$$(\delta')_l^m = \begin{cases} 1 & \text{if machine } m \text{ is considered operational at the beginning} \\ & \text{of the } l^{th} \text{ epoch from now} \\ 0 & \text{if machine } m \text{ is not considered operational at the be-} \\ & \text{ginning of the } l^{th} \text{ epoch from now} \end{cases}. \tag{5.8}$$

There are two reasons that the word 'considered' is adopted in this context:

(a) $(\delta')_l^m$ is only defined by whether a machine is supposed to be undergoing preventive maintenance or replacement, but neglects the possibility of machine breakdowns at any point in time due to random shocks;

(b) the asset degradation is a stochastic process, it is not possible to obtain the exact point in time for when an asset needs preventive maintenance and we cannot be sure of such estimation. As shown in Figure 5.4, the orange dots only represent the expected rather than an accurately predicted time of the next preventive maintenance/replacement. To allow for some leeway for such uncertainty, a hypothetical 'buffer' is inserted both before and after the estimated time point. The size of the buffer is denoted as B, as shown in Figure 5.4.

The introduction of a buffer is inspired by the work of Hao et al. [42], where they have developed an optimisation model that actively controls the degradation of machines by

adjusting the workload assigned. One of the constraints in their optimisation model is that the predicted residual lives of any two machines that are to fail consecutively should have a difference greater than the repair time. While their approach indeed manages to mitigate loss of production, it does not capture the economic consequences. Moreover, it has not considered the uncertainty of the degradation of assets by using only point estimates for the predicted residual lives. Therefore, we have taken the alternative of setting a buffer with an adjustable size B which would enable the proposed approach to:

(a) take into account the uncertainty associated with the predicted time to the next PM/RP of each machine where a larger B characterises the model being less confident with its predictions;

(b) control the amount of attention that the coordinator devotes to preventing concurrent machine breakdowns where a larger B indicates less willingness to face the consequences of loss of production.

More detailed analysis and discussion of the impacts of the buffer size B will be provided in Section 5.4 and 5.6.

Equation 5.9 gives the approach to calculate L as shown in Figure 5.4, the length of time horizon to be considered.

$$L = \max_{m\in\{\mathbb{N}:1\leq m\leq M\}}(\mathrm{E}(T_{t_0,N_0+1}|x_0,t_0,N_0)^m + t^m_{pm/rp}) + B. \tag{5.9}$$

The first item in the above equation is the latest date of completion of the immediate next preventive maintenance or replacement among all machines, and the second item is simply the 'buffer size'. The concept of a 'buffer' as well as those related to L is illustrated in Figure 5.4, where it also depicts the situation under which a machine is 'considered' operational or not operational.

At every time epoch, even for a machine 'considered' to be operational, random shocks can happen, as explained in Section 4.5.4. A machine will thus be in one of the three states: 1) not considered operational due to preventive maintenance or replacement; 2) considered operational and not failed due to shocks; 3) considered operational but failed due to shocks. Here we use a vector variable S_l to denote a specific combination of machine statuses at the beginning of the l^{th} time interval from now, and $\mathbf{S}_l$ to denote the set of all combinations. $P(S_l)$ is the probability of S_l happening and $\bar{U}(S_l)$ is the total maximum production of machines under S_l.

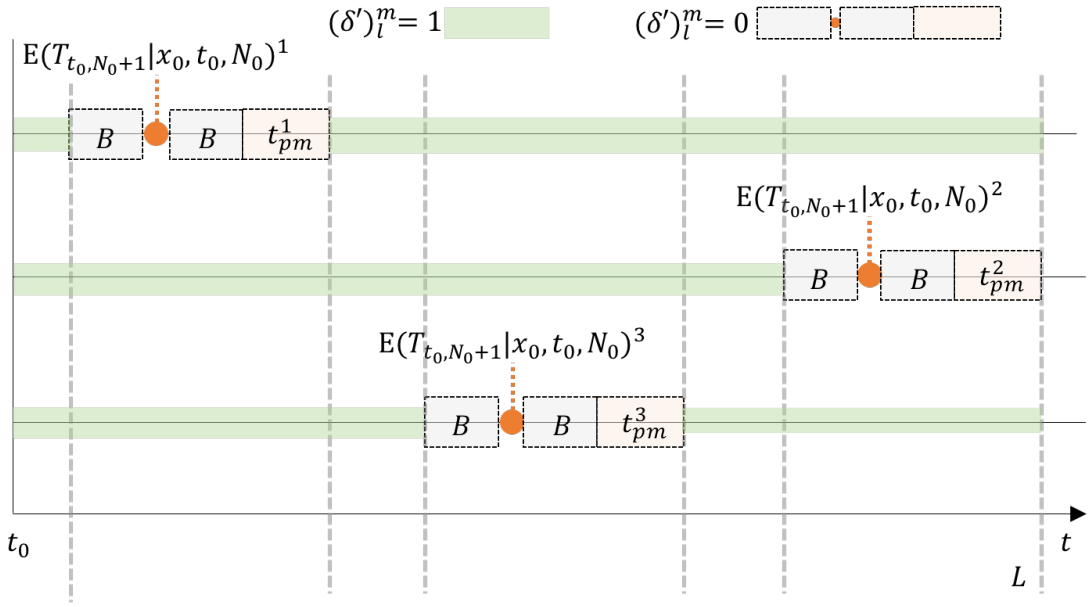


Fig. 5.4 Illustration of constraints for the coordinator-level optimisation

An example is now used to demonstrate the calculation of $P(S_l)$ and $\bar{U}(S_l)$. Consider a system consisting of three units with initial degradation levels $[x_0^1, x_0^2, x_0^3]$, identical capacity W, workload u_0, failure threshold F, gamma degradation process parameters $[\kappa_0, \theta_0, s]$, and random shock parameters $[\lambda_0, \lambda_1]$. Units 1 and 2 are considered to be operational at the beginning of the l^{th} time interval from now on, and Unit 3 is not considered to be operational due to preventive maintenance being expected around that time. The expected degradation level for Unit 1 and 2 at the l^{th} time epoch is given by

$$\mathrm{E}(x_l^1) = x_0^1 + \kappa_0 l \theta_0 (\exp(\frac{u_0}{W}))^s, \text{ and } \mathrm{E}(x_l^2) = x_0^2 + \kappa_0 l \theta_0 (\exp(\frac{u_0}{W}))^s$$

Using equation 4.4, the probability of shocks happening to Unit 1 and 2 are

$$1 - \exp(-\lambda_0 \exp(\lambda_1 \frac{\mathrm{E}(x_l^1)}{F})), \text{ and } 1 - \exp(-\lambda_0 \exp(\lambda_1 \frac{\mathrm{E}(x_l^2)}{F})), \text{ respectively}$$

Then for a status combination S_l with Unit 1 failed due to shocks, Unit 2 being operational as normal, and Unit 3 not being operational due to preventive maintenance,

$$\begin{aligned} P(S_l) &= P(\text{Unit 1 shocks}) \cdot P(\text{Unit 2 no shocks}) \cdot P(\text{Unit 3 not operational}) \\ &= \left(1 - \exp(-\lambda_0 \exp(\lambda_1 \frac{\mathrm{E}(x_l^1)}{F}))\right) \cdot \left(\exp(-\lambda_0 \exp(\lambda_1 \frac{\mathrm{E}(x_l^2)}{F}))\right) \cdot 1 \end{aligned}$$

$$\bar{U}(S_l) = 0 + W + 0 = W$$

In terms of constraints imposed on the decision variables, the first constraint, equation 5.4 limits the range of the assigned workload to a machine be between 0 and the maximum capacity of that machine. The second constraint, equation 5.5 forces machines in maintenance to take on 0 workload. Note that though here we assume that u^m can be any real number between 0 and W^m, the model is equally applicable to cases where the value of u^m is discrete.

3. **Second-round machine-level calculation**
   The last step in the decision-making process is to have the machine agents recalculate their corresponding $[(H^*)_k^m, (N^*)_k^m]$ for the assigned workload u_k^m at the current decision epoch ε_k. The obtained $[(H^*)_k^m, (N^*)_k^m]$ will later be used to check if any maintenance task needs to be performed on machine m at the beginning of the next decision epoch ε_{k+1}.

The next section will provide a brief description of the optimisation algorithm used at the coordinator level.

5.3 Optimisation Algorithm

In view that in reality the levels of workload that an asset can take up can be either discrete or continuous, the optimisation algorithm to be adopted by this study needs to have the capability to solve mixed integer problems. Furthermore, the proposed methodology has not specified the form of the penalty function for production losses, a relatively generic optimisation technique is preferred to allow for more flexibility. Moreover, as can be inferred from the formulation of the optimisation problem, the algorithm will need to be able to deal with various types of constraints. For the reasons mentioned above, we adopted the mixed-integer genetic algorithm (GA) based on the approach named MI-LXPM proposed by Deep et al. [29] in order to solve the coordinator-level optimisation problem.

The optimisation algorithm follows the basic 6-step process of GA: initialising candidate pool, calculating fitness values, selecting parents, crossover and mutation, truncating off-springs, and checking stopping criteria. The pseudocode of the algorithm is given in Figure 5.5, followed by a brief description of the parameters and techniques used in each of the steps.

- **Laplace crossover:** A pair of off-springs are produced from two parent chromosomes using parameters generated from Laplace distribution. The related input parameters are

MI-LXPM: Find the optimal workload allocation vector $[u_k^1, u_k^2, \cdots u_k^M]^*$

Input: Problem Parameters

Output: The optimal workload allocation vector $[u_k^1, u_k^2, \cdots u_k^M]^*$

begin

Initialise the maximum number of iterations N_g, and the iteration count $C_g = 1$

Generate the 1st generation chromosomes – O_g workload allocation vectors

Check stopping criteria

while $C_g < N_g$ **and** stopping criteria not met **do**

foreach workload allocation in the current generation **do**

Calculate its corresponding fitness value – the negative value of the coordinator objective function

end

Check stopping criteria

Generate the $(C_g+1)^{th}$ generation chromosomes with crossover and mutation

Conduct truncation on those that fail to meet the integer constraints

Replace the current generation with the $(C_g+1)^{th}$ generation

Increase C_g by 1

end

Record the chromosome with the best fitness value in the current generation as the optimal workload allocation vector $[u_k^1, u_k^2, \cdots u_k^M]^*$

end

Fig. 5.5 The pseudocode for the Genetic Algorithm adopted by the coordinator

the location parameter a_{GA}, and the scale parameter b_{GA} that can be different (b_{GA}^{real} and b_{GA}^{int}) for real and integer decision variables. The probability of crossover is denoted as P_c.

- **Power mutation:** An off-spring chromosome is created at the vicinity of a parent chromosome using random variable generated from Power distribution. The related input parameter is ρ_{GA} which controls the degree of perturbation. ρ_{GA} can also take different values (ρ_{GA}^{real} and ρ_{GA}^{int}) for real and integer decision variables. The probability of mutation is denoted as P_m.

- **Tournament selection:** Tournament selection approach is applied here as it has been shown by Goldberg and Deb [38] to have better or at least equivalent efficiency while compared with other selection techniques. Tournament selection requires tournament to be played between k_{GA} candidates and the one with the largest fitness value is placed into the mating pool.

- **Truncation procedure:** As the crossover and mutation operators adopted in the algorithm do not guarantee that the obtained off-springs still conform to the integer constraints, truncation is performed on decision variables where needed. For any off-spring $o = (o_1, o_2, \cdots, o_{n_{GA}})$, if o_i should be an integer but is not, its value will be reset to either $[o_i]$ or $[o_i] + 1$ with equal probability. $[o_i]$ denotes the integer part of o_i.

In the next section, a two-machine system will be used as an example for further discussions on model characteristics.

5.4 Numerical Examples and Discussion

The aim of this section is to illustrate the model in action by giving a couple of numerical examples.

5.4.1 Numerical Examples

The objectives of the numerical example are to:

1. provide an overview of different aspects of the modelling results;

2. conduct an analysis on the impact of the long-term penalty-related component (the last item in the coordinator objective function equation 5.3), which adds long-term and system-level views to the decision-making process, on the behaviour of the proposed model.

In order to achieve these objectives, the results between the following two versions of the decision-making model are compared:

1. in the first version, which is referred to as the **basic version**, the last item in equation 5.3, also known as the long-term penalty-related component, is excluded from the objective function used by the coordinator;

2. in the second version, which is referred to as the **complete version**, adopts equation 5.3 the way it is to be the coordinator objective function. For the complete version, the buffer size B can be perceived as an coefficient that controls the weight of the long-term penalty-related component in the objective function. In order to have a comprehensive understanding of the model behaviour, a series of experiments have been run with the buffer size B taking a range of integer values between 0 and 8.

A two-unit system is considered in the numerical example. The parameters for the individual asset model and coordinator-level optimisation are given in Table 5.1. We set a constant demand $D_k = D$ for this example and assume a linear relationship between the penalty cost q_k and the amount of unfulfilled demand $max(0, D_k - U_k)$: $q_k = q \cdot \max(0, D_k - U_k)$. As D is equal to the capacity W of either units, the system is able to meet the demand as long as one of the units is up and running.

Table 5.1 Individual asset model parameters for two units

Unit 1 parameters	κ_0	θ_0	s	λ_0	λ_1	X_0	N_0	F
	2	1	1	0.002	6.907	**0**	0	100
	c_{mr}	c_{pm}	c_{rp}	α_0	β_0	b	W	
	1000	2500	10000	2	12	1	100	
Unit 2 parameters	κ_0	θ_0	s	λ_0	λ_1	X_0	N_0	F
	2	1	1	0.002	6.907	**40**	0	100
	c_{mr}	c_{pm}	c_{rp}	α_0	β_0	b	W	
	1000	2500	10000	2	12	1	100	
Coordinator parameters	D	q	B					
	100	90	0-8					

It can be noticed that the two units share most of the parameters except for one thing: unit 2 has a worse initial condition than unit 1, represented by a larger X_0. The parameters related to the genetic algorithm are given in Table 5.2. The parameters are chosen by taking into account both the optimality of the solution as well as the time taken to achieve such optimality. The values for b_{GA}^{int} and ρ_{GA}^{int} have been omitted since continuous workload is assumed.

Multiple replications of simulation have been run using this set of parameters, and the results from one typical replication are discussed to demonstrate the decision-making rationale adopted by the model. As the interest of this research is not in the steady-state system performance for an infinite time horizon, the discussion of the model characteristics is limited to the point when all units have been replaced for the first time.

Table 5.2 Parameters for the genetic algorithm used for optimisation

GA parameters	a_{GA}	b_{GA}^{real}	b_{GA}^{int}	ρ_{GA}^{real}	ρ_{GA}^{int}	k_{GA}	P_c	P_m
	0	0.1	-	10	-	5	0.85	0.015

Basic Version - Results and Discussions

In this particular replication, the basic version of the model is implemented. It can be observed in Figure 5.7 that Unit 2 is replaced between the 38^{th} and 40^{th} time interval and Unit 1 is replaced between the 44^{rd} and the 46^{th} interval. The workload allocation, unit degradation over time, and the times at which minimal repairs (MR), preventive maintenance (PM), and replacement (RP) take place are shown in Figure 5.6, 5.7, and 5.8, respectively. By examining these three figures, some interesting observations can be made;

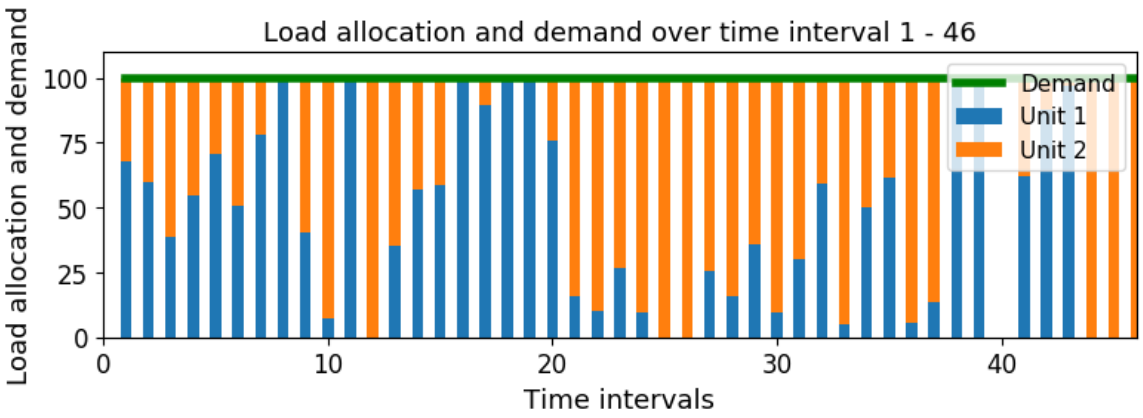


Fig. 5.6 Load allocation and demand over time interval 1-46 for a two-unit system

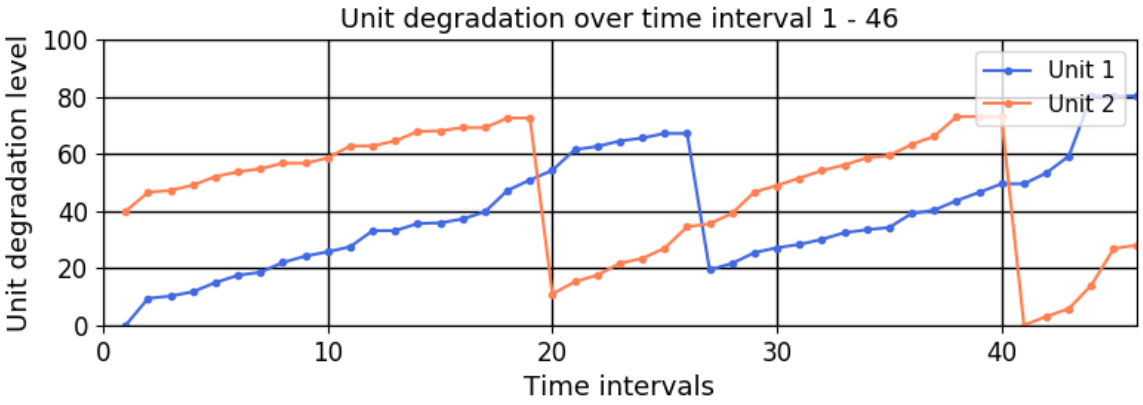


Fig. 5.7 Unit degradation process over time interval 1-46 for a two-unit system

1. For most of the time, the model exhibits a strong tendency to fulfil the demand as best as it can in order to avoid the penalty for loss of production. For instance, at the 8^{th} and 11^{th} intervals, Unit 1 is fully loaded due to shocks happening to Unit 2, as shown in Figure 5.8. This is due to the fact that the penalty for production loss is greater than the additional maintenance cost generated from a heavier workload. Specifically, at the 8^{th} interval, the following relationship between the optimal maintenance cost rate Q^* and workload u is obtained for Unit 1:

$$Q^*(u) = 287.03 + 193.29\frac{u}{100} + 88.85(\frac{u}{100})^2$$

The derivative of the above function is $1.933 + 0.018u$, implying that the largest possible increase in maintenance cost rate is $1.933 + 0.018 \times 100 = 3.733$, which is still much less than the penalty cost for a single unit of production loss $q = 90$. When such condition exists, the model will always prioritise production over self-protection.

2. A clear transition in load allocation preference is seen every time the relative state between two units changes, which usually happens when one of the units goes through PM or RP. To be specific, From the 1^{st} to the 17^{th} interval before Unit 2 is preventively maintained for the first time, the model tends to allocate more workload to Unit 1. However after the first PM of Unit 2, a significant amount of workload is shifted from Unit 1 to Unit 2. Then another wave of redistribution happens after the replacement of Unit 2 has been completed at the 40^{th} interval.

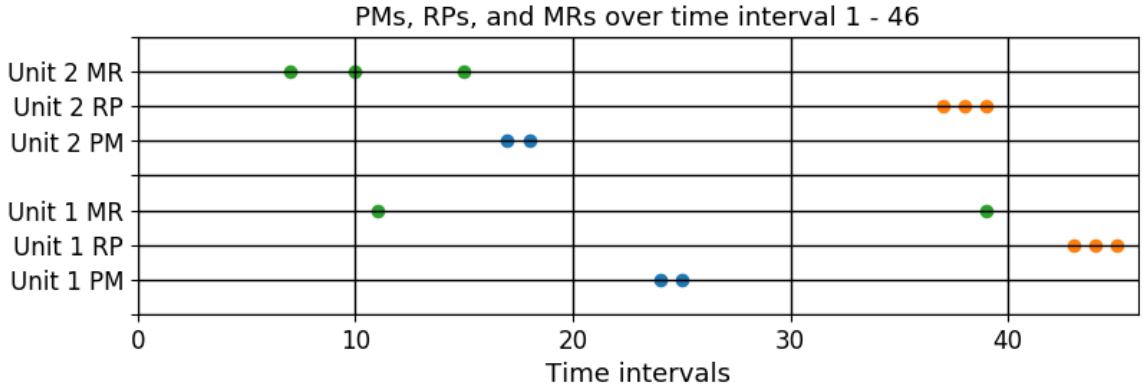


Fig. 5.8 PMs, RPs, and MRs over time interval 1-46 for a two-unit system

The main reason for the second observation is that the coordinator does not allocate workload by comparing the absolute values of maintenance cost rates among the two units. Instead, its decision is based on how fast the maintenance cost rate of units changes with increasing

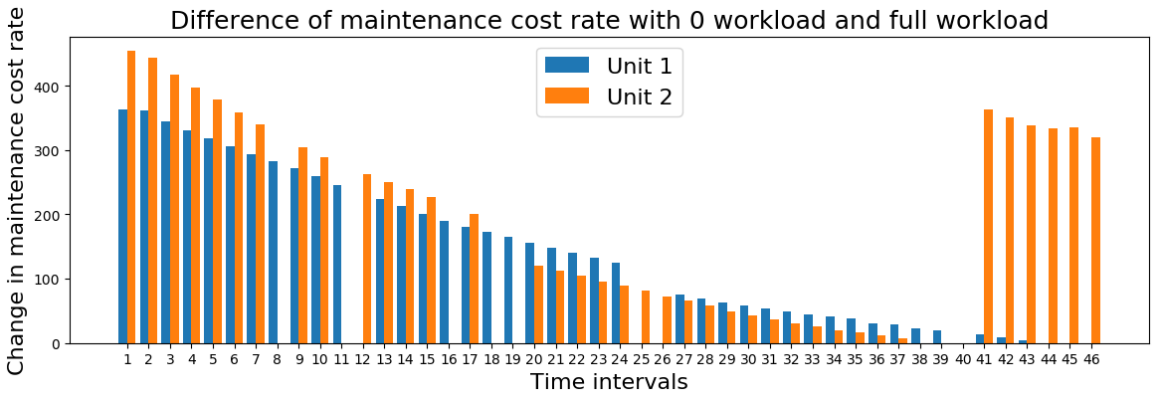


Fig. 5.9 Difference in unit maintenance cost rates with 0 workload and full workload

workload. The difference between maintenance cost rates with no workload and full workload for both units is plotted in Figure 5.9. It is clear that the relative magnitude of the change in maintenance cost rates exactly reversely reflects that of the workload allocation. Also note that the cost rate gap keeps decreasing for both units until a complete replacement, which is in accordance with the fact that the unit-level objective function is the average expected remaining maintenance cost over the life cycle. As an unit approaches the end of its life-cycle, the impact of workload adjustment on improving maintenance cost rate gradually diminishes. Such characteristics would result in the coordinator assigning higher workloads to units expecting a full repair in order to make full use of them.

The above discussion is related to the decision-making rationale behind the basic version of the model. Next the optimisation results from the complete version with different values for the buffer size B will be presented and discussed.

Complete Version - Results and Discussions

The objective of this part of the section is to highlight the role of the last component, which is the long-term penalty-related component in the coordinator objective function as given by equation 5.3. Recall that the long-term penalty-related component attempts to measure the expected production losses caused by too many machines being taken off-line at the same time, and that the buffer size B is used to set the level of conservativeness of the model, where a larger B corresponds to the model being less confident with its predictions for the expected time to the next PM. It also controls how much effort is devoted to minimising the potential production losses in the optimisation process. Based on the above discussion, we can expect to have the following logically-related observations:

1. As B increases, the long-term component will have a larger contribution towards the coordinator objective function.
2. In order to minimise the long-term component, the coordinator will attempt to widen the gap in the expected time to the next PM/RP between the two units.
3. One effective approach to widen such a gap is to have the more degraded unit degrade even faster. This can be achieved by allocating higher workload to the more degraded unit.
4. For a larger B, the optimal load allocation will render a lower system-level risk of production losses.

The rest of the subsection will explore the four observations in more details by analysing the simulation results of a typical replication.

Observation 1

By definition, a greater B implies that the units are not 'considered' operational for a larger number of time intervals. As a result, the same workload allocation will mean a higher potential production losses to a coordinator that is given a larger B. The proportion taken by the long-term component in the total objective function for various buffer sizes for the first optimisation interval in this repetition is presented in Table 5.3. The actual figure increases monotonically from 7.84% for $B = 0$ to 74.05% for $B = 8$.

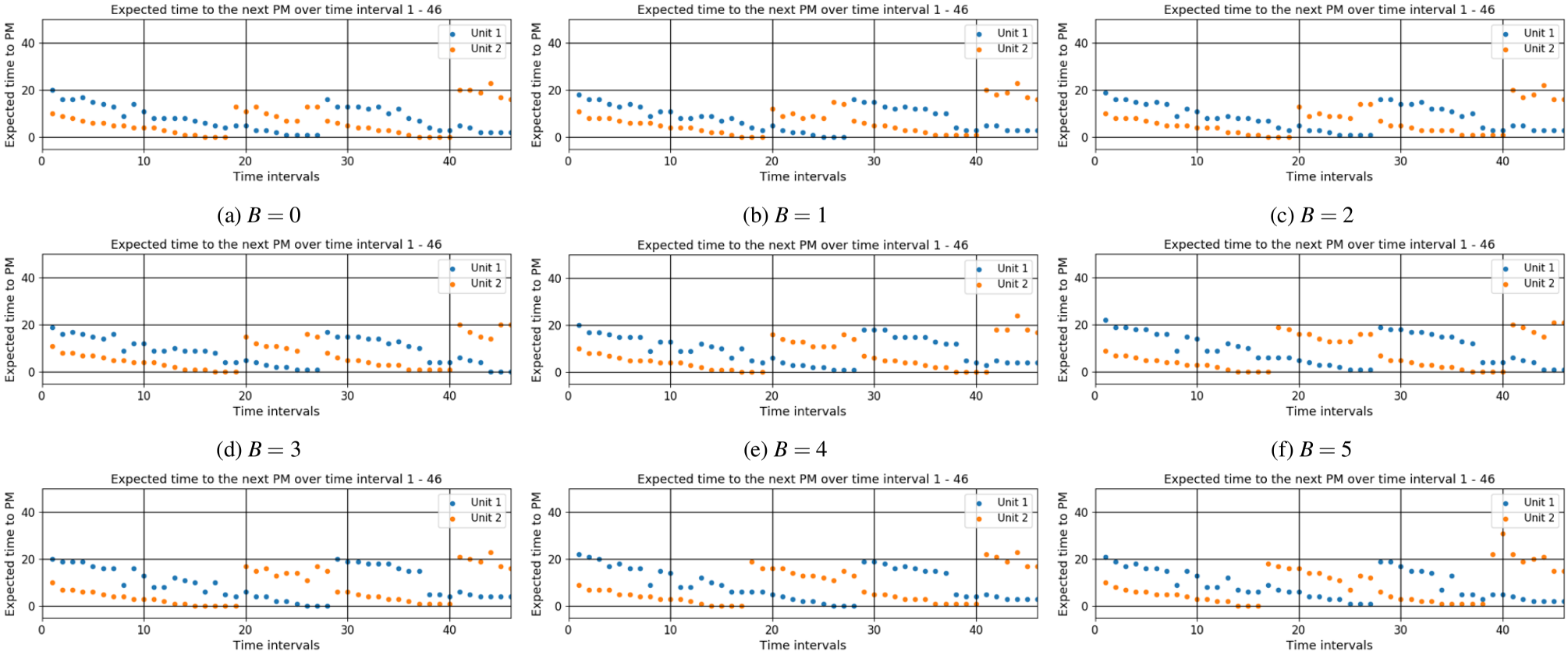
Table 5.3 Proportion of objective function taken by the long-term component

Buffer size B	0	1	2	3	4	5	6	7	8
Long-term %	7.84	11.37	15.78	20.35	26.35	43.52	62.35	65.57	74.05

Observation 2

The expected time to the next PM/RP of both units for various buffer sizes are plotted in Figure 5.10. It could be observed that the gap between the estimated time to the next preventive maintenance of the two units widens as B increases, which is characterised by the distance between a blue dot and an orange dot at the same time interval. For the purpose of a clearer presentation of the results, we split the time horizon considered into periods based on when maintenance tasks are carried out on each of the units in the system. Specifically, the time horizon considered in each scenario here can be roughly divided into three periods. The first period starts from the very beginning of the time horizon to the point when Unit 2 is preventively maintained for the first time; the second period runs till the first PM of unit 1;

the third period ends when a replacement is performed on either unit. A more quantitative comparison of the three periods between different buffer sizes is presented in Figure 5.11. As B increases from 0 to 6, an upward trend of the gap could be spotted for all three periods. However, such trend is reversed from $B = 6$ onwards. When B takes relatively small values, the role of the long-term component is not significant enough in the optimisation process, whereas an excessively large B weakens the impact of workload assignment since both units will not be 'considered' operational for the near future regardless of how the workload is allocated. An indication of such behaviour of the model is that there might exist an optimal B that directs the most appropriate amount of attention from the coordinator to the potential production losses. A closer examination will be conducted on the impact of the buffer size B on the model performance later in Section 5.4.2.



(a) $B = 0$ (b) $B = 1$ (c) $B = 2$

(d) $B = 3$ (e) $B = 4$ (f) $B = 5$

(g) $B = 6$ (h) $B = 7$ (i) $B = 8$

Fig. 5.10 Estimated time to the next PM/RP of various buffer sizes B for a typical repetition

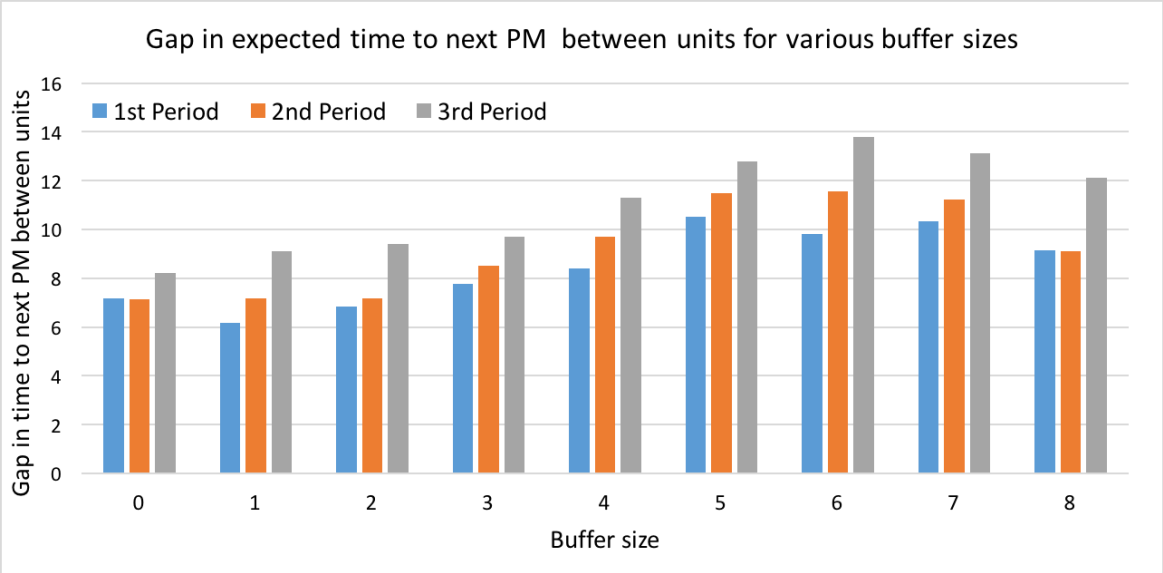


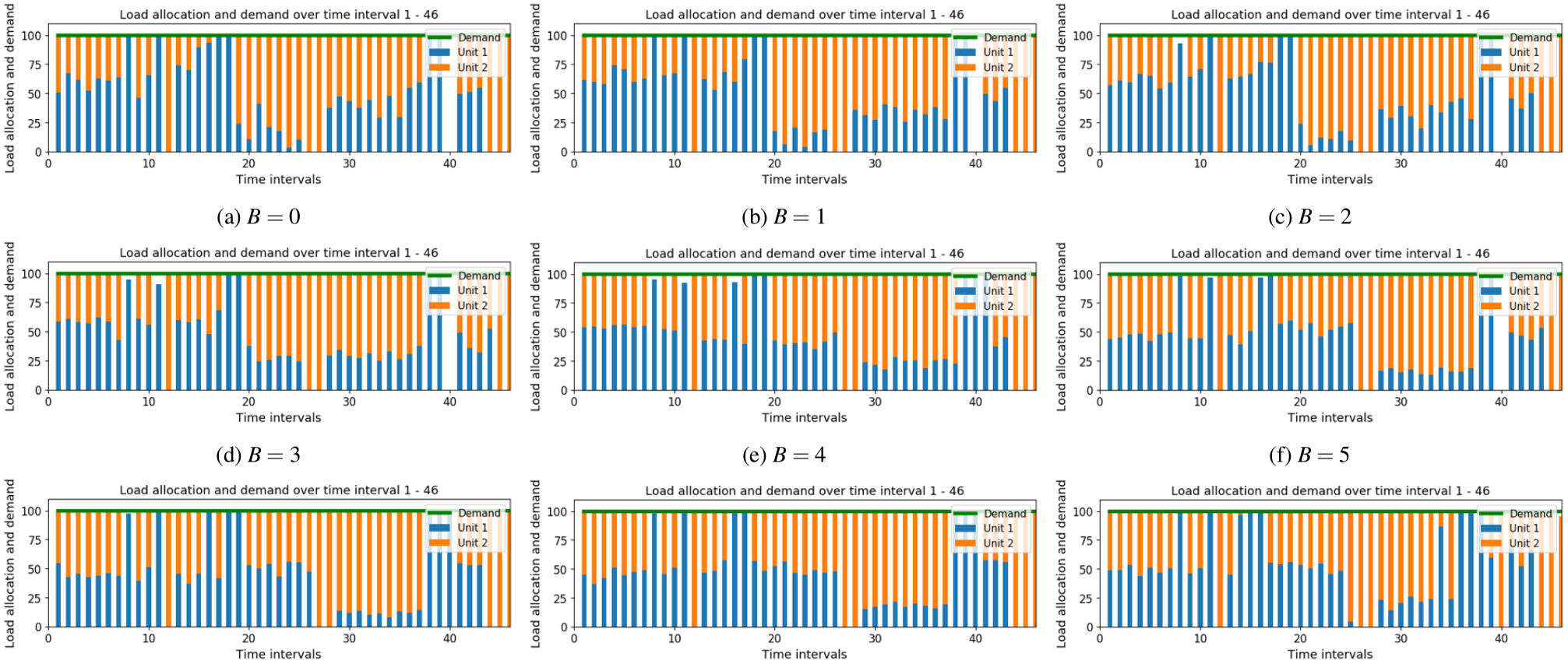
Fig. 5.11 Gap in the expected time to the next PM/RP between units

Observation 3

The workload allocation among the two units for buffer sizes 0-8 are plotted in Figure 5.12. Following the previous approach of dividing the considered time horizon into three time periods, the average workload assigned to Unit 1 and its standard deviation for the resultant periods are presented in Figure 5.13. Combining these two charts, we have the following observations:

- The trend of workload allocation alters among the three time periods: in the first and third periods, the workload allocated to Unit 1 goes down as B increases, whereas for the second time period the trend is completely opposite. Recall that the initial condition of Unit 1 is much better than Unit 2, for the first time period, reducing the workload allocated to Unit 1 implies an even later PM for Unit 1, which meets the requirement imposed by a larger B for a wider time gap between the PMs of two units. After a PM task has been completed on Unit 2 and the system enters the second time period, Unit 2 becomes the healthier of the two units and a larger B would lead to lower workload being assigned to Unit 2. The same argument is also applicable to the third period.

- For all three time periods, the workload allocation becomes more stable as B goes up, as evidenced by decreasing standard deviations measured by the error bars in Figure 5.13. It can then be inferred that apart from shifting the attention of the coordinator between various components in its objective function, another function of the buffer size B is to smooth over unnecessary fluctuations in workload allocation caused by over-confident estimations for the expected PM time slots.



(a) $B = 0$ (b) $B = 1$ (c) $B = 2$

(d) $B = 3$ (e) $B = 4$ (f) $B = 5$

(g) $B = 6$ (h) $B = 7$ (i) $B = 8$

Fig. 5.12 Load allocation of various buffer sizes B for a typical repetition

- The trends of both the workload allocation and variance in workload allocation are reversed when B surpasses 5 or 6. This again is an indication of the diminishing effects of B taking excessively large values.

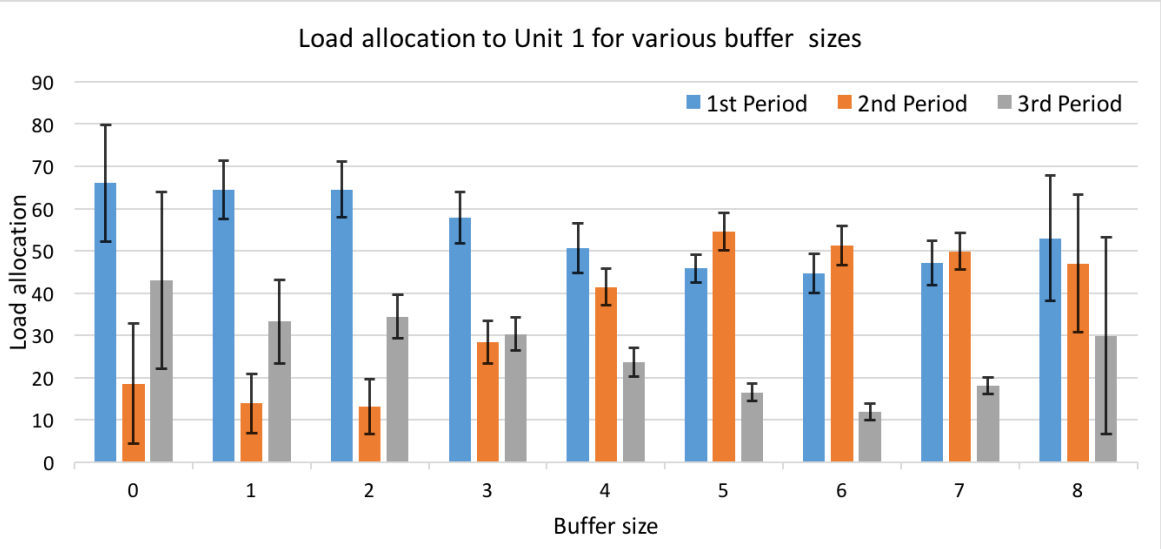


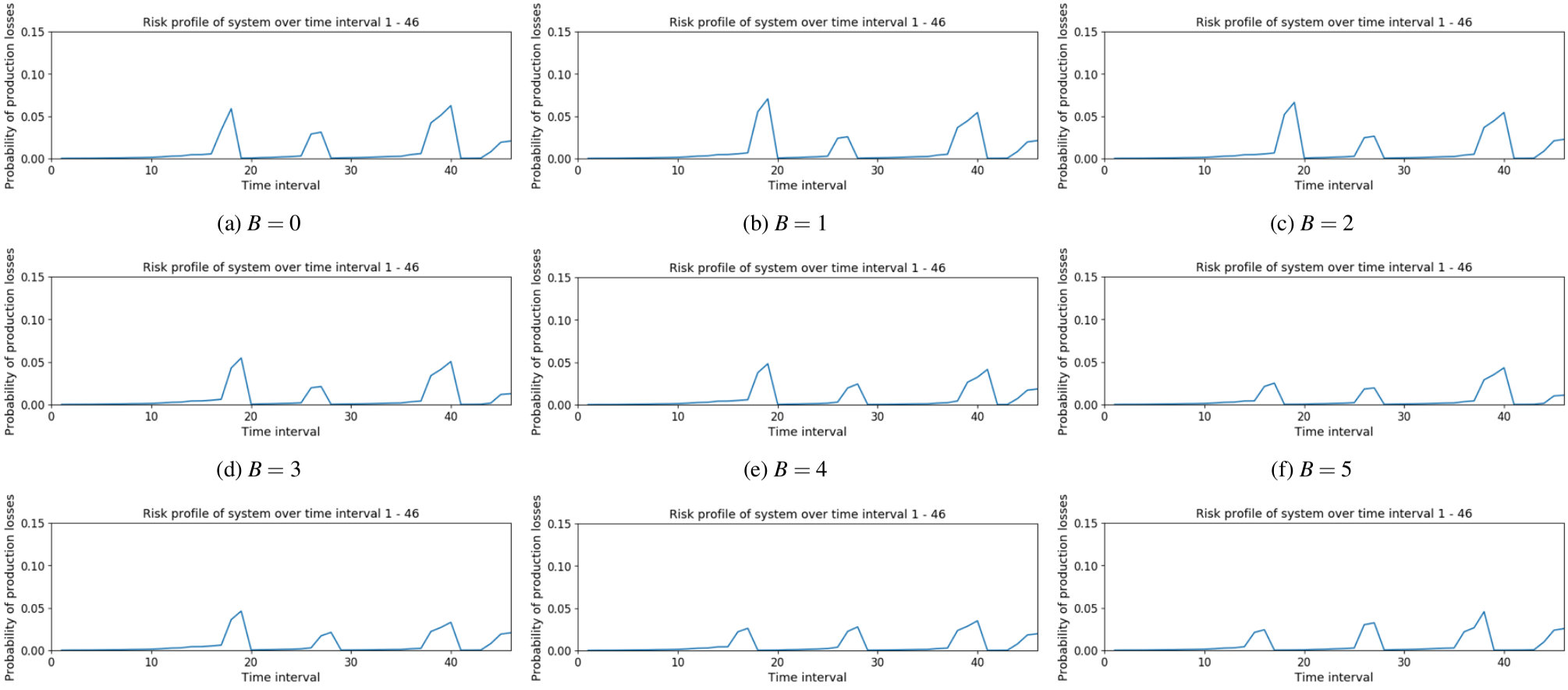
Fig. 5.13 Workload allocation to Unit 1 under various buffer sizes

Observation 4

Although in reality both machines may be in an operational state for a specific time interval, it will appear to the coordinator that one of them is taken off-line for maintenance due to the existence of buffers B, and the probability of production losses occurring over that interval is perceived to be higher by the coordinator. By assigning different values to B, the optimal workload allocation under each scenario is obtained. Using the obtained optimal workload allocation, the risk profile, which is the actual probability of production losses for the considered time horizon, can be found in Figure 5.14. It can be inferred from the graph that increasing the buffer size B has a mitigating effect on the probability of production losses.

5.4.2 Discussion

The aim of this subsection is to provide a short summary of the important findings from the above analysis. As up till now the discussion of model behaviour is centred around a single replication, this subsection also acts as a source of forecast for the model performance measured based on results from multiple replications of the simulation.



(a) $B = 0$ (b) $B = 1$ (c) $B = 2$

(d) $B = 3$ (e) $B = 4$ (f) $B = 5$

(g) $B = 6$ (h) $B = 7$ (i) $B = 8$

Fig. 5.14 Risk profile of various buffer sizes B for a typical repetition

1. The coordinator-level objective function essentially consists of three parts: a) instant production losses; b) marginal maintenance costs; and c) potential future production losses (long-term penalty-related component). The first part is only sensitive to the total workload of all units and not affected by how the workload is assigned among units. The other two elements are largely affected by workload allocation, and generally move in opposite directions. To be specific, in the numerical example, maintenance costs tend to increase with more workload being allocated to the more degraded unit (Unit 2), while the potential future production losses tend to decline. The aim of the coordinator is then, to reach the optimal trade-off between the three components.

2. There are in fact two types of workload allocation in the model. The explicit one is the decision made by the coordinator at each time epoch that designates each unit to take on a certain amount of workload. The implicit one is that even for the same unit, workload is being shifted along the timeline, which is a result of the explicit workload allocation. For instance, in Figure 5.13, after being heavily loaded for the first time period, Unit 1 tends to undertake lower workload in the second time period.

3. Apart from parameters quantifying the physical characteristics of the system, the buffer size B also has an important role in the trade-off process. The three major functions of B are: a) to take into account the uncertainty associated with the predicted time to the next PM/RP; b) to adjust the weight of the long-term component in the coordinator objective function so that it more closely characterises the consequences of loss of production; c) to reduce the unnecessary fluctuations in the optimal solution for workload allocation resulted from the stochastic degradation process of units. We can therefore expect that the model performance would be influenced by the choice of B and that a different B might be required to achieve the best result if changes have been made to other parameters in the problem setting. A systematic approach to choosing an appropriate B is for now out of the scope of this research project. . However, two potential ways of choosing B will be proposed in Chapter 6 when the proposed approach is applied in the context of a real case example in the oil and gas industry. The choice of B will also be discussed in Section 7.6.1.

5.5 Impact of Decentralised Approach on Efficiency

As mentioned in Section 4.2, while it is still feasible to place all computational burdens on only one entity, the applicability is restricted to cases with very few parallel assets or when there is no time constraint. An approach that decentralises the solution-seeking task is thus

preferred in this research project where flexibility and high efficiency is emphasised. This section aims to provide a quantitative description of how the computational time on one entity changes as the number of parallel assets gets larger, which in a way gives clues on the time-saving effect of a decentralised approach.

The device used here is a MacBook Pro released in late 2013, and all the programmes are written in Python with part of the codes directed to lower-level C functions to provide a speedup in the calculation of the integrals. Detailed specifications of the device is given in Table 5.4. Due to the availability of multiple cores, parallel computing is used wherever possible in the programme. In Section 5.2.2, the decision-making process is divided into three steps: 1) first-round machine-level calculation and data preparation; 2) coordinator-level workload allocation; 3) second-round machine-level calculation.

In the analysis, the run-time for the three steps is recorded separately to give a complete map of how the computational efforts are distributed.

Table 5.4 Specifications of the device used for computation

Item	Processor	Memory	No. of Physical Cores
Specification	2.4 GHz Intel i5	8 GB 1600 MHz DDR3	4

We start with $M = 2$ and gradually increase the number of parallel units by 2 each time until $M = 16$. Note that here all calculation is actually performed by one entity. If each machine was indeed represented by a machine-agent on a separate computer, the run-time for the first and third steps would be roughly $\frac{1}{M}$ of what is shown in Table 5.5, whereas the computational task involved in the second step cannot be distributed. The proposed approach is therefore preferred if the first and third step take longer to complete than the second step.

The following insights can be drawn progressively from Table 5.5:

1. Despite small fluctuations, both the total run-time as well as the time taken to complete each of the optimisation steps increase with the number of parallel units M.

2. Initially the first step, whose calculation is performed at the unit level, makes the biggest contributor to the total run-time. However, it gradually gives way to the coordinator optimisation ($0.17\% \rightarrow 59.60\%$) step as M increases from 2 to 14. Such trend continues as the computational time needed for the second step increases at a faster rate than the first step.

3. The most computation-intensive part in the machine-level optimisation is various integrals used to obtain the expected life cycle and cost components related to each

machine. As a result, a roughly linear relationship exists between the time taken by the first step and the number of machines M. This also applies to the third step, where the machines recalculate the optimal maintenance solutions after being assigned their own workload.

4. In the second step, a significant amount of time is devoted to the calculation of the long-term penalty-related component, which involves a separate consideration for every possible combination of machine statuses in the near future. As the number of combinations is a factorial function of M, the calculation is bound to take substantially longer time for a larger M.

Table 5.5 Computational time (in seconds) needed for M from 2 to 16

M	2	4	6	8	10	12	14	16
First step	81.27	93.01	161.79	149.79	233.05	315.92	327.45	232.15
Second step	0.16	1.09	1.54	5.45	21.50	119.34	573.37	3049.45
Third step	16.48	17.29	28.09	28.77	39.93	49.73	61.24	43.11
Total time	97.91	111.40	191.43	184.01	294.48	484.99	962.06	3324.71
$\frac{\text{Second step}}{\text{Total time}}$	0.17%	0.98%	0.81%	2.96%	7.30%	24.61%	59.60%	91.72%

Though it may seem that the decentralised approach will surely lose its edge when M is big enough, the range within which it is noticeably more efficient, however, is affected by various factors. Generally speaking, any factor that lengthens the time taken by the first and last steps or shortens that taken by the second step tends to widen such range. To be specific, the run-time of the first step is positively correlated with the length of the life-cycle of machines. A longer life-cycle leads to a wider integral interval, whose calculation is more time-consuming. Apart from M, one determining factor of the run-time needed at the coordinator level is the least number of machines required to fulfil customer demand, as this controls the scale of machine status combinations that will lead to production losses.

In summary, adopting the proposed decentralised approach leads to time-saving effects compared with centralised approaches that place all computational burden on one entity. The effects, however, tend to diminish as the number of parallel machines becomes too large due to the run-time of the second step being more sensitive to M. The diminishing rate of the time-saving effects is affected by multiple factors, such as the length of machine life-cycle and the degree of redundancy of the system. Though the actual impact of the decentralised approach on computational efficiency can only be determined on a case-by-case basis, it can be considered quite significant for a reasonable range of M.

5.6 Performance Comparison with Traditional Strategies

In this section, the performance of the decision-making approach developed in this research project is compared with some of the traditional approaches currently adopted by the industry or in the literature while they are trying to make a decision on workload allocation. The following three strategies are the most commonly mentioned:

1. **Uniform workload allocation:** uniform allocation ensures that workload is evenly split among assets in the fleet. This is very typical in load-sharing systems presented in academic literature. Recall in Section 3.2, Yu and Chan [105] studied the the load sharing mechanism in the case of multiple chillers in a HVAC system, where each chiller gets an equal volume of water flow. Keizer et al. [56] discussed the phenomenon of uniform workload allocation observed in a pumping facility.

2. **Random workload allocation:** in this case, there is no well-defined set of rules that governs how workload should be allocated. Instead, the decision is made on an ad-hoc basis and often influenced by unexpected random factors. For instance, in the exploratory case study described in Section 3.3.2, the operator chooses not to use certain vessels due to inconvenient bash-wash procedures or broken valves, neither of which affects the capacity the vessels can achieve.

3. **More-on-new-machine workload allocation:** in general, when this approach is adopted, more workload is assigned to newer machines and less to older ones. Being intuitively correct, this approach has been adopted both in the academic literature and by practitioners in the industry. For instance, in the task-dependent condition-based maintenance model proposed by Celen and Djurdjanovic [13], it is assumed that the least-degraded chamber always gets the most demanding task. Moreover, it has been mentioned in the Rapid Gravity Filter (RGF) case provided by Ye et al. [102] that the aim of the operator is to balance the degradation level of the RGFs by choosing appropriately which ones to be turned on, indicating more frequent usage of the newer ones.

In order to isolate and highlight the advantage of the proposed workload assignment model at the coordinator level, the same machine-level load-dependent maintenance optimisation model is also adopted for the three traditional workload allocation strategies. Note that while separate experiments have been run for both random and uniform workload allocation strategies, we did not explicitly build a case for the more-on-new-machine strategy mainly because of imprecise introduction on how such a strategy should be implemented. However, the basic version of the model as mentioned in Section 5.4.1, whose objective function

excludes the long-term penalty-related component, can be perceived as an exemplar of the more-on-new-machine workload allocation strategy. In the basic version, the optimal solution is largely determined by the total maintenance cost rate, which tends to increase with workload being shifted from newer machines to older ones. Consequently, the basic version of the model prefers to allocate more workload to newer machines in order to minimise total maintenance cost rate.

5.6.1 Experiment Settings

Similar to the numerical examples presented in Section 5.4.1, a two-unit system is considered here. The model parameters given in Table 5.1 are also adopted in this section. The flowchart of the uniform and random workload allocation process can be found in Figure 5.15. Note that no information on penalty costs is needed in these two strategies and that no optimisation task is performed at the coordinator level.

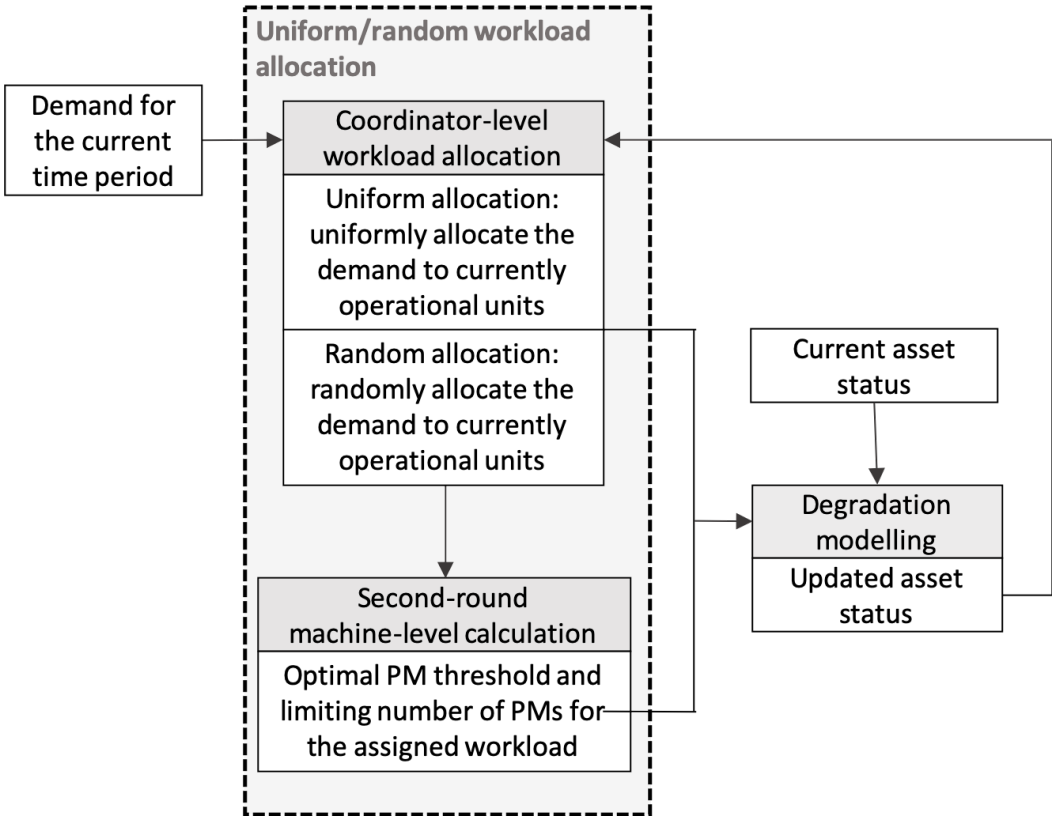


Fig. 5.15 Flowchart of the uniform/random workload allocation strategy

In order to have a thorough evaluation of the proposed model, the complete version with buffer size ranging from 0-9 have been benchmarked against the basic version of the

model and the other two traditional strategies. We have run 150 replications of simulation for each case, and each replication consists of 100 time intervals. It will be shown later that 150 replications is sufficient to largely eliminate from the results fluctuations caused by the stochastic nature of the production system considered. The time horizon, based on which the model performance has been measured, is chosen to be 65 intervals, since all units will have been replaced by this point in time in every possible scenario.

5.6.2 Results and Discussions

Due to the stochastic nature of the system, it takes a large number of replications to reach an absolutely steady results. For the purpose of this study, which is to develop a decision-making model and evaluate its performance with existing strategies, the relative relationship between various cases is of greater concern. As a result, the number of replications is considered sufficient when the relative relationship, especially that between the proposed model and the traditional strategies, has stabilised. The accumulative average maintenance cost rate across replications from the 50^{th} replications on-wards for the three traditional strategies and a representative group of B values is given in Figure 5.16. It can be seen that the relative relationship hardly changes after the 110^{th} replication. In the end, 150 replications have been run for each case to obtain the final results.

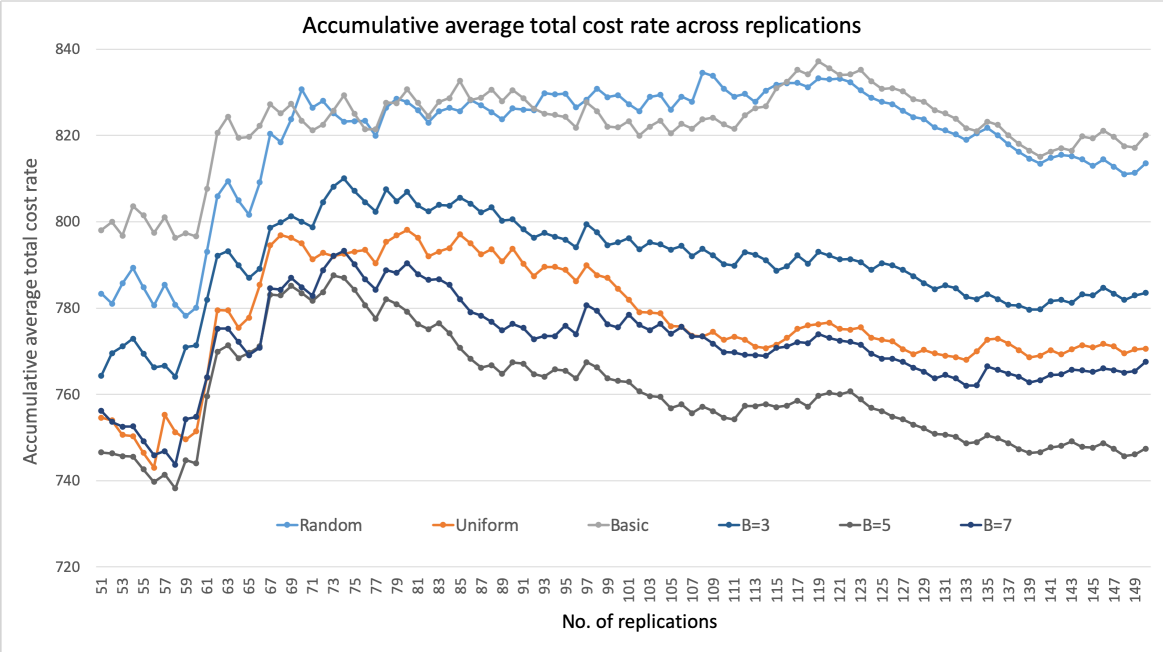


Fig. 5.16 Accumulative average total cost rate across replications

Figure 5.17 shows the average total cost rate from the 1^{st} to the 65^{th} time interval for various strategies. A couple of interesting observations can be made from the plot:

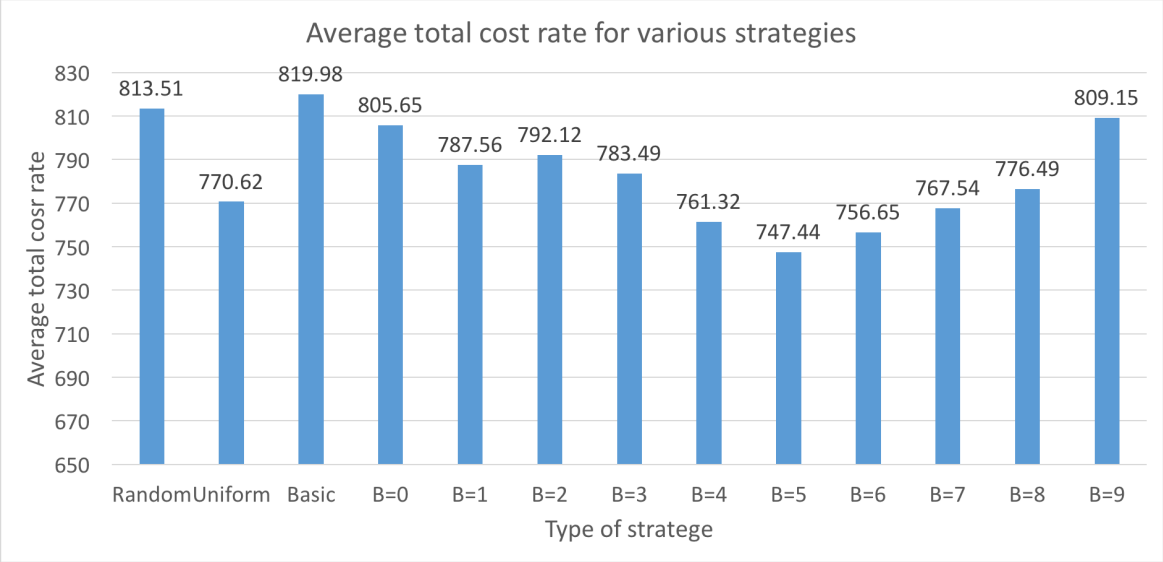


Fig. 5.17 Comparison of the average total cost rate for various strategies

1. The basic version ends up with the worst performance amongst all strategies, rendering a total cost rate of 819.98, which is even higher than that of random workload allocation. Recall that the basic version of the model deliberately considers only the effect of a certain workload allocation on maintenance cost rate, such under-performance should be expected due to the partial and myopic view of the coordinator in this case.

2. The complete version of the model with a buffer size $B = 5$ gives the lowest average total cost rate of 747.44, which is 8.12%, 3.00%, and 8.85% lower than random, uniform, and the basic version workload allocation, respectively. This is in accordance with a previous suspicion that an optimal B might exist for a specific problem setting. Though our experiment stops as $B = 9$, it could be deduced from the definition of buffer size that increasing B will not improve the situation any further. When B is large enough, no units will be 'considered' operational for every time interval in the near future. This would result in the coordinator being indifferent to the possible ways of allocating workload when it comes to potential future production losses. Consequently, decisions will be made based solely on the magnitude of the maintenance-related component - exactly what the basic version does.

3. Though the proposed complete version of the decision-making model outperforms the random allocation strategy for all values of B, it only beats the uniform allocation

strategy for B between 4 and 7. This could possibly be attributed to the following two factors:

(a) In Section 5.4.1, we discussed the stabilising effect of B on the optimal workload allocation. Improvement in model performance is seen with less fluctuations in workload allocation. Such stabilising effect is intrinsic in the uniform allocation strategy while the demand is constant. This has led to the under-performance of the proposed model while the stabilising effect provided by B is not sufficient to match that of uniform allocation. Due to the time constraint of this study, we will stop exploring the details of this aspect of the model behaviour. Directions on how further work can be conducted on this matter, however, will be given in Chapter 7 which concerns future research.

(b) Uniform workload allocation is in fact not very far off from the actual optimal solution under the parameter setting of this particular numerical example. As discussed earlier, one approach to mitigate the possibility of production losses is to separate in time the need for maintenance of all units in the system. In this numerical example, as there is a significant gap between the initial degradation level of the two units (0 vs 40), uniform workload allocation tends to retain such difference and thus naturally helps to avoid clashes of unit breakdowns.

A further implication of factor b) is that the proposed strategy will outperform the traditional strategies by a larger margin if initially the units are in similar condition, where uniform workload allocation exacerbates risks of production losses. This scenario will be further explored in the model sensitivity analysis in Section 5.7.

With the random allocation being taken as the basis for comparison, the difference between the other strategies and the random allocation is plotted in Figure 5.18 for each of the four cost rate components: preventive maintenance (PR), replacement (RP), minimal repairs (MR), and penalty for loss of production. Note that while saving in penalty costs is almost equally important to the total cost saving as that in maintenance-related components in the case of uniform allocation, it has a much more dominant role in the complete version of the model. It can be seen from Table 5.6 that for $B = 0$ and $B = 9$, the saving in penalty costs even surpasses that of the total costs.

The degradation levels of the two units as well as their average at the end of the 65^{th} time interval can be found in Table 5.7. Still using the random allocation as the basis for comparison, in Figure 5.19 we have plotted the difference between the other strategies and the random allocation for the sum of maintenance-related cost rate as well as the average degradation level of the two units at the end of the considered time horizon. Not surprisingly,

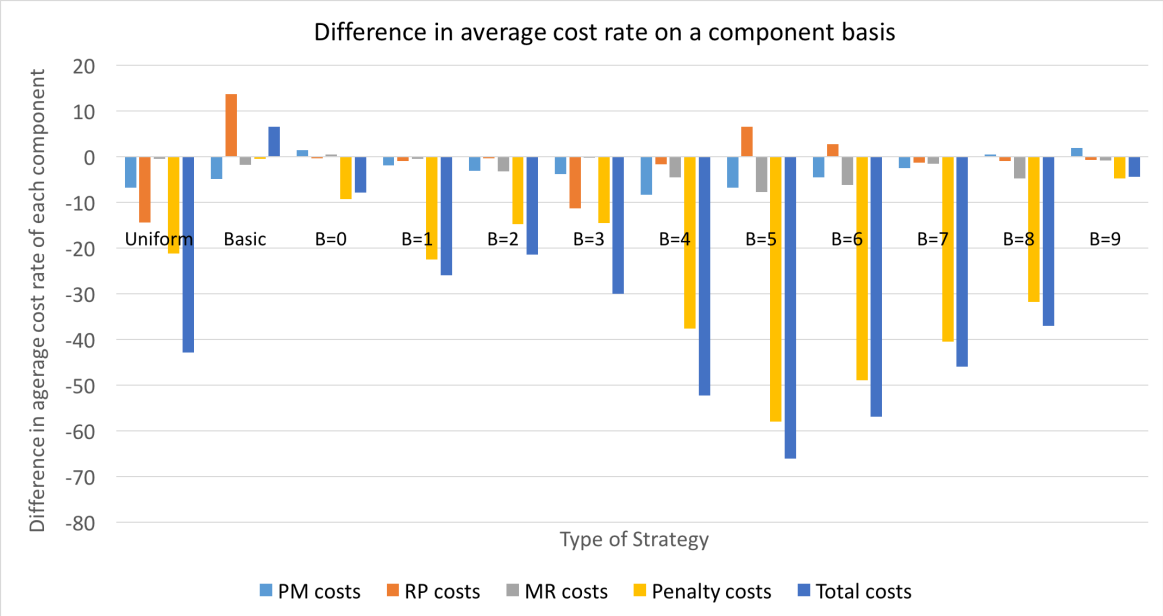


Fig. 5.18 Difference in the average cost rate for various strategies on a component basis

Table 5.6 Proportion of total costs saving from saving in penalty costs

Case	Uniform	Complete version - buffer size									
		0	1	2	3	4	5	6	7	8	9
Percentage %	49	119	87	69	49	72	88	86	88	86	110

opposite trends can be observed between the two lines, implying that while a strategy incurs higher maintenance costs, it results in healthier asset condition. That being said, the difference in average asset condition at the end of the 65^{th} time interval is rather insignificant, with a low of 44.13 and a high of 48.18. This has again, however, confirmed a previous finding discussed in Section 5.4.2: for the entire system, workload allocation happens between different assets; for a single unit, workload is being shifted along the timeline. It is worth mentioning that the results obtained in this section are based on two units that are mostly identical except for their initial condition, it might not be the same case if the system consists of units with heterogeneous degradation behaviours. For instance, if Unit 1 is much more sensitive to workload (e.g., a larger s in equation 4.2), the proposed model is likely to allocate much more workload to Unit 2, which deteriorates less than Unit 1 for the same amount of output. Under such parameter setting, a significantly lower average unit degradation level at the end of the considered time horizon can be expected from the proposed strategy.

Table 5.7 Unit degradation level at the end of the considered time horizon

Case	Random	Uniform	Basic	Complete version - buffer size									
				0	1	2	3	4	5	6	7	8	9
Unit 1	45.2	44.5	43.1	43.1	45.8	44.1	46.5	47.2	42.2	42.9	43.4	43.4	42.6
Unit 2	49.9	47.6	45.2	47.5	48.9	49.6	49.9	48.4	46.3	47.4	48.3	48.5	45.7
Average	47.6	46.1	44.1	45.3	47.4	46.8	48.2	47.8	44.2	45.1	45.8	45.9	44.2

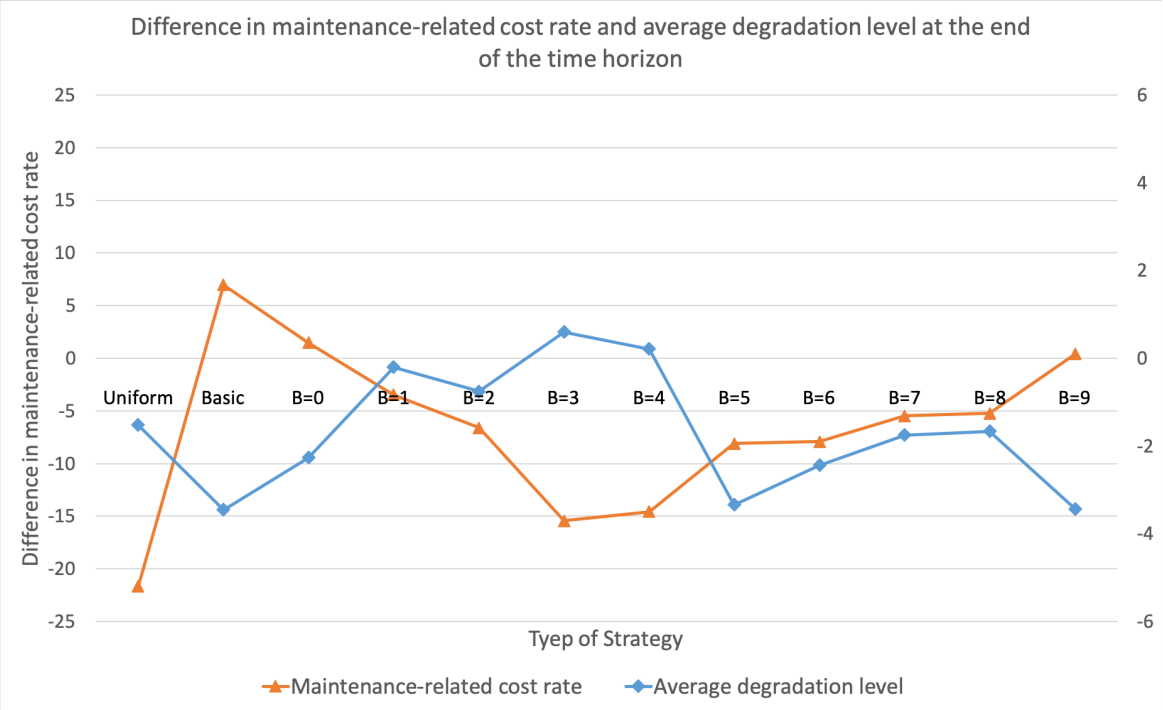


Fig. 5.19 Difference in the maintenance-related cost rate and the average asset degradation level at the end of the 65^{th} interval for various strategies

In the next section, sensitivity analysis will be conducted on how the model behaviour changes with two important sets of parameters - the initial degradation level of units X_0^m, and the penalty cost for a unit of production loss q.

5.7 Sensitivity Analysis of Model Parameters

In this section, sensitivity analysis is carried out to gain deeper insights into the behaviour and working mechanism of the proposed decision-making model. The main objective here is to advance the understanding of the applicability and limitations of the model.

All the scenarios to be considered for the sensitivity analysis are based on a two-unit system, and use largely the same parameters as given in Table 5.1 except for one variation in each scenario. The details of the variations are given in Table 5.8. These scenarios are also numbered for ease of reference.

Table 5.8 Description of the scenarios considered in sensitivity analysis

Name of scenario	Parameters	-
Base Scenario	Table 5.1	-
Name of Scenario	**Difference from Base Scenario**	**Variation**
Scenario 1	Initial degradation level of units X_0^m	$[0,40] \rightarrow [0,0]$
Scenario 2	Penalty cost for a unit loss of production q	$90 \rightarrow 60$
Scenario 3	Penalty cost for a unit loss of production q	$90 \rightarrow 30$
Scenario 4	Penalty cost for a unit loss of production q	$90 \rightarrow 10$

5.7.1 Initial Condition of Units

In the base scenario, Unit 1 and 2 start with a degradation level of 0 and 40, respectively. For the sensitivity analysis in this subsection, Scenario 1 is adopted - the experiments have been conducted on an initial degradation level of 0 for both units. Since it is possible for a fleet of assets to be in any condition when a maintenance/operation strategy is adopted, it is important to understand the model performance for assets with various initial degradation levels.

The total cost rate for a selection of strategies in this case is shown in Figure 5.20. Similar to the experiment results presented earlier in Section 5.6.2, the best performance is given by the complete version of the model. The lowest average total cost rate, 687,71, is observed when $B = 5$, which is 9.16% and 4.52% lower than the random and uniform allocation strategy, respectively. This is a more substantial improvement compared with 8.12% and 3.00% achieved in the base scenario.

Using random workload allocation as basis for comparison, the difference in the average cost rate on a component basis for a selection of strategies is plotted in Figure 5.21. As is the case with the base scenario, saving in penalty costs clearly takes a dominant role in reducing total costs.

Recall in Section 5.6.2 an inference has been made that the proposed model is likely to outperform the uniform allocation strategy by a larger margin if units start in more

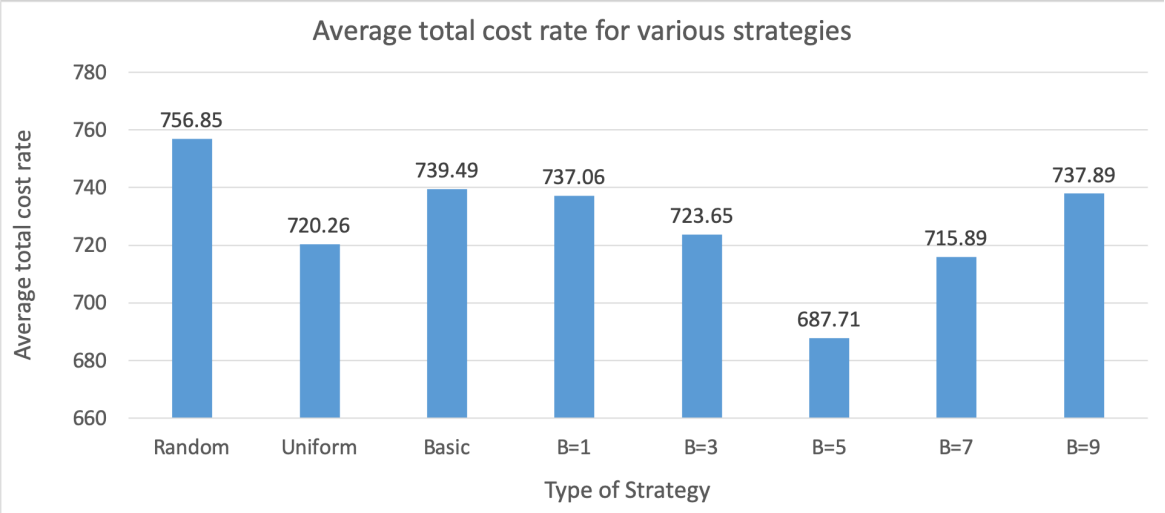


Fig. 5.20 Comparison of the average total cost rate in initial condition sensitivity analysis

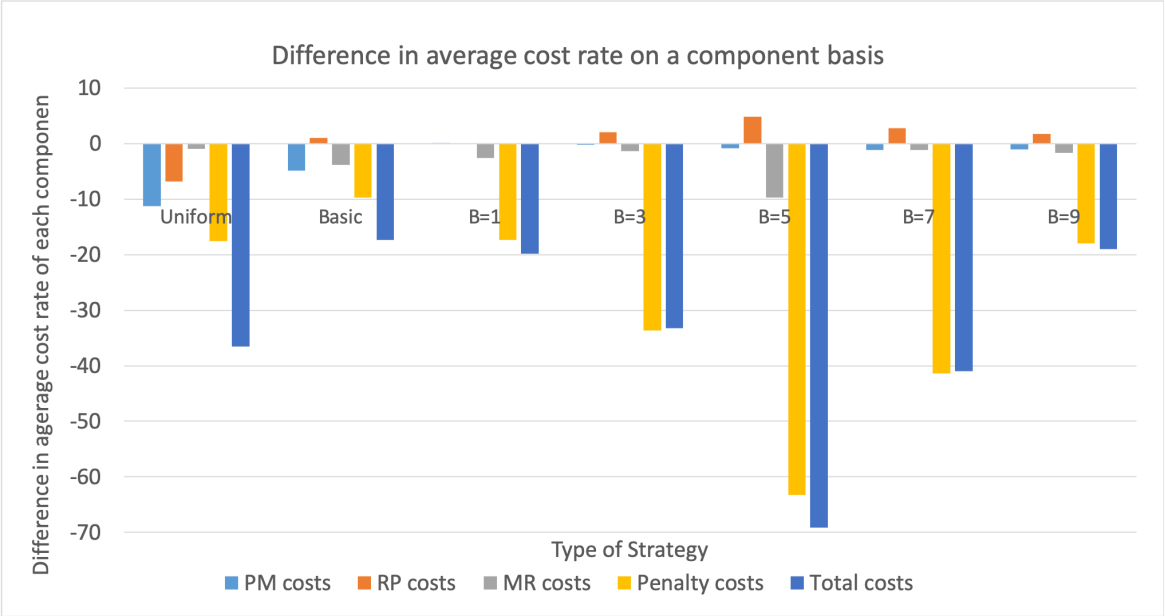


Fig. 5.21 Difference in the average cost rate on a component basis in initial condition sensitivity analysis

homogeneous condition, where uniform allocation tends to exacerbate loss of production incurred by simultaneous machine breakdowns. Here we attempt to verity this inference by benchmarking the performance of the proposed model against that of uniform allocation. Figure 5.22 shows how the proposed decision-making model differs from the uniform allocation strategy in total cost rate in both absolute and percentage terms. In five out of six cases presented, the proposed model either beats the uniform allocation strategy by a

wider margin or underperforms by a smaller margin in Scenario 1 than that in the base scenario in both absolute and percentage terms. A further decomposition of the total cost rate in Figure 5.23 reveals that compared with the base scenario, in Scenario 1, the proposed model has demonstrated a stronger tendency to generate less penalty costs than the uniform allocation strategy.

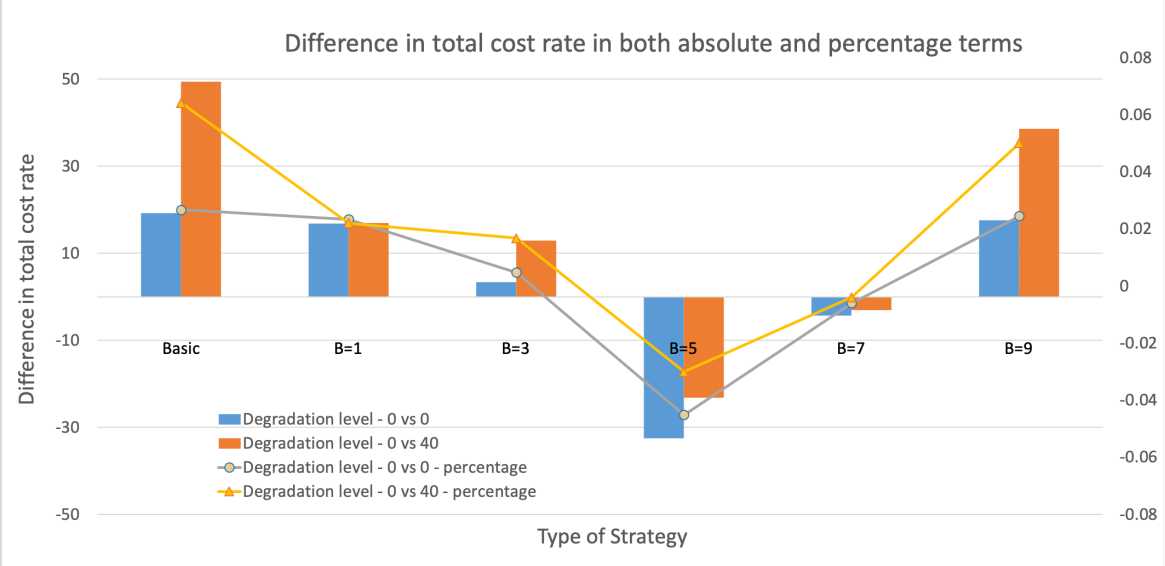


Fig. 5.22 Difference in total cost rate on in initial condition sensitivity analysis - benchmarked against uniform allocation

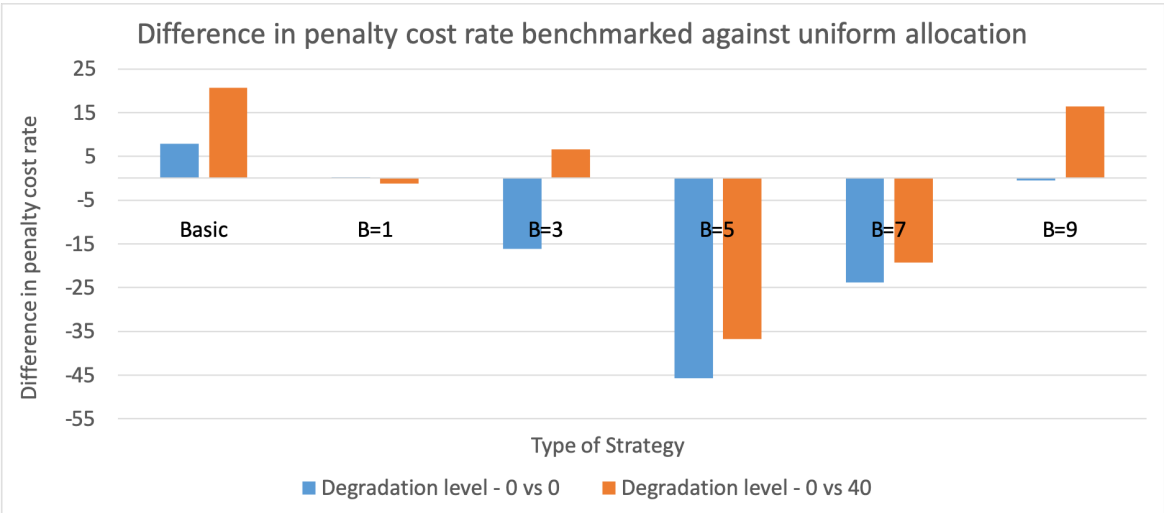


Fig. 5.23 Difference in penalty cost rate in initial condition sensitivity analysis - benchmarked against uniform allocation

The above analysis shows that the proposed model is able to generate lower total costs than those rendered by traditional strategies with an appropriate choice of buffer size B in both scenarios with homogeneous and heterogeneous initial asset condition.

5.7.2 Penalty Cost for Production Losses

Three different values have been used as the penalty cost for a unit of production loss - 60, 30, and 10. To some extent, the magnitude of penalty cost can be seen as characteristics of different industries. For instance, a higher penalty cost might correspond to the process industry which has much less tolerance for a drop in system availability than other industries. The results of the sensitivity analysis, therefore, has practical implication on whether the proposed decision-making model is an attractive option to plant owners in specific industries. As the proposed model and the traditional strategies are all based on the same machine-level maintenance optimisation model, a large part of the potential benefits of the proposed model is more likely to be realised in the form of reduction in production losses. As a result, it is expected that the advantage of the proposed model over traditional strategies will diminish as loss of production gets cheaper.

Table 5.9 summarises the results obtained for the scenarios considered in this subsection. In each of the scenario, five rows of data are presented. The first row gives the total cost rates for different types of strategy. The second and fourth rows are the absolute difference in total cost rate compared with random and uniform allocation, respectively, and the third and fifth rows are the difference in percentage terms. For each scenario, the best performance achieved by the proposed model is highlighted in red. As q shrinks from 90 in the base scenario to 10 in Scenario 4, the change in total cost rate also declines gradually from a reduction of 8.12% to 1.68% while compared with random allocation, and from a reduction of 3.01% to an increase of 2.12% while compared with uniform allocation. It can be very tempting to attribute the performance deterioration to fewer costs being saved with a smaller q even if the same amount of production loss has been avoided. If such suspicion were valid, we would expect to see the proposed strategy leading to a penalty cost rate lower than that given by uniform allocation regardless the scale of q. Further inspection into the experiment results, however, points to something different. The difference in penalty cost rate benchmarked against uniform allocation for all scenarios is plotted in Figure 5.24. It can be observed that in Scenario 4 when $q = 10$, the proposed strategy fails to come up with a lower penalty cost rate in any case. This is an indication that the scale of penalty costs has other impacts on the decision-making model. While q is large, the coordinator is motivated to direct more attention to ensuring that the numerical value of the long-term component stays low. Since small shifts of workload might result in significant changes in the long-term component,

Table 5.9 Summary of results for sensitivity analysis of penalty cost for production losses

Name	Type of strategy												
	Rnd.	Uni.	Basic	Complete version - buffer size B									
				0	1	2	3	4	5	6	7	8	9
Base	813.5	770.6	820.0	805.6	787.6	792.1	783.5	761.3	747.4	756.7	767.5	776.5	809.1
	Com. Rnd.		6.5	-7.9	-26.0	-21.4	-30.0	-52.2	-66.1	-56.9	-46.0	-37.0	-4.4
	Com. Rnd. %		0.80	-0.97	-3.19	-2.63	-3.69	-6.42	-8.12	-6.99	-5.65	-4.55	-0.54
	Com. Uni.		49.4	35.0	16.9	21.5	12.9	-9.3	-23.2	-14.0	-3.1	5.9	38.5
	Com. Uni. %		6.41	4.55	2.20	2.79	1.67	-1.21	-3.01	-1.81	-0.40	0.76	5.00
Sce. 2	748.8	718.4	757.6	734.6	739.0	742.6	728.9	707.3	707.9	713.0	716.9	727.8	745.6
	Com. Rnd.		8.8	-14.2	-9.8	-6.3	-19.9	-41.6	-40.9	-35.8	-31.9	-21.0	-3.2
	Com. Rnd. %		1.18	-1.90	-1.31	-0.83	-2.66	-5.55	-5.46	-4.78	-4.26	-2.81	-0.43
	Com. Uni.		39.2	16.2	20.7	24.2	10.5	-11.1	-10.5	-5.4	-1.4	9.4	27.3
	Com. Uni. %		5.46	2.26	2.88	3.37	1.47	-1.55	-1.46	-0.75	-0.20	1.31	3.79
Sce. 3	686.7	658.0	702.7	679.6	676.1	673.2	671.9	663.2	666.9	670.0	673.1	673.3	683.7
	Com. Rnd.		15.9	-7.2	-10.6	-13.6	-14.9	-23.5	-19.8	-16.7	-13.6	-13.4	-3.0
	Com. Rnd. %		2.32	-1.04	-1.54	-1.98	-2.17	-3.43	-2.88	-2.44	-1.98	-1.96	-0.44
	Com. Uni.		44.7	21.6	18.1	15.2	13.9	5.2	8.9	12.0	15.1	15.3	25.7
	Com. Uni. %		6.79	3.28	2.76	2.30	2.11	0.79	1.36	1.82	2.30	2.33	3.91
Sce. 4	644.5	620.5	693.5	664.5	645.8	648.8	640.7	634.8	637.5	633.6	643.0	645.1	649.1
	Com. Rnd.		49.1	20.0	1.3	4.3	-3.8	-9.7	-7.0	-10.8	-1.4	0.6	4.6
	Com. Rnd. %		7.61	3.11	0.20	0.66	-0.59	-1.50	-1.08	-1.68	-0.22	0.09	0.71
	Com. Uni.		73.1	44.1	25.3	28.3	20.2	14.4	17.0	13.2	22.6	24.6	28.6
	Com. Uni. %		11.78	7.10	4.08	4.56	3.26	2.32	2.75	2.12	3.64	3.96	4.61

a relatively stable workload is maintained at consecutive decision-making intervals. As discussed in Section 5.6.2, a reduction in the level of fluctuations in workload allocation leads to improvement in model performance. This stabilising effect, however, is lost with a very small q. As the fluctuations in the maintenance cost component and the penalty cost component in the coordinator objective function are of roughly the same magnitude, they tend to cancel out each other while workload is being shifted. Consequently, along the timeline being considered, more variation exists in the workload allocation solution provided

by the proposed strategy given a smaller q. Data from one replication is given as an example: the approach used in Section 5.4.1 is also adopted here to divide the considered time horizon into three periods, the average and standard deviation of workload allocated to Unit 2 for each period is:

Base scenario: [53.4, 3.4], [83.2, 1.89], [81.7, 1.7];

Scenario 4: [51.4, 4.0], [69.5, 3.85], [80.9, 2.5].

It is clear that for all three periods, Unit 2, which is the older of the two units, takes on higher workload in the base scenario. Even so, Scenario 4 has larger standard deviation in its allocation in absolute terms.

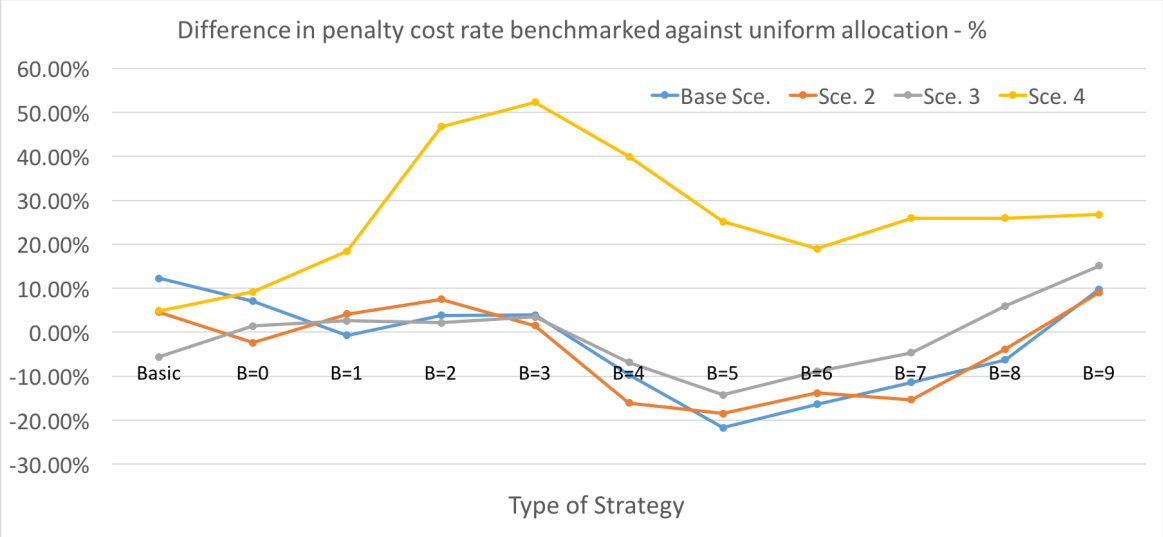


Fig. 5.24 Difference in penalty cost rate in penalty cost sensitivity analysis - benchmarked against uniform allocation

5.7.3 Cross Comparison

The previous two subsections have respectively studied the impact of initial asset condition and penalty cost for a unit of production loss on the performance of the model. In order to 1) gain deep insights into the joint effects of the two factors; 2) further validate the findings of the previous two subsections, this subsection studies the model performance under different combinations of q and X_0^m. Apart from the five scenarios defined in Table 5.8, experiments have been run for three more sets of parameters. Similarly, only their difference from base scenario is presented:

Scenario 5 initial condition X_0^m: [0, 0]; penalty cost for a unit of production loss $q = 60$;

Scenario 6 initial condition X_0^m: [0, 0]; penalty cost for a unit of production loss $q = 30$;

Scenario 7 initial condition X_0^m: [0, 0]; penalty cost for a unit of production loss $q = 10$.

For each scenario, the difference in total cost rate of the best case given by the proposed strategy benchmarked against the two traditional strategies are summarised in Table 5.10. The performance superiority of the proposed model gets more obvious with a larger q for scenarios with either initial degradation level. It can also be noticed that for $q = 30, 60$ and 90, the proposed model generates better results for an initial degradation level of [0, 0] than [0, 40], which is in accordance with discussions earlier in this chapter that the uniform allocation strategy will lose its edge if units start with less heterogeneous condition. One of the implications here is that the proposed model is likely to bring more benefits to companies that have very high costs for production losses, such as those in the oil and gas industry.

Table 5.10 Difference in total cost rate benchmarked against traditional strategies

	Comparison with Rnd.		Comparison with Uni.	
Initial degradation level	[0, 0]	[0, 40]	[0, 0]	[0, 40]
$q = 10$	-0.95%	-1.68%	2.70%	2.12%
$q = 30$	-3.53%	-3.43%	0.43%	0.79%
$q = 60$	-7.26%	-5.55%	-2.96%	-1.55%
$q = 90$	-9.13%	-8.12%	-4.52%	-3.01%

5.8 Chapter Summary

The main aim of this chapter is to answer the third research question, which is to quantify the impact of a specific workload allocation on maintenance and production at the system-level, and figure out the types of information needed in order to do so. A decision-making model for the integrated optimisation of condition-based maintenance threshold and workload allocation among a fleet of asset has been developed. This is achieved by combining the load-dependent individual asset maintenance model developed in the previous chapter and a coordinator-level workload allocation strategy proposed in this chapter.

Following the discussion in Chapter 4, a multi-agent structure is believed to be suitable for formulating such an integrated model. Two types of agents are needed for the decision-making process - machine agents that monitor units and seek optimal solutions for individual asset models, and a coordinator agent tasked with system-level workload allocation. The two types of information that the coordinator will need from machine agents in order to make a

conscious decision are identified, and the coordinator objective function is formed based on the rationale that any decision made at the moment will have an instant effect on the system as well as long-term consequences. To be specific, the first component is the instant penalty for production losses, the second one accounts for the marginal maintenance-related costs, and the last component concerns potentially penalty costs in the near future.

A mixed-integer genetic algorithm is then introduced to solve the coordinator-level optimisation problem. Numerical examples are given to highlight the role of the long-term penalty-related component in shaping the decision-making process. An important parameter in the objective function, the buffer size B is found to be have significant impacts on the performance of the model. The concept of a buffer is initially introduced into the model to 1) compensate the point estimate of the time to the next PM/RP of each machine by creating a range around it; 2) adjust the amount of attention that the coordinator directs to ensuring a satisfactorily small numerical value is taken by the long-term penalty-related component. Closer inspection has revealed that increasing B within a reasonable range has a stabilising effect that reduces the degree of fluctuations in the workload allocation between consecutive decision-making epochs, which is positively correlated with the model performance. It follows that an appropriate value needs to assigned to B to allow the proposed approach to realise its greatest potential. When B is too small, the optimal solutions obtained tend to be very sensitive to the stochastic degradation process of machines, whereas too large a B renders the model indifferent between different workload allocations. The numerical examples have also demonstrated the capability of the model to widen the time gap between the need for maintenance between units and to improve the system-level risk profile.

The performance of the proposed model is compared with that of three traditional strategies - uniform workload allocation, random workload allocation, and more-on-new-machines workload allocation. Sensitivity analysis is conducted on two important sets of model parameters - the initial degradation condition of units X_0^m and the penalty cost for a unit of loss of production q. The combined results from these experiments have proved the superiority of the proposed model in generating a lower total cost rate benchmarked against traditional strategies. Such improved performance is achieved by appropriately shifting workload both between units and along the timeline for a single unit, which leads to substantial saving in penalty costs. It has also been revealed that the advantage of the proposed model is more significant when loss of production is expensive and when units start in less heterogeneous degradation condition. These findings can serve as a guidance for asset owners to evaluate the usefulness of the decision-making model based on their industry and asset characteristics.

Despite the improved performance achieved by the proposed model, some limitations have also been identified. For instance, there is a lack of systematic approach to select the optimal buffer size B. It is also yet to be explored whether deliberately stabilising the workload allocation will further enhance the model performance. However, two potential ways of choosing B will be proposed in Chapter 6 when the proposed approach is applied in the context of a real case example in the oil and gas industry. These limitations and what future research can be conducted to address such limitations will be discussed in Chapter 7.

The next chapter will test the practicality and performance of the proposed decision-making model in a real industrial context.

Chapter 6

Case Example

6.1 Introduction

This chapter aims to demonstrate the practicality of the proposed decision-making model by applying it to a real-industry example - a fleet of vessels in a secondary effluent water treatment (SET) plant that belongs to an oil refinery. First, in Section 6.2 the industrial context is established and the problem faced by the case company is presented, after which a description is given of the information extraction and pre-processing approach for obtaining necessary data as model inputs. Then the decision-making methodology is tested and the results are analysed for two different scenarios in Section 6.3 and 6.4, respectively. The first scenario is a closer representation of reality where the vessels only have on/off flow control whereas the second scenario assumes that continuous flow control is enabled. In both scenarios, the impact of long-term considerations on how workload is allocated amongst the vessels and therefore on model performance is discussed. Apart from comparison with traditional strategies, the model performance under the two scenarios are compared against each other to emphasise the benefits brought by having reasonable flexibility in workload control. The chapter is closed with a summary in Section 6.5.

6.2 Case Description: SET Plant Vessels

This section gives a detailed description of the industrial context of the case study and presents the data that will later be used as model inputs.

6.2.1 Problem Description

The basic information regarding the case company and the assets that are of interest to this research study have already been discussed in Section 3.3.2 in the exploratory case studies chapter. Here a brief recapture of the problem background is provided. The subject of this case study is a fleet of vessels in an oil refinery. A diagram of a typical vessel can be found in Figure 6.1. These vessels are mainly used for filtering effluent water before it can be discharged into the river. A very important health indicator of a vessel is the status of its internal lining. The lining is often subject to multiple effects (such as ageing, erosion, and scratching) that lead to accumulative damage. If the lining is left to degrade without any maintenance interventions, and the water treatment system reaches a state of high risk, the plant might suffer from severe loss caused by failure to meet environmental standards.

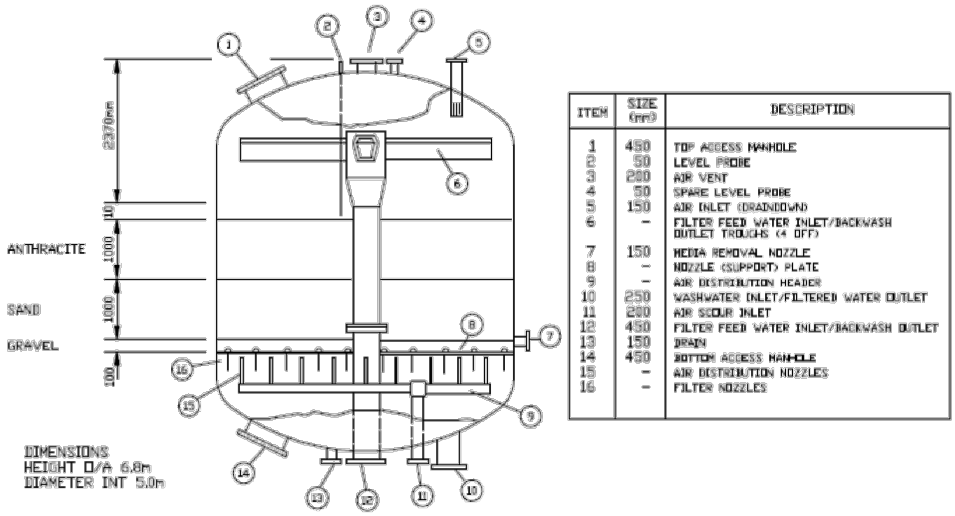


Fig. 6.1 Diagram of a typical vessel in a SET plant (provided by the company)

The SET plant consists of seven identical vessels operating in parallel. As an operational requirement, each vessel needs to go through a back-wash process every four hours, which takes around forty minutes to complete. Due to the redundancy and flexibility embedded in such configuration, the system can easily maintain its desired capacity with two vessels taken off-line. A 'run-to-failure' strategy has been in use until recently when all vessels have become heavily degraded and the plant is at high risk. A major re-lining programme is currently being carried out on all vessels and the reliability manager is interested in developing a maintenance strategy for the vessels for their new life cycle. It is expected that the strategy can potentially take advantage of any workload adjustment capability the plant possesses.

6.2.2 Data Extraction and Pre-processing

The inputs for the decision-making model are presented in Table 6.1. Since it is impossible to have all required parameters readily available, estimates and guesses have to be made based on incomplete information and expert advice. A label is given to each of the parameters to show its source (A=actual figure, E=well-informed estimate, and I=rough estimate). Parameters are mostly shared by all vessels except for the initial degradation condition X_0. The rest of the subsection gives a detailed explanation of how these model inputs are obtained.

Table 6.1 Parameters used for case studies on vessels

Production-related	$W(\mathrm{m}^3/\mathrm{month})$	$D(\mathrm{m}^3/\mathrm{month})$	M	$y(\cdot)(£)$					
	E	E	A	E					
	288k	1,440k	7	**[2]**					
Deterioration-related	λ_0	λ_1	κ_0	θ_0	s	X_0	N_0	F	
	E	E	E	E	E	E	A	I	
	3.51e-5	11.34	0.50	1.37	0.20	**[1]**	0	360	
Maintenance-related	c_{mr}	c_{pm}	c_{rp}	t_{mr}	t_{pm}	t_{rp}	α_0	β_0	b
	E	A	A	E	A	A	E	E	E
	£182,125	£364,250	£782,990	1	2	6	2.00	7.50	0.87

[1] [0, 9.6, 19.2, 28.8, 38.4, 48.0, 57.6]

[2a] For continuous flow control: $y(U, D) = 0.44643 \cdot \max(0, \frac{6}{5}D - \sum_{m=1}^{7} u^m)$

[2b] For on/off flow control: $y(N^f) = \begin{cases} 0 & \text{if } N^f = 0 \text{ or } N^f = 1 \\ 107,143 & \text{if } N^f = 2 \\ 214,286 & \text{if } N^f = 3 \\ 321,430 & \text{if } N^f = 4 \\ 428,570 & \text{if } N^f = 5 \\ 535,716 & \text{if } N^f = 6 \\ 642,860 & \text{if } N^f = 7 \end{cases}$

Production-related Data

Each of the vessels is designed to have a full effluent water treatment capacity of $540\mathrm{m}^3/\mathrm{h}$. However due to limitation in the efficiency of the filter media and pumping capacity, the actual throughput of one vessel is around $400\mathrm{m}^3/\mathrm{h}$. The yearly average flow rate of waste

water is around $1400\text{m}^3/\text{h}$. This figure can vary due to fluctuation in production and weather conditions. In order to have extra insurance, the reliability manager had requested that an hourly demand of $2000\text{m}^3/\text{h}$ be chosen as the model input. Namely, the plant needs to have at least five vessels up and running to fulfil the demand. Furthermore, since each vessel is programmed to go through back-washes every four hours due to operational requirements, the system can only afford at most one vessel being under maintenance at any point in time without failing to meet the production expectations.

In this particular case, the penalty cost here is partly caused by the reduction in revenues and partly due to potential pollution resulting from failing to supply the expected amount of end products, such as petrol, marine fuels, heating oil, and diesel. In order to write the penalty cost as a function of production loss of the vessels (unfulfilled demand for treating wasted water). We start with an estimate of monthly losses in monetary terms provided by the management team of the plant and make the following assumptions as suggested by the reliability manager:

1. There exists a linear relationship between the amount of end products and the volume of wasted water that needs to be treated by the SET plant.

2. There also exists a linear relationship between the monetary losses and the amount of end products that can be produced by the plant.

If all seven vessels are down for maintenance for one month, the plant will generate a total loss of £642,860, whereas no loss will be incurred if no more than one vessel fails. With the two assumptions made above, the penalty costs incurred by any number of vessel failures can be deduced. The plant operates 24/7 and the amount of effluent water that can be treated by one vessel per month (30 days) is calculated by

$$400\text{m}^3/\text{h} \times 24\text{h} \times 30 = 288,000\text{m}^3$$

Then the total capacity provided by 5 vessels in one month is $288,000 \times 5 = 1,440,000\text{m}^3$. It then follows that the penalty cost incurred by a unit of production loss will be

$$642,860 \div 1,440,000 = £0.44643/\text{m}^3$$

Since the proposed methodology will be applied to two case scenarios in later sections, two different forms of penalty functions are derived. It is worth noting that the time taken to perform back-washes should also be taken into consideration while deriving these functions. Specifically, the filter media in each vessel in operation needs to be back-washed every four hours and the process takes forty minutes. Namely, the vessels will only be able to realise

$5/6$ of their full capacity even in their most healthy state. For the second scenario where continuous flow control is assumed, $y(\cdot)$ is simply a linear function of the production loss incurred.

$$q_k = 0.44643 \cdot \max(0, D_k - \frac{5}{6}\sum_{m=1}^{7} u^m) \tag{6.1}$$

A coefficient of $\frac{5}{6}$ is introduced before $\sum_{m=1}^{7} u^m$ in order to incorporate the impact on production caused by back-washes. Note that in calculating $O(\cdot)$ in equation 5.3, the assigned workload u^m will be scaled back to be $\frac{5}{6}u^m$. Similarly, the capacity of unit is set to be $\frac{5}{6}W$ while calculating the long-term penalty-related component in equation 5.3.

For the first scenario, $y(\cdot)$ will be a step function since a vessel can only be either operating at full capacity or on standby. It is obvious that when there is only one vessel breakdown, no production loss will be incurred as six vessels are more than enough to provide the required capacity. While two vessels undergo maintenance simultaneously, the monthly maximum capacity five vessels can reach is calculated as follows

$$288,000 \times 5 \times \frac{5}{6} = 1,200,000\text{m}^3$$

The monthly penalty cost for this two-breakdown case is thus

$$(1,440,000 - 1,200,000)\text{m}^3 \times £0.44643/\text{m}^3 = £107,143$$

The monthly penalty cost for cases where there are more overlapping machine breakdowns can be calculated with the same approach and is given by

$$q_k = \begin{cases} 0 & \text{if } N_k^f = 0 \text{ or } N_k^f = 1 \\ 107,143 & \text{if } N_k^f = 2 \\ 214,286 & \text{if } N_k^f = 3 \\ 321,430 & \text{if } N_k^f = 4 \\ 428,570 & \text{if } N_k^f = 5 \\ 535,716 & \text{if } N_k^f = 6 \\ 642,860 & \text{if } N_k^f = 7 \end{cases}, \tag{6.2}$$

where N_k^f denotes the number of machines undergoing maintenance over the k^{th} month.

Deterioration-related Data

In this case, there are two ways to know the condition of the internal lining, ultrasonic-based technology and in-person visual inspection. However as neither of these actions have been performed since the installation of the vessels, no condition monitoring data is available. Here we rely on the lifetime estimates provided by the reliability manager in order to formulate a load-dependent deterioration model for the vessels. The raw estimates are given Table 6.2 in the form of Weibull distribution parameters for three different usage frequencies. It can be

Table 6.2 Estimates of Weibull distribution parameters for vessels

Parameters (yr)	Degrees of usage		
	Always in use ($r = 1$)	Rarely in use ($r = 0.2$)	Always on standby ($r = 0$)
β_W-shape parameter	14	18	20
η_W-scale parameter	37.5	44	45

observed that more frequently used vessels are expected to have a shorter characteristic life denoted by a small η_W. In order to make the proposed model directly applicable to this data set, the lifetime Weibull distribution parameters need to be converted to equivalent Gamma degradation process parameters. The rationale behind this conversion is that for all three workload conditions, the resultant Gamma degradation process should give at least a similar lifetime distribution to the Weibull one. First without loss of generality, the failure threshold F is assumed to be 360. According to equation 4.1 and 4.2, three parameters are needed to fully characterise the load-dependent Gamma process - shape parameter κ_0, scale parameter θ_0, and coefficient of sensitivity to load s. The following steps are followed to determine these parameters:

1. Generate the $1-100^{th}$ Weibull distribution percentile for all three degrees of usage.
2. Define a wide enough range for κ_0, θ_0, and s, within which potential candidates for these three values are generated.
3. Generate a large number of value combinations of κ_0, θ_0, and s within the range.
4. For the Gamma distribution characterised by each combination generated in the previous step, calculate its aggregated distance $D_W = \sqrt{D_{W1}^2 + D_{W2}^2 + D_{W3}^2}$ from the Weibull distributions, where D_{W1}, D_{W2}, and D_{W3} are the distance between the Gamma distribution and the Weibull distributions with $r = 0$, $r = 0.2$, and $r = 1$, respectively. Here we take D_{W1} as an example to demonstrate how the distance is calculated:

- For each of the $1-100^{th}$ percentile of the Weibull distribution with $r=0$, find its survival rate with the Gamma distribution.
- Then D_{W1} is simply the distance between the vector of survival rates and the vector $(0.01, 0.02, \cdots, 1.00)$.

5. Select the combination $[\kappa_0^*, \theta_0^*, s^*]$ that has the smallest distance.

In this particular case, as the reliability manager is least confident in his estimates for when $r=0.2$, $D_W = \sqrt{D_{W1}^2 + 0.3 \times D_{W2}^2 + D_{W3}^2}$ is used, which leads to a smaller weight being assigned to the second term, for calculating the aggregated distance. Using the described approach, the following Gamma distribution parameters are obtained: $[\kappa_0^*, \theta_0^*, s^*] = [0.5, 1.3684, 0.2]$. Figure 6.2 gives the comparison between the cumulative distribution function of vessel lifetime generated using the Weibull distribution and the calibrated Gamma degradation process. It can be observed that the outcomes produced by Weibull distribution and Gamma process exhibited consistent similarity across all three usage scenarios.

The company has a maintenance practice manual named Equipment Degradation Documents (EDD), which provides general degradation and risk information of various assets being used in its plant. Here the parameters in the proposed model that are related to shock rates will be deduced from the failure probabilities given in EDD for epoxy linings. EDD places epoxy linings into five categories $[E, D, C, B, A]$ based on the severity of degradation, with increasing failure probability $[0.0001, 0.0003, 0.003, 0.03, 0.3]$. In order to relate the failure probability to the random shocks in the proposed model defined by a non-homogeneous Poisson process with intensity given by equation 4.4, we evenly divide the possible degradation level between 0-360 into five stages and equate their shock rates at the beginning of each state to $[0.0001, 0.0003, 0.003, 0.03, 0.3]$. The resultant values for λ_0 and λ_1 are 3.51e-5 and 11.34, respectively. A comparison between the failure probabilities given by EDD and shock rates generated by the Poisson Process is presented in Figure 6.3. It can be observed that most of the time the calibrated Poisson Process is able to closely track the EDD trend.

It has been mentioned previously in this chapter, a full repair has just been completed on one of the vessels, and the other vessels will be undergoing a relining process in the next six years at a rate of one vessel per year. The initial condition will thus be different among the vessels in six years when the maintenance strategy is initiated. Specifically, with the characteristic life of a vessel always in use being 37.5 years and assuming a roughly linear degradation process, the yearly deterioration increment will be $\frac{F}{37.5} = 9.6$. The initial degradation level of the vessel fleet at the beginning of the 7^{th} year is therefore [0, 9.6, 19.2, 28.8, 38.4, 48.0, 57.6].

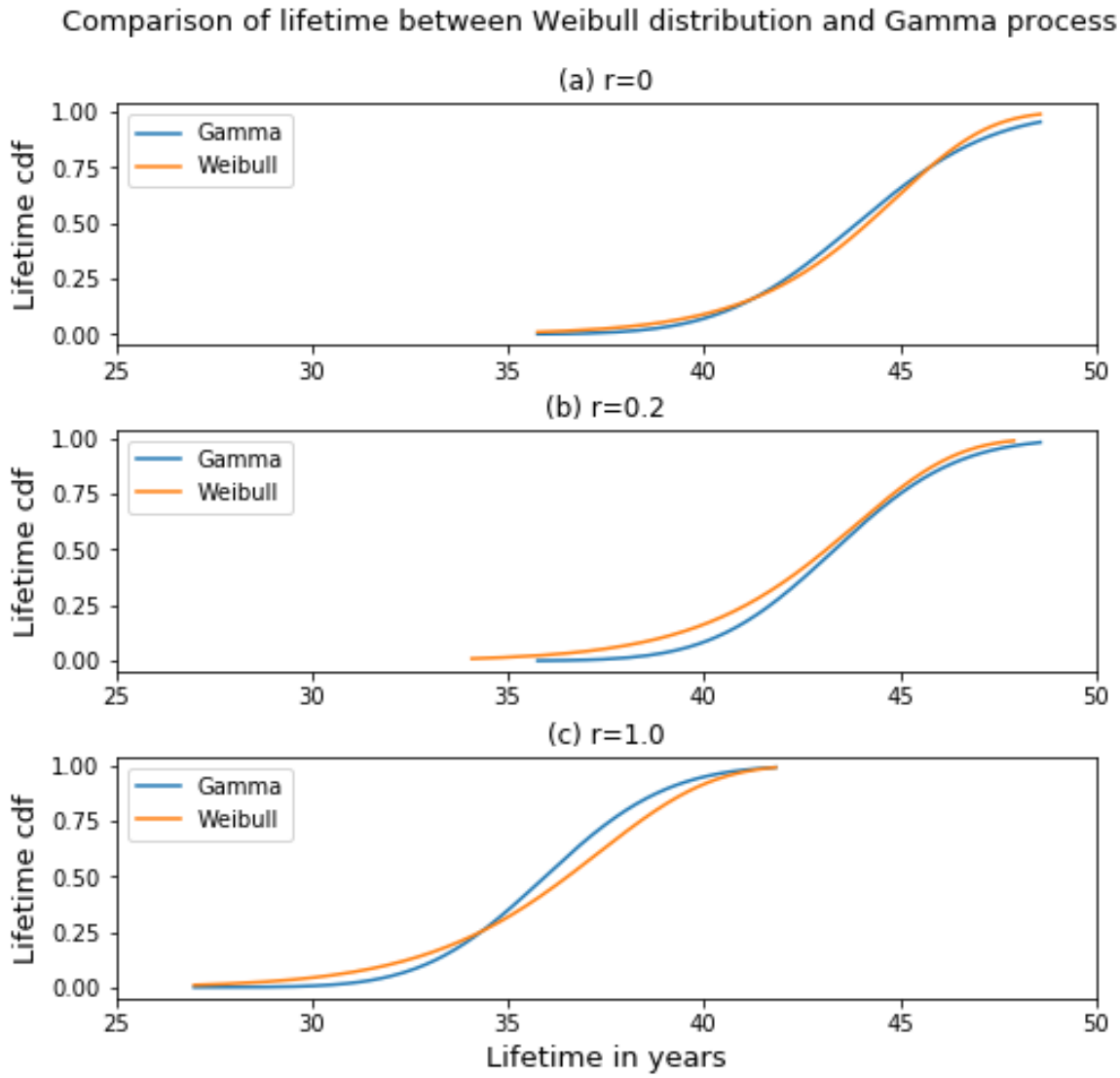


Fig. 6.2 Comparison of lifetime generated by Weibull distribution and Gamma process

Maintenance-related Data

Two types of maintenance actions can be performed on the vessels: patch repairs and full repairs. Full repairs refer to a six-month thorough relining of the internal walls, which bring the vessels back to the 'as-good-as-new' state. The cost for a full repair is £782,990. Patch repairs are used to treat localised corrosion of the lining. These areas are typically patched with Belzona, a type of protective coating. Furthermore, patch repairs can be either expected or unexpected. For expected patch repairs, which costs £364,250 and takes two months, weak areas on the lining are carefully treated. However, when unexpected localised lining breakdown happens, patch repairs are done rather fast to 'get the vessel up and running for

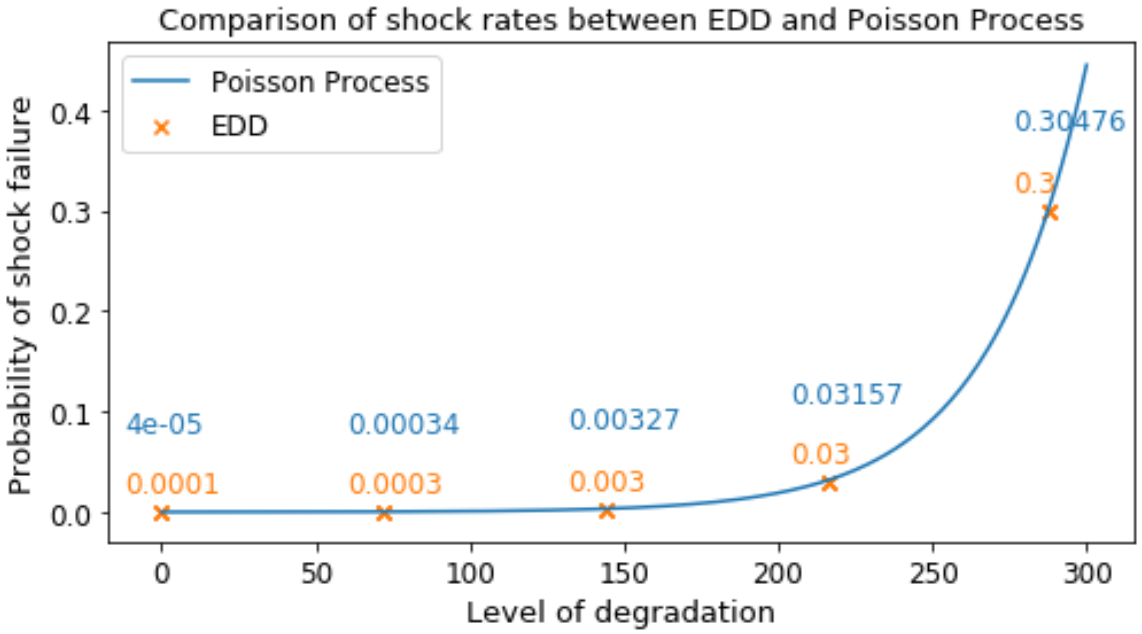


Fig. 6.3 Comparison of shock rates given by EDD and Poisson process

at least a little while'. The first former patch repairs can thus be perceived as imperfect preventive maintenance, whereas the latter is more similar to minimal repairs. As there has been no readily available data for the cost and time duration of minimal repairs, a figure of £182,125 and one month are assumed in this study, both of which are half of what imperfect preventive maintenance incurs.

As mentioned in Chapter 4, patch repairs of internal linings of large vessels requires the set-up and tear-down of scaffolds, which might damage the linings and result in areas of accelerated corrosion. Therefore, imperfect preventive maintenance in this case tend to have diminishing rectification effects. Recall that in the individual asset model, the mean and variance of $\frac{X(R_i^+)}{X_(R_i^-)}$ are given by equation 4.9 and 4.10 respectively, which ultimately are determined by three parameters: α_0, β_0, and b. A rough estimate of the rectification effect of preventive maintenance is an age reduction of four to six years. In view of the fact that the cost for preventive maintenance is rather high for the vessels, frequent maintenance actions are not expected. Here we assume that a vessel will undergo its first preventive maintenance in its 10^{th} year in service. Namely, $\mathrm{E}(\frac{X(R_1^+)}{X_(R_1^-)})$ is set to be $\frac{10-6}{10} = 0.4$. For the same reason, it is not worth having too many preventive maintenance tasks before a full repair, the rectification effect is expected to fast diminish. Therefore, $\mathrm{E}(\frac{X(R_2^+)}{X_(R_2^-)})$ are $\mathrm{E}(\frac{X(R_3^+)}{X_(R_3^-)})$ are set to be 0.6 and 0.8 respectively. Using equation 4.9, we have $\alpha_0 = 2.00$, $\beta_0 = 7.50$, and $b = 0.87$.

In the next section, the input data obtained above is applied to numerical case studies to test the practical applicability of the proposed model.

6.3 Case Scenario 1: On/Off Flow Control

This case scenario is a closer representation of the reality compared with the second scenario in the next section. Since for now no variable-flow pumps are installed to control the volume of effluent water flowing into each vessel, here the decision will be whether to utilise a vessel to its full capacity or have it on standby. Namely, the load ratio the coordinator chooses for each vessel will be binary - 0 or 1. The model inputs for this scenario can be found in Table 6.1.

6.3.1 Results and Discussion of a Single Replication

Here we also consider both the basic version, which is an exemplar of the more-on-new-machine allocation strategy, and the complete version of the proposed model. In the complete version, the buffer size B takes the following values [0, 2, 4, 5, 6, 8, 10]. As the number of parallel assets in this case study is much larger than that used in the numerical examples in Chapter 5 (7 instead of 2), only parts of the analysis carried out previously can be applied directly here. One such type of analysis is how the trend of workload allocation among assets with different initial degradation levels changes as B increases. For demonstration purposes, the workload allocation before the first preventive maintenance is performed in each case is taken as the unit of analysis. The average load ratio allocated to each vessel is given in Table 6.3. The table also presents the difference in the workload assigned between vessel 7 and 1.

It can be observed that in the basic version of the model, which is an exemplar of the more-on-new-machines allocation strategy as mentioned in Section 5.6, the workload is first allocated to the newer vessels (1-5), and the rest goes to the older ones (6 and 7). This is a natural outcome as the workload allocation decision made by the basic version is solely based on the sum of maintenance cost rates across vessels and newer vessels tend to generate less maintenance costs. Also note that the difference in workload assigned to vessel 7 and 1 tends to go up first and then down as B is increased from non-existent in the basic version to 10 in the complete version. Such difference can be perceived as characterising the willingness of allocating more workload to older vessels. With the older vessels heavily loaded, the gaps between the maintenance slots of the fleet are widened in time and the likelihood of overlapping machine breakdowns is thus reduced. It will be shown later in Section 6.3.2 that

Table 6.3 Average workload allocated to each vessel before the first PM

Case	Vessel No.							Difference between 7 and 1
	1	2	3	4	5	6	7	
Basic	1.00	1.00	1.00	1.00	1.00	0.72	0.28	-0.72
$B = 0$	0.42	0.82	0.99	1.00	1.00	0.89	0.88	0.46
$B = 2$	0.24	0.76	1.00	1.00	1.00	1.00	1.00	0.76
$B = 4$	0.17	0.83	1.00	1.00	1.00	1.00	1.00	0.83
$B = 5$	0.13	0.89	0.98	1.00	1.00	1.00	1.00	0.87
$B = 6$	0.20	0.81	0.99	1.00	1.00	1.00	1.00	0.80
$B = 8$	0.35	0.65	1.00	1.00	1.00	1.00	1.00	0.65
$B = 10$	0.94	0.06	1.00	1.00	1.00	1.00	1.00	0.06

the buffer size $B = 5$ which corresponds to the biggest difference between 7 and 1 also leads to the most cost saving in penalty costs resulted from production losses.

The next section further compares the performance of the proposed model with that of random and uniform workload allocation averaged over multiple replications of simulation.

6.3.2 Performance Comparison with Traditional Strategies

For random workload allocation, at each decision epoch, 6 out of 7 vessels are picked randomly and will operate at full capacity. For uniform workload allocation, the constraint of on/off flow control is temporarily released and all functional vessels will be set to fulfil an equal share of the total demand. We have run 110 replications of simulation for each of the cases considered here. This number is chosen such that the relative relationship between various scenarios is stabilised. The accumulative average total cost rate across replications for a selective group of cases is plotted in Figure 6.4, which indicates a consistent relationship between various cases after the 85^{th} replication. The time horizon, based on which the model performance has been measured, is chosen to be 475 intervals (equivalent to 39.6 years), since all vessels will have been replaced by this point in time in every possible scenario.

The total cost rate for each case considered can be found in Figure 6.5. Similar to the results presented in the numerical examples in Chapter 5, random workload allocation leads to the highest average total cost rate amongst all cases. The basic version of the model outperforms random allocation but fails to beat uniform allocation. The complete version generates lower total cost rates benchmarked against all three traditional scenarios under all buffer sizes ranging from 0 to 10, with the best performance given by $B = 6$ whose cost

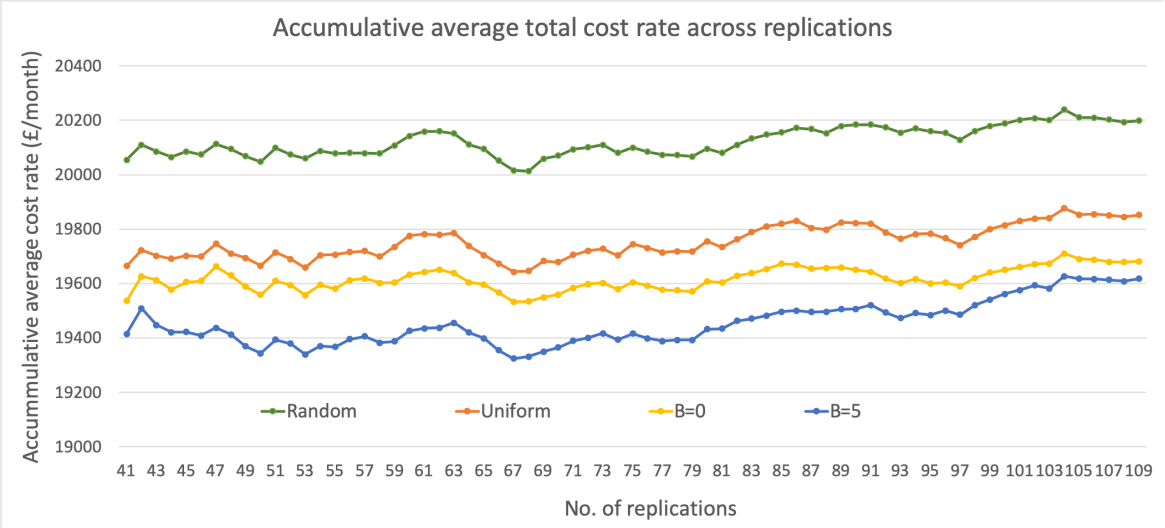


Fig. 6.4 Accumulative average total cost rate across replications

rate is 3.01%, 1.34%, and 1.90% lower than that of uniform, random, and more-on-new-machine workload allocation, respectively. Recall in Section 5.6.2 the proposed model only outperforms uniform allocation for B within a certain range, and it was argued that uniform allocation could gain extra advantage over the proposed model with its intrinsic stabilising effect. Here due to the constraint imposed by on/off flow control, at each time epoch the proposed model also tends to stick with the workload allocation determined by the previous epoch unless significant changes happen to the vessel fleet. Consequently, uniform workload allocation will lose its edge under such constraints. The observation that the basic version again fails to beat uniform allocation and the complete version of the model can be perceived as evidence that it is not necessarily optimal to assign higher workloads to newer machines. In the basic version of the model, the oldest vessel - vessel 7, has been put on standby until vessel 6 reaches almost the same degradation level as vessel 7, after which the workload is shifted in iteration between 6 and 7. As a result, the basic version ends up with a much higher variance in the load ratios assigned to vessel 6 and 7 than other strategies. The data obtained from one typical simulation replication is presented in Table 6.4 for the purpose of demonstration. Since the system performance is to some extent negatively correlated with the fluctuation in workload assigned to the vessels, such characteristics of the basic version are not preferred.

Since the performance of random workload allocation and the basic version of the model is far worse than any other strategy, the focus here will be on the comparison between uniform workload allocation and the proposed complete version of the model from this point onwards. The difference between the two in total cost rate is further broken down into two components

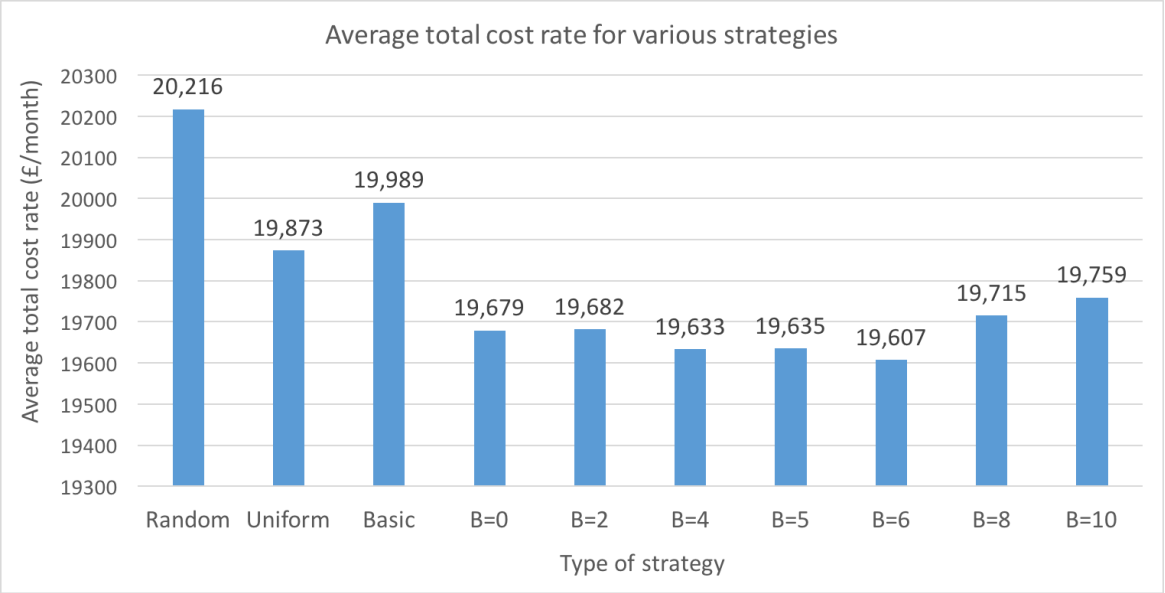


Fig. 6.5 Comparison of the average total cost rate for various strategies

Table 6.4 Standard deviation of the workload ratio of vessel 6 and 7 before the first PM

Case	Complete version							Basic version
	B=0	B=2	B=4	B=5	B=6	B=8	B=10	
Vessel 6 (avg)	0.89	1.00	1.00	1.00	1.00	1.00	1.00	0.72
Vessel 6 (std)	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.45
Vessel 7 (avg)	0.88	1.00	1.00	1.00	1.00	1.00	1.00	0.28
Vessel 7 (std)	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.45

in Figure 6.6: maintenance-related cost rate and penalty cost rate. In the previous section it was mentioned that the largest difference (0.87) in average workload assigned to vessel 7 and 1 is found while $B = 5$, implying the most efforts dedicated to avoiding production losses. Not surprisingly, here having $B = 5$ leads to the highest saving in penalty cost rate (£64/month) among all cases.

Unlike the example presented in Section 5.6.2 where saving in penalty cost contributes to a large proportion of the gap, here most of the saving in total cost rate results from lower maintenance-related costs. Such divergence in the modelling results indicates the difference in the relative magnitude of the two cost components in the total costs generated by the production system. Take the optimal solution of the first decision-making epoch as an example, the ratio of the penalty cost rate and the maintenance-related cost rate in Section 5.6.2 for $q = 30$, $q = 60$, and $q = 90$ grow from 2.82% to 87.2%, 5.04% to 188.40%,

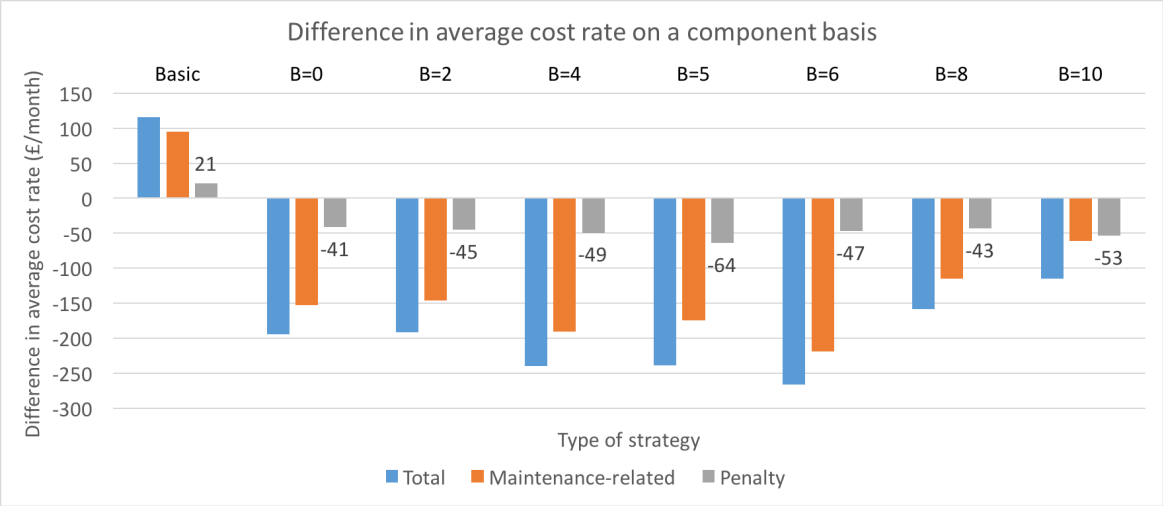


Fig. 6.6 Difference in the average cost rate for various strategies on a component basis

and 8.28% to 273.72%, respectively, as B is increased from 0 to 8. The number for case scenario 1 here is 0.73% and 48.05%. It is obvious that maintenance-related costs take up much higher proportions of total costs in case scenario 1 than the hypothetical numerical examples in the previous chapter.

6.3.3 Issues with Choosing Parameter B

The way via which B is chosen depends on the amount of data, time, and computational power available before the decision-making system is installed.

- **Case 1:** When there is enough time and computational power for running simulations, B should be determined via simulations. In the best-case scenario where sufficient and accurate data regarding the physical properties of the production system and maintenance tasks is available, simulations can be run for the entire life cycle of machines using a range of B values and whichever B that generates the lowest system-level cost rate should be picked. When the amount and quality of the data available is not readily available, simulations need to be run using estimated data for the life cycle of machines using a range of B values. In this case, however, when the decision-making system is installed and more data becomes available as time goes by, the operators need to rerun the simulation using updated data to decide whether the initial B needs to be changed.

- **Case 2:** In cases where time and computational power are limited, B can still be determined using simulation. Instead of running simulation for the life cycle of

machines, the operator only runs the simulation for a short period of time. Specifically, for each B value in a given range, a trial of several epochs is run (the number of epochs varies depending on the length of the asset life cycle and can be set as a proportion of the life expectancy of assets). Recall that the buffer has a stabilising impact on the workload allocation recommended by the proposed model for consecutive time epochs and that the model performance is positively correlated with the stability of the workload allocation. Following this principle, whichever B that results in the least fluctuation in the workload allocation for consecutive epochs should be chosen.

The question that remains is how the range of B values should be determined to provide some guarantee that it includes the optimal B. As B acts as a safety net against the uncertainty regarding the estimated time to the next PM/RP, it is reasonable to relate the range of B to the confidence interval of such estimation. However, to go through all B values bounded by multiples of the standard deviation of the estimated time to the next PM/RP is not always necessary, as the model might be indifferent between various workload allocations before B even takes that value. Here we recommend that the lower limit of B should be set to be the of the same length as the PM duration, and the upper limit of B to be twice the standard deviation of the estimated time to the next PM/RP. In Case 1, run simulations starting with the lower limit of B; gradually increase the value of B; stop when 1) all B values are used; or 2) when there is significant increase of total cost rate; choose the one that generates the lowest total cost rate. In Case 2, run simulations starting with the lower limit of B; gradually increase the value of B; stop when 1) all B values are used; or 2) when there is significant increase of fluctuations in workload allocation; choose the one that results in the least fluctuation in the workload allocation for consecutive epochs.

For the industrial case presented in this section, different B values would be recommended depending on whether simulation can be conducted for the life cycle of vessels. If there is sufficient time and computational power for such simulation (Case 1), it is obvious from the results as presented in Figure 6.5 that $B = 6$ is the best choice, which generates the lowest total cost rate among all strategies. If we were to deal with Case 2, due to the on/off constraint imposed on the system which forces the model to stabilise the recommended workload allocation, even by having $B = 0$ would result in perfectly stable workload allocation for the early epochs. The value of B would therefore be set to be 0 in this case. While the resultant total cost rate is higher than when the optimal B value is used, it is still lower than the traditional strategies.

6.3.4 Guidance on Implementation

In order to make use of the proposed approach in the operation of these SET vessels, two major types of preparation work need to be carried out:

1. **Preparing the data required:** referring back to Table 6.1, we can see that some of the inputs used in this case study are estimated using historical records or from expert opinions. Such estimates might be out dated or inaccurate, and thus additional data needs to be collected. Specifically, the company needs to improve both the quality and quantity of a) degradation-related data, which includes both the initial condition as well as the Gamma process parameters of the vessels. Such data can be obtained by having more frequent inspections on the vessels as well as from the OEM of the internal lining material; b) PM-related data that quantifies the rectification effect of imperfect PMs. This could be achieved by a close comparison between the condition of vessels before and after PM.

2. **Installing the decision-making system:** the company can install the optimisation model on a computer and input the required parameters into the model. With the current on/off flow control capability, a buffer size of $B = 6$ should be adopted. It is also recommended that separate simulation model is made available on the side so that the model parameters can be refined from time to time with new data accumulating.

Once all preparation work has been completed, the system is ready for use. The operator is supposed to run the optimisation model at the beginning of each month to decide on the vessels that are to be put on-line for this time period. Every one or two year, it is recommended that the operator run the separate simulation model with updated inputs to check if changes need to be made to B. The parameters in the simulation model should also be updated if they now differ significantly from their previous values.

6.3.5 Case Summary and Discussion

This case example has demonstrated the usefulness of the proposed decision-making model in a practical context. The proposed approach is used to dynamically determine the workload allocation and condition-based maintenance threshold for a fleet of seven vessels in order to minimise the total cost rate over the life-cycle of the system. In this scenario, only on/off flow control is enabled on the vessels, which represents the reality the way it is. It has been demonstrated through the case study that by considering the workload-dependent degradation behaviour of SET plant vessels and consciously making decisions on whether to turn on the flow control valves of each vessel, both maintenance costs and penalty costs incurred due to

loss of production across the system would be less than that generated by traditional strategies such as uniform and random workload allocation. The extent to which the total system-level cost rate can be reduced is dependent on the value of the buffer size B, where the best performance is achieved with $B = 6$. Furthermore, it has been shown that the basic version of the model can be perceived as a strategy that intuitively allocates full workload to newer vessels and sequentially moves to the older ones, which is also outperformed by the complete version of the model. Depending on whether sufficient data, time, and computational power is available, different approaches have been recommended for choosing an appropriate buffer size B. A brief discussion is made on how such approaches can be adopted to select B in this particular case example, after which guidance on the implementation of the proposed model in the context of this vessel fleet is provided.

6.4 Case Scenario 2: Continuous Flow Control

This case scenario is focused on the same fleet of vessels as the the first scenario. It is however further assumed that continuous flow control is enabled on all vessel valves such that the volume of effluent water flowing into each vessel can be any real number between 0 and the full capacity 288,000m^3/month.

6.4.1 Results and Discussion of a Single Replication

From both the numerical examples presented in Section 5.6.2 and the modelling results in Scenario 1, we can see that in most if not all cases the best performance of the proposed model is rendered by the complete version of the model with an appropriate buffer size B. Furthermore, in view of the fact that it is very time-consuming to run multiple simulation replications of the entire life-cycle for the vessel fleet, here experiments are only done on a selected set of B values, [4, 5, 6], for the complete version of the model. These three B values are chosen based on the observation that they are the top three best practices in Scenario 1. Since except for the integer constraint imposed on decision variables most input parameters are the same in Scenario 1 and 2, it is highly probable that they will result in the lowest total cost rates in Scenario 2 as well.

Similar to Case Scenario 1, here we compare the trend in workload allocation, as a ratio of the vessel capacity. For a specific replication, the average load ratio allocated to each vessel before the first PM is shown in Table 6.5.

Again, similar observations can be made from the data provided. As B increases from 4 to 6, directing the attention of the coordinator to fighting against production losses, the

Table 6.5 Average load ratio allocated to each vessel before the first PM

Case	Vessel No.							Difference between 7 and 1
	1	2	3	4	5	6	7	
$B = 4$	0.82	0.85	0.85	0.87	0.86	0.87	0.87	0.04
$B = 5$	0.80	0.86	0.88	0.86	0.86	0.86	0.88	0.08
$B = 6$	0.79	0.86	0.86	0.87	0.87	0.87	0.89	0.10

workload of vessel 1 drops from 0.82 to 0.79 whereas that of vessel 7 is increased from 0.87 to 0.89. The trend of shifting workload from newer vessels to older ones is better represented by the difference between vessel 7 and 1, which increases from 0.04 to 0.10. In the following section, we will show that the case where $B = 6$ has indeed led to the lowest penalty costs out of all cases considered.

6.4.2 Performance Comparison with Traditional Strategies

For random workload allocation in Scenario 2, at each decision epoch, the total monthly demand is randomly assigned amongst all operational vessels, as opposed to Scenario 1 where 6 vessels are picked out of 7 to operate at full capacity. Similarly, 110 replications of simulation are run for each case and the time horizon considered is 475 intervals.

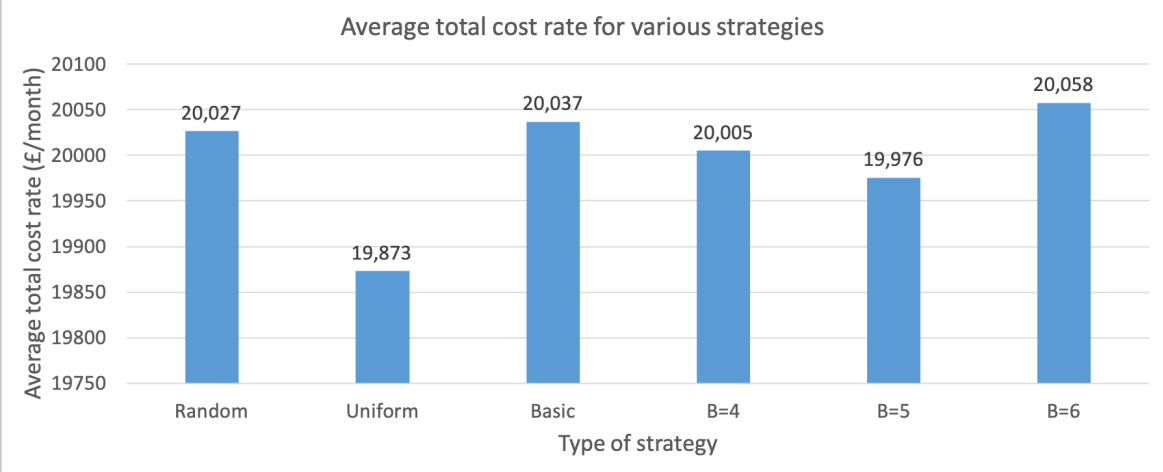


Fig. 6.7 Comparison of the average total cost rate for various strategies

The total cost rate for various strategies can be found in Figure 6.7. Interestingly, when the integer constraint is removed which adds more flexibility to the model, the performance has not been improved. Recall that while in Scenario 1 Section 6.3.2 the best practice of the proposed model leads to a reduction of 1.34% in total cost rate while benchmarked against

uniform workload allocation, this is not the case in Scenario 2. Though the lowest total cost rate achieved by the proposed model, at £19,976/month, outperforms random allocation and the basic version of the model, it fails to beat the £19,873/month given by uniform allocation. When the difference in total cost rate between various strategies and uniform workload allocation is divided into two components - maintenance-related and penalty cost rate, as shown in Figure 6.8, it becomes clear that, the same as Scenario 1, the main contributor to such difference is maintenance-related costs. It is worth noting that indeed the penalty cost rate is lowest for $B = 6$ among the three cases in Scenario 2.

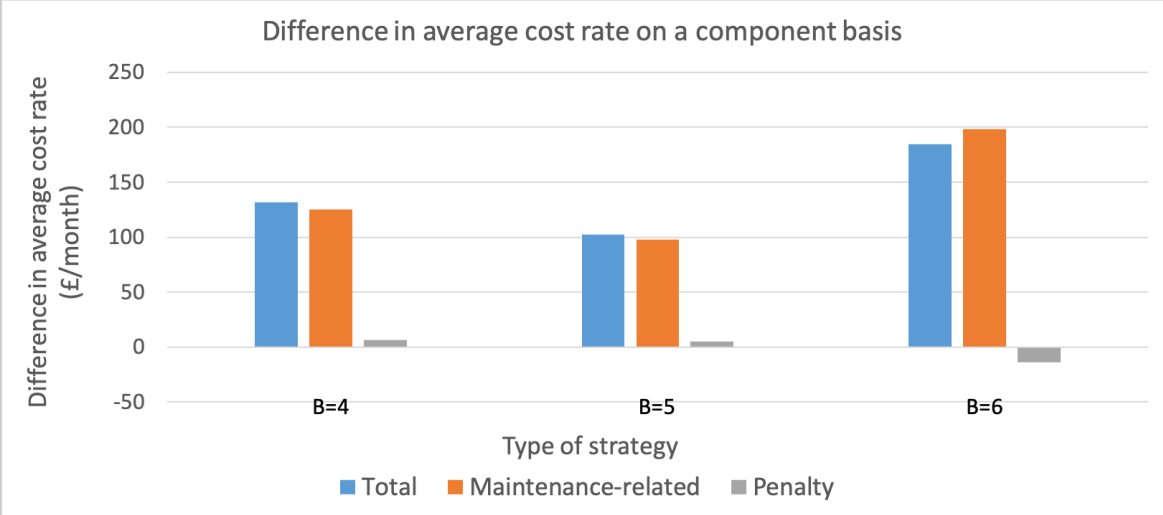


Fig. 6.8 Difference in the average cost rate for various strategies on a component basis

In order to find out why the proposed model fails to generate better solutions with a larger candidate pool in Scenario 2, which means the model should give solutions at least as good as that found in Scenario 1 (e.g., degenerate to [0, 1]), we will need to first briefly review how the model works, and then understand the impact of the efficiency of the optimisation algorithm adopted here on the model performance.

The model, although 'aims' to minimise the total cost rate over the life-cycle of machines, makes decisions about relatively short-term, which spans from the present until the latest next PM to be conducted. Based on the current knowledge, the short-term optimal solution is indeed a partial workload rather than [0, 1]. However, once the system is run using that optimal partial workload, as time progresses, the real deterioration takes on, and the model recalculates the next workload allocation, which again according to its knowledge is a partial workload. Even if the model is given [0, 1] to start with at the first epoch, it will later for sure divert from this solution as it is not the optimal based on its current knowledge. As it now has been clear why the model did not recommend [0, 1] in Scenario 2, we move to explain

why the model has failed to generate solutions better than the traditional strategies. Note that the Genetic Algorithm used is a heuristic technique which does not guarantee global optimal and can get trapped in different local optimum from time to time, especially now it is dealing with more decision variables than just two as presented in Section 5.6. This means that there tends to be more fluctuations in the workload allocation recommended by the model for consecutive time intervals. It can indeed be shown that without the on/off restriction, the fluctuations in the solutions found are in fact exacerbated. Since we are trying to compare the relative stability of the solutions obtained with and without the binary constraint, variance is no longer a suitable metric.

To illustrate this point, consider two workload allocation series having the same average 0.5 - [1, 1, 1, 0, 0, 0] and [0, 0.5, 1, 0, 0.5, 1]. The two series have variance of 0.25 and 0.17, respectively. However, it does not make sense to perceive the second series as more stable than the first one - the workload changes only once from 1 to 0 for the first series whereas it never stays the same between any consecutive intervals in the second series. Here the comparison will be based on the absolute value of the change in load ratio between consecutive intervals. The average of such changes of individual vessels in various strategies are laid out in Table 6.6. At the individual level, the workload assigned to all vessels in Scenario 2 tend to fluctuate more frequently and with much more significant magnitudes than its counterpart in Scenario 1. Note that in Scenario 1 for all B values, the workload assigned to vessel 4-7 has stayed constant almost all the time. At the fleet level, the difference can be as huge as 55 (0.111/0.002) times in the case where $B = 4$. Recall the discussion in Section 5.6.2 that the relative stability has a positive impact in workload allocation, it is not surprising to see the underperformance of the proposed model in Scenario 2. From the above discussion we can suspect that if the workload allocation is forced to fluctuate only reasonably around the actual global optimal from the first epoch, it is possible for the proposed model to outperform the traditional strategies. This will be included in Chapter 7 as one of the directions for future research.

6.4.3 Case Summary and Discussion

In this case example, the on/off flow constraint is released. It is instead assumed that the rate of effluent water flow can be any real number between 0 and the maximum capacity of a single vessel. Despite signs that the proposed approach indeed behaves as it should - shifting more workload from newer vessels to older ones and fighting to mitigate production losses as B increases, something counter-intuitive has also happened. While it is expected that the added flexibility will enable better model performance, the proposed model fails to beat neither uniform allocation nor any of its counterparts in Scenario 1. This has been attributed

Table 6.6 Average change in workload allocated between consecutive intervals

Case	Vessel No.							Fleet average
	1	2	3	4	5	6	7	
S1: $B = 4$	0.007	0.000	0.000	0.000	0.000	0.000	0.007	0.002
S1: $B = 5$	0.106	0.085	0.035	0.000	0.000	0.000	0.000	0.032
S1: $B = 6$	0.063	0.070	0.007	0.000	0.000	0.000	0.000	0.020
S2: $B = 4$	0.127	0.113	0.100	0.107	0.100	0.111	0.118	0.111
S2: $B = 5$	0.136	0.099	0.093	0.088	0.111	0.098	0.117	0.106
S2: $B = 6$	0.144	0.113	0.116	0.108	0.116	0.105	0.117	0.117

to a lower level of stability in workload allocation in Scenario 2, which is a consequence of the heuristic optimisation algorithm failing to stick to the global optimal solution and getting trapped in local optimum from time to time. The results obtained in this case example are very insightful in that they have proved that by introducing reasonable stability into the optimal solutions, the model performance could be further improved. This sheds light on a potential future research on developing a semi-dynamic integrated model, the details of which will be discussed in Chapter 7.

6.5 Chapter Summary

This chapter has demonstrated the usefulness of the proposed approach in real industrial settings. First, the process of converting industrial data into model inputs has been elaborated. Then the model has been applied to a fleet of seven vessels that are used to treat effluent water in a refinery. Specifically, two scenarios have been studied: 1) on/off flow control as the way it is in the refinery; and 2) continuous flow control where the binary constraint has been removed. The key findings from the two examples are summarised as follows:

- It is evident in Scenario 1 that by actively exploiting the interrelationship between workload and maintenance, additional saving could be achieved in both maintenance costs and penalty costs resulted from loss of production for the fleet of vessels considered.
- One of the factors that affect whether saving in maintenance costs or penalty costs will be the bigger contributor to performance improvement is the relative magnitudes of the two components, which turns out to be maintenance costs in the case examples.
- While having more flexibility in its choice of workload allocation, the proposed model fails to outperform the traditional strategies in Scenario 2. It can be concluded by

comparing the model performance in the two scenarios and referring back to the numerical examples presented in Chapter 5 that the higher level of fluctuation in workload allocation in Scenario 2 is held accountable for the unsatisfactory performance of the model. This is largely due to the fact that the heuristic optimisation algorithm does not stick to the global optimal solution and can get trapped in local optimum from time to time. It is therefore recommended that future research be conducted to improve the optimisation algorithm adopted. Furthermore, it is important to ensure a certain level of stability in the solution-seeking process of the model, which will be discussed in Chapter 7.

This chapter has brought into the theoretical model an industrial element and demonstrated a streamlined process of the actual implementation of this approach to optimise the performance of assets in typical parallel configuration.

Chapter 7

Conclusions and Future Research

7.1 Introduction

This chapter aims to provide a summary of this research project. The research questions and the approach taken to address them are first reviewed. Then the key findings, implications, and limitations of this research are presented. The chapter is closed with some recommendations for future research.

7.2 Summary of Research

This research was dedicated to developing a dynamic integrated decision-making model to improve the system-level performance of a fleet of parallel assets. The aim of the model was to realise the potential benefits, mainly in the form of lower maintenance costs and reduced penalty costs incurred due to loss of production, by simultaneously optimising workload allocation and condition-based maintenance threshold. The model was developed on the following basis: 1) an asset is likely to degrade faster when a higher workload is assigned to it, which implies that the asset owner has active control over the degradation progress of an asset; and 2) in cases where the production system consists of a fleet of parallel assets, workload can be shifted among the assets based on their degradation levels, current demand for production, the requirements for future system availability, and the inter-dependency between these factors.

A detailed review was conducted on related research work and case studies in the academic literature, as well as a series of exploratory case studies with companies in the manufacturing domain, which were elaborated in Chapter 2 and Chapter 3, respectively. The literature review indicated that neglecting the workload-dependent degradation behaviour

could lead to sub-optimal maintenance decisions, and that there has been growing research interest in formulating mathematical models that explicitly tie together how an asset degrades with the type of task or amount of workload it takes on. It was also revealed in the analysis that it is not feasible to gain a holistic view of the system dynamics if the various factors involved in a manufacturing process are treated independently. Despite the fact that workload and condition-based maintenance are both closely linked to asset degradation, few attempts have been made to explore the potential benefits of integrated decisions of the two for an asset fleet. The exploratory case studies indicated that in order for the practitioners to make a conscious decision on either workload allocation or maintenance arrangement, it is important to understand how the system-level performance is affected by an ad-hoc change at the individual asset level both in the short and long term.

The decision-making model was implemented in an agent-based system involving two types of agents - 1) machine agents that reside within each individual machine; and 2) a coordinator agent that oversees the entire system. The two constitutive components of the integrated decision-making model were - 1) a workload-dependent condition-based maintenance optimisation model at the asset level implemented through a machine-agent; and 2) a workload allocation strategy at the system level implemented by a coordinator agent. The individual asset model was developed in Chapter 4 that seeks to minimise the expected maintenance cost rate by calculating the optimal condition-based maintenance threshold for any given workload, and feeds the solution together with other relevant information to the coordinator. For the purpose of analysis, the degradation of assets was modelled as a Gamma process, which is suitable for characterising continuously accumulating damage or faults.

The second constitutive component, the system-level workload allocation strategy adopted by the coordinator was presented in Chapter 5. The coordinator seeks to minimise the expected total cost rate of the entire fleet by reaching an optimal balance between maintenance costs and penalty for production losses, as well as a trade-off between short-term and long-term system performance. Numerical examples were used to demonstrate the rationale and mechanism behind the decision-making process of the coordinator. The performance of the proposed strategy was benchmarked against that of traditional uniform, random, and more-on-new-machine allocation strategy. The various factors affecting the margin by which the proposed strategy outperforms the traditional strategies were also explored in this chapter. The results also showed the condition under which the proposed approach was likely to bring more significant benefits, so that asset owners can make decisions on whether to adopt this model based on their own asset characteristics and production needs. Finally the proposed integrated decision-making model was applied in a real industrial context. The case studies

were presented in Chapter 6 for the purpose of demonstrating the practical applicability of the proposed approach.

7.3 Key Findings

In this section, a recap is given on how the proposed decision-making model answers the research questions identified in Chapter 1 and Chapter 2, followed by a discussion of the key research findings.

7.3.1 Recap of Research Questions

Research Question 1: What are the constitutive components for a dynamic optimisation model for joint decision making of CBM threshold and workload allocation for an asset fleet?

In Chapter 4 the two constitutive components of the decision-making model were identified to be 1) a condition-based maintenance optimisation model for individual assets; and 2) a system-level workload allocation strategy that utilises relevant intermediate results calculated by the first component. Such a structure is adopted for the following reasons: 1) existing approaches mostly seek the steady-state solution, which lacks the flexibility of dynamically adjusting the workload assigned to asset when additional benefits could be realised by doing so, and 2) one of the challenges faced by the industry concerns the impact of individual asset performance on system-level performance, an entity at a higher level is needed to collect necessary information and establish a comprehensive understanding of the fleet.

Research Question 2: How can we quantify the impact of workload on the degradation behaviour of individual assets, and how can such knowledge be used to set the most appropriate CBM threshold?

The discussion in Chapter 2 revealed that for discrete stochastic degradation models, the effect of workload is reflected in the scale of transitive probability between states, whereas for continuous degradation it is accounted for by the different choice of distribution parameters. As shown in Chapter 4, Gamma Process can be adapted to characterise the asset degradation process, where the scale parameter can be used to model degradation as a function of workload and a sensitivity factor that quantifies the degree to which the degradation process is affected by workload. Based on the workload assigned and the current condition of asset, the CBM threshold is chosen in order to minimise the expected remaining maintenance costs to be incurred averaged over the expected life cycle of asset. An optimisation algorithm was designed to solve the machine-level optimisation problem and the optimality of the solution

found by the algorithm was mathematically proved. The innovative feature of the model, which is being able to base its decision on the costs to be incurred in view of its current state, allows for dynamic optimisation of the CBM threshold under varying production condition. Results and discussions in Chapter 4 revealed that the impact of workload on CBM decisions is captured by the model via trade-offs between the lengthened expected life cycle and the increased shock probability induced by a higher threshold.

Research Question 3: How can we quantify the impact of a specific workload allocation on maintenance and production at the system-level, and what type of information is needed in order to do so?

This research question corresponds to the formulation of the second constitutive component, which is the workload allocation strategy at the coordinator level as presented in Chapter 5. The discussion in Chapter 5 indicates that for a specific workload allocation, the asset decides on its own the optimal CBM threshold, which also determines the maintenance cost rate over its life cycle. The maintenance cost rate at the system-level is thus the sum of that of each individual asset. In the proposed model, the impact on system-level production is divided into two parts: 1) in the short-term, the system will either be able to meet the demand for the current time period or instantly incurs penalty costs resulted from insufficient production; 2) in the long-term, the future degradation status and time window for maintenance tasks of the asset fleet is largely determined by the workload allocation, based on which the future system availability can be deduced. It was also revealed in Chapter 5 that the maintenance cost rate incurred by each individual asset as an approximated function of the workload needs to be provided to the coordinator in order to calculate the first part. Furthermore, information that generates insights into how the expected time till the next preventive maintenance or replacement of an asset varies with workload should also be available to the coordinator in the form of an approximated function. The aim of the coordinator is then, to reach the optimal trade-off between the maintenance cost, the present penalty cost, and the future penalty cost.

7.3.2 Research Findings

This research project seeks to explore the possibility of generating additional benefits to asset owners by exploiting the intrinsic interconnections between workload and condition-based maintenance arrangement. The approach taken is a simultaneous optimisation of the workload allocation and condition-based maintenance threshold for a fleet of parallel assets. The analysis in Chapter 4, 5, and 6 indicated that considering the workload-dependent degradation behaviour of assets reduces both maintenance costs and penalty costs incurred by loss of

production. The key findings can be categorised into two main areas: 1) factors that affect the usefulness of the proposed approach in maintenance cost saving; and 2) factors that affect the performance of the proposed approach in penalty cost saving while benchmarked against traditional strategies. Note that the proposed approach almost always outperforms random and more-on-new-machine workload allocation by a large margin, the findings in the second type of factors are more concerned about comparison with uniform workload allocation.

Factors related to maintenance cost saving

The saving in maintenance costs can come from two sources: 1) using workload-dependent condition-based maintenance thresholds rather than a uniform one for any workload; and 2) appropriately allocating workload between assets to minimise maintenance costs.

1. **Sensitivity of asset degradation to workloads**
   The more sensitive the degradation behaviour of an asset is to the workload assigned, the more cost saving can be realised at the individual asset level. The analysis in Chapter 4 indicated that workload has a strong impact on asset degradation, and there tends to be a significant difference in the optimal CBM threshold when an asset takes on different levels of workload. Treating such degradation behaviour of assets as independent, therefore, leads to maintenance actions that diverge severely from the optimal plan, incurring costs that could have been saved with workload-dependent optimisation models (refer to Section 4.7.3).

2. **Relative magnitude of maintenance costs and penalty costs for production losses**
   It is more likely to realise saving in maintenance-related costs while penalty costs take up a relatively smaller proportion of the total costs. While the opposite is true, the potential saving in maintenance costs is sacrificed for a more significant reduction in penalty costs (refer to Section 6.3.2).

Factors related to penalty cost saving

- Objective physical characteristics of the system

  1. **Heterogeneity of assets**
     Though heterogeneity in assets can be understood as assets belonging to different categories, here specifically, we would restrict our discussion of heterogeneity to the difference in the initial condition and degradation behaviour of assets in the same fleet. An asset fleet consists of multiple assets belonging to the same category that share similar technical features and missions, as defined by Table 2.1. The performance of the proposed model is summarised in the matrix

given in Figure 7.1. The model is of greater use in Case 1 where the initial condition and the degradation behaviour of assets are both fairly homogeneous; and Case 4 where the initial condition and the degradation behaviour of assets are both fairly heterogeneous. Since one of the key points to lower system-level risk of production loss is to keep the maintenance time windows of assets far apart from each other, the additional benefits brought by the proposed model will not be that significant in Cases 2 and 3, where gaps will result naturally in-between these windows even with simple uniform workload allocation (refer to Section 5.7).

	Homogeneous initial degradation level	Heterogeneous initial degradation level
Homogeneous degradation behaviour	Case 1 ✓	Case 2 -
Heterogeneous degradation behaviour	Case 3 -	Case 4 ✓

Fig. 7.1 Summary of model performance with asset heterogeneity

2. **Penalty cost for a unit of production loss**
   The proposed model has a more obvious advantage over traditional strategies when loss of production is expensive. This is a direct consequence of the model deliberately taking into consideration both the instant loss and the potential future loss incurred by a specific workload allocation. A side effect of a larger penalty cost has also indirectly contributed to the improved model performance. While loss of production is expensive, the model is motivated to direct more attention to ensuring that the long-term penalty-related component stays low. Since small shifts of workload might result in significant changes in the long-term component, a relatively stable workload is maintained at consecutive decision-making intervals, which, as discussed in Section 5.6.2, leads to better model performance (refer to Section 5.7).
3. **Redundancy of the system**
   The redundancy of the system has a negative impact on the amount of saving in penalty costs that can be potentially realised via the proposed approach. Here

the meaning of the redundancy of the system goes beyond simply the number of assets that can be down before any production loss is generated. It is extended to include the probability as well as the consequence of production losses. For instance, although both systems analysed in Section 5.4.1 and Section 6.3.2 start to suffer from loss of production once two machines are taken offline, the former in general has a higher probability of losing two machines at the same time as well as a more severe consequence. While the former delivers nothing, the latter is still able to meet 5/6 of its production demand (refer to Section 5.6.2 and Section 6.3.2).

- Subjective choice of non-physical model parameters (buffer size B)
  Though the buffer size B has no position in describing the physical characteristics of the production system, it plays a critical part in the decision-making model. The two important functions of B are to: 1) adjust the amount of attention that the model directs to ensuring a small numerical value is taken by the future penalty component; and 2) control the degree of fluctuations in workload allocation between consecutive decision-making epochs. For a specific problem setting, an optimal B exists that maximises the saving in penalty cost at a maintenance cost comparable with traditional strategies. A B which is too small does not assign enough weight to the potential future production losses in the system-level objective function and cannot provide the stabilising effect required, whereas a B too large renders the decision-making model indifferent to various workload allocations when it comes to future penalty costs (refer to Section 5.6.2 and 5.8). Guidance on how future research can be carried out to establish a systematic approach to find an appropriate B will be given in Section 7.7.

7.4 Novelty of Research

The novelty of this research is summarised as follows:

- The use of Gamma process to model the workload-dependent degradation behaviour of an asset, where the scale parameter of Gamma distribution is set to be a function of the workload assigned to and the maximum capacity of an asset.
- The design of a condition-based maintenance optimisation objective function based on the expected future maintenance costs to be incurred for the remaining life-cycle of an asset.

- The design of an agent-based approach for dynamic integrated optimisation of workload allocation and condition-based maintenance threshold for a fleet of parallel assets.

7.5 Contributions of Research

This section aims to summarise the contributions made by this research to the academic domain as well the industrial practice.

7.5.1 Academic Contributions

- **A model of workload-dependent continuous degradation process**
  Although there have been studies devoted to characterising the workload/task-dependent degradation behaviour of assets, they are either in the form of discrete-state models or assume simple linear relationships between the incremental degradation and the workload assigned to an asset. This research has formally proposed a workload-dependent degradation model based on Gamma process for continuous-state situation. This approach can be generalised to add workload-dependent traits to other stochastic processes that are widely-adopted for describing asset degradation behaviour.

- **A CBM optimisation model based on the life-cycle of an asset**
  This is one of the two constitutive components of the overall approach developed in this study. Unlike most maintenance models that seek to optimise a steady-state performance metric such as long-run cost rate or long-run system availability, this study has proposed a model that aims to minimise the total maintenance cost rate over the life-cycle of an asset based on the current condition of and the workload assigned to the asset. Such characteristics of the maintenance model allows decisions to be updated and refined if situation changes or additional information of the production system is obtained both at the individual asset level and the system level.

- **A dynamic workload allocation strategy for parallel assets**
  This is the other constitutive component of the overall approach. The proposed strategy explicitly takes into account the interactions between maintenance and production, and enables decisions to be made dynamically and concurrently on condition-based maintenance threshold and workload allocation among a fleet of parallel assets. The inclusion of the impact of workload allocation decision on long-term maintenance costs, short-term production losses, and long-term production losses helps to improve system-

level performance by reducing total maintenance-related costs as well as penalty costs resulted from loss of production.

- **Basis for semi-dynamic decision-making models for CBM threshold and workload allocation**
  In view of one particular finding from this research project which digs into the stabilising effect of buffer size B on model performance, the study has provided the basis for possible new research directions of semi-dynamic decision-making models. A semi-dynamic model is expected to retain the ability of the proposed model to exploit the interrelationships of maintenance and production, but also be able to stick to a relatively stable workload allocation by only updating its decision when noticeable changes take place (e.g. a preventive maintenance is carried out on one machine which significantly improves its condition).

7.5.2 Industrial Contributions

Before going into the details of the industrial contributions, it is worth mentioning a few real-world situations (apart from the SET vessels studied in Chapter 6) where the proposed approach can be of use so as to set the problem in a practical context. For instance, the parallel chillers in HVAC systems are used to lower the temperature of cooling water. The proposed approach can help with optimise the amount of cooling water to be directed to each chiller based on the condition of the chillers, which might lead to both cost and energy saving as more degraded chillers are not as efficient as newer ones. The escalators in metro stations are another use-case for the proposed model. Some of the escalators need to be switched off at off-peak times, and this model can assist operators in deciding which ones to be switched off. In future smart motorways, the model can be useful for lane management in ways in that it could make suggestions on which lanes a heavy/light vehicle should be directed to while considering the degradation level of the lanes.

It can be seen from the examples mentioned that the systems to which the proposed approach is applicable tend to possess the following characteristics: 1) consisting of an asset fleet as defined in Table 2.1; 2) having certain level of redundancy so that there is room for shifting workload; 3) having relatively constant demands that need to be fulfilled on time (backlogs are rarely allowed). For organisations that have systems with the aforementioned characteristics, the industrial implications of this study can be classified against the organisational decision-making context in three levels: strategic, tactical, and operational.

- **Strategic: Guidance on whether to invest in workload control capabilities**
  One of the requirements of implementing the proposed model in an industrial context is

that the assets considered need to have workload control mechanism, and investments need to be made if such mechanism is not readily available, such as adopting variable flow valves in the case of effluent water treatment vessels. This research project has provided an approach to evaluate whether it is economically beneficial for asset owners to make this type of investment by comparing the additional costs incurred and the potential saving in both maintenance and penalty costs. Practitioners can also draw insights from this study on factors that impact the usefulness of the proposed approach.

- **Tactical: Approach for evaluating the long-term impacts of sudden events**
  The model formulated in this research mainly follows the rationale that decisions should be made with a forward-looking attitude based the situation at hand. This trait of the model allows practitioners to evaluate the long-term impacts of sudden events such as an increase in demand, a decrease in the number of redundant machines, and a change in the cost of maintenance actions.

- **Operational: Assistance on decision-making of workload allocation and CBM thresholds**
  The proposed model can provide assistance to operation and maintenance managers to make decisions on the optimal combination of workload allocation and maintenance plans for assets in the production system.If automatic condition monitoring system and workload control mechanism are also implemented, the shifting of workload between assets can be automated based on the information received, which might lead to saving in labour costs.

7.6 Limitations

This section discusses the limitations regarding both the formulation of the mathematical model as well as the analysis of modelling results presented in this study. These limitations are discussed in details in the rest of the section.

7.6.1 Limitations of the mathematical model

- **Modelling assumptions**
  A couple of assumptions have been made in the process of formulating the mathematical model. These assumptions might only be partially true in reality. For instance, it is assumed that the time taken to shift workload between assets is negligible and that no costs are associated with such re-allocation. In reality, however, it is possible that

neither of the conditions will hold. Take parallel stations in the job shop as an example, it usually requires man power to move components and subassemblies around, which does not only take time, but also leads to labour costs. Further work is needed to incorporate these factors into the model to eliminate this limitation.

- **Choice of the model parameter B**
  It is found in the previous analysis that one factor that has noticeable influence on the performance of the proposed model is the choice of buffer size B. In all the numerical examples and case studies presented in this thesis, the most appropriate B has been found through trial and error using simulation. As simulation is time-consuming and software programmes are not always readily available for running simulations, a systematic and efficient approach is needed to decide on the optimal value of B to be used if the model were to be implemented in a real industrial context. One promising way to go might be an analysis into the relative magnitude of the penalty costs due to loss of production and the total maintenance-related costs. Details regarding of this approach are given in Section 7.7.

- **Computational complexity**
  Though adopting the proposed approach leads to time-saving effects compared with centralised approaches that place all computational burden on one entity, such effects tend to diminish as the number of parallel units becomes too large. This is due to the run-time of the coordinator-level decision making having a factorial relationship with the number of machines M. Reasonable as it might be, the range of M to which this approach is applicable can still be considered as a limitation of the model proposed in this research.

7.6.2 Limitations of the analysis of modelling results

Apart from limitations related to the mathematical model itself, limitations also exist in the analysis of modelling results.

- **The impacts of relative stability**
  One speculation of this study is that relative stability in the solutions is likely to improve the performance of the model, and such stability could be realised by an appropriate choice of B. The ultimate rationale that has led to this observation is yet to be explored and further analysed. How such analysis can be carried forward will be discussed in the section regarding recommendations for future research - Section 7.7.

- **Capital costs of enabling workload control mechanism among assets**
  Neither the model nor the analysis has taken into consideration the capital costs needed to enable the production system to control workloads assigned to each asset. However, the potential saving given that such mechanism is already there can be inferred from the model, which can be used to decide whether it is advisable to make this investment.

- **Fluctuations in production demand**
  The analysis of the model has been carried out in a scenario where the production demand stays constant. In reality, production demands tend to fluctuate for various reasons. Further analysis is yet to be conducted to see whether the proposed model can cope with varying demands.

- **Duration of maintenance actions**
  It can be inferred from the way penalty costs are generated that the duration of maintenance actions is likely to have reasonable impacts on the usefulness of the proposed model. This limitation can be eliminated by running more simulations with a wide range of values for maintenance duration. It can also be implicitly addressed by observing how the model performance varies with different buffer sizes.

7.7 Recommendations for Future Research

This research study is carried out to address the problem of simultaneously optimising the condition-based maintenance threshold and workload allocation for a fleet of assets. Moreover, the mathematical approach developed here serves as the foundation for further research in the area of integrated production and maintenance decision making. For reference purposes, we have identified the following directions in which the proposed approach can be extended:

- **Developing a systematic approach for the choice of B**
  Since the value of B has a significant impact on the performance of the proposed model, it is very important that a systematic approach is at hand via which a proper B value can be chosen. Here a potential approach is proposed, which can be further verified and explored in the future. The two guiding rules followed the approach are: 1) relative stability should be observed in the workload allocation solution found by the model; and 2) a larger B value is preferred while loss of production is expensive. Though an operator might not have the benefits of seeing simulation results before implementing this model, an appropriate B value can be found through trial and error. Specifically, the model can first be implemented with a reasonable guess for B. The operator can

then observe the workload allocation produced by the model for a number of time intervals. If too much fluctuation exists in the workload allocation for consecutive intervals, the operator can increase B and repeat the process until he is comfortable with the variance in the workload allocation recommended by the model.

- **Introducing more decentralisation into the decision-making process**
  The proposed approach is essentially a two-step decision-making process, where the first step is done at the individual asset level and the second at the coordinator level. The computational burden is still rather heavy, however, if the number of assets involved in the system keeps growing. Furthermore, in view of the development of the Internet of Things, it is thus recommended here that more decentralisation ought to be introduced into the second step of decision-making process. Some interesting work that has been done along this line can be taken as a starting point, such as the machine-to-machine communication and knowledge sharing mechanism adopted by Palau et al. [74], and the well-known bidding mechanism in Contract Net Protocol [87].

- **Extrapolating the model to deal with more complex configurations**
  The focus of the model developed in this study is on a fleet of parallel assets. It is hardly the case that these assets will stand in isolation in a real manufacturing plant. Extrapolating the model to deal with more general configurations such as a serial-parallel system is likely to improve the applicability of the proposed approach. The model can also be extended to incorporate more than one type of task.

- **Turning the proposed dynamic model into a semi-dynamic one**
  This recommendation is closely related to the benefits that could be potentially realised by reducing the fluctuations in the solution proposed by the model. In the current approach, the decision is updated at pre-defined time epochs given the situation at hand, which, based on the previous analysis, is not necessarily the best thing to do. This can be improved by turning the proposed model into a semi-dynamic one via one of the following two routes: 1) to develop a method to choose the optimal decision-making epoch; and 2) to come up with a criteria to decide whether to stick to the previous solution at consecutive time epochs.

- **Incorporating workload shifting costs into the model**
  As mentioned in Section 7.6.1, both the time and costs taken to shift workload among assets are assumed to be negligible. It is however hardly the case in reality. Here we recommend that the model should be extended to incorporate the time and costs associated with workload shifting to more closely represent the reality. For instance, it

can be assumed that both the time and costs are functions of the amount of workload being shifted. This can be linked back to the previous point of making the model semi-dynamic, as it encourages the model to maintain a relatively stable workload allocation across time.

- **Exploring ways to improve the efficiency of the optimisation algorithm**
  The analysis in Section 6.4.2 has shown that the Genetic Algorithm used by the model is inefficient fail to stick to the global optimal solution and can get trapped in local optimum from time to time when the number of decision variables gets larger. This has resulted in a lower level of stability in workload allocation and thus underperformance of the proposed model in Scenario 2. Therefore, it is recommended an optimisation algorithm be used, which can be achieved via various approaches such as 1) adopting advanced techniques to reduce the chance of the algorithm being trapped in local optimum; 2) using commercial optimisation software.

This research project proposed a decision-making approach to concurrently optimise workload allocation and condition-based maintenance threshold for a fleet of assets. With the above extensions recommended on the basis of the work done in this study, it is possible to deliver more comprehensive approaches that further enhance the benefits that can be brought to asset owners by exploiting the intrinsic relationship between production and maintenance.

References

[1] Ahmad, R. and Kamaruddin, S. (2012). An overview of time-based and condition-based maintenance in industrial application. *Computers and Industrial Engineering*, 63(1):135–149.

[2] Alaswad, S. and Xiang, Y. (2017). A review on condition-based maintenance optimization models for stochastically deteriorating system. *Reliability Engineering and System Safety*, 157:54–63.

[3] AlDurgam, M. M. and Duffuaa, S. O. (2013). Optimal joint maintenance and operation policies to maximise overall systems effectiveness. *International Journal of Production Research*, 51(5):1–12.

[4] Allen, J. and Rabiner, L. (1977). A unified approach to short-time Fourier analysis and synthesis. *Proceedings of the IEEE*, 65(11):1558–1564.

[5] Bahnfleth, W. and Peyer, E. (2004). Varying views on variable-primary flow chilled-water systems. *HPAC Heating, Piping, AirConditioning Engineering*, 76.

[6] Barlow, R. and Proschan, F. (1996). *Mathematical Theory of Reliability*. Society for Industrial and Applied Mathematics.

[7] Ben Ali, M., Sassi, M., Gossa, M., and Harrath, Y. (2011). Simultaneous scheduling of production and maintenance tasks in the job shop. *International Journal of Production Research*, 49(13):3891–3918.

[8] Biswas, A., Sarkar, J., and Sarkar, S. (2003). Availability of a periodically inspected system, maintained under an imperfect-repair policy. *IEEE Transactions on Reliability*, 52(3):311–318.

[9] Bonissone, P. P. and Decker, K. S. (1986). Selecting uncertainty calculi and granularity: An experiment in trading-off precision and complexity. In KANAL, L. N. and LEMMER, J. F., editors, *Uncertainty in Artificial Intelligence*, volume 4 of *Machine Intelligence and Pattern Recognition*, pages 217 – 247. North-Holland.

[10] Bouzidi-Hassini, S. and Benbouzid-Sitayeb, F. (2013). Multi-agent based joint production and maintenance scheduling considering human resources. *2013 5th International Conference on Modeling, Simulation and Applied Optimization, ICMSAO 2013*, pages 0–4.

[11] Bunks, C., Mccarthy, D., and Al-Ani, T. (2000). Condition-Based Maintenance of Machines Using Hidden Markov Models. *Mechanical Systems and Signal Processing*, 14(4):597–612.

[12] Castanier, B., Grall, A., and Bérenguer, C. (2005). A condition-based maintenance policy with non-periodic inspections for a two-unit series system. *Reliability Engineering & System Safety*, 87(1):109 – 120.

[13] Celen, M. and Djurdjanovic, D. (2012). Operation-dependent maintenance scheduling in flexible manufacturing systems. *CIRP Journal of Manufacturing Science and Technology*, 5(4):296–308.

[14] Celen, M. and Djurdjanovic, D. (2015a). Integrated Maintenance Decision-Making and Product Sequencing in Flexible Manufacturing Systems. *Journal of Manufacturing Science and Engineering*, 137(4):1–15.

[15] Celen, M. and Djurdjanovic, D. (2015b). Integrated Maintenance Decision-Making and Product Sequencing in Flexible Manufacturing Systems. *accepted for publication in the ASME Journal of Manufacturing Science and Engineering*, 137(August 2015):1–15.

[16] Chan, G. K. and Asgarpoor, S. (2006). Optimum maintenance policy with Markov processes. *Electric Power Systems Research*, 76(6-7):452–456.

[17] Chandra, A., Ahsan, M., Lahiri, S., Panigrahi, S., Manupati, V. K., and Costa, E. (2017). Degradation modeling to predict the residual life distribution of parallel unit systems on benchmark instances. In *Proceedings of the World Congress on Engineering 2017*, volume 2, London, U.K.

[18] Chen, D. and Trivedi, K. S. (2005). Optimization for condition-based maintenance with semi-Markov decision process. *Reliability Engineering and System Safety*, 90(1):25–29.

[19] Chen, N., Ye, Z.-S., Xiang, Y., and Zhang, L. (2015). Condition-based maintenance using the inverse gaussian degradation model. *European Journal of Operational Research*, 243(1):190 – 199.

[20] Chinnam, R. B. and Baruah, P. (2003). Autonomous diagnostics and prognostics through competitive learning driven hmm-based clustering. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 4, pages 2466–2471.

[21] Chinnam, R. B. and Baruah, P. (2004). A neuro-fuzzy approach for estimating mean residual life in condition-based maintenance systems. *International Journal of Materials and Product Technology*, 20(1-3):166–179.

[22] Choi, G. and Choi, G. (1996). Application of minimum cross entropy to model-based monitoring in diamond turning. *Mechanical systems and signal processing*, 10:615–631.

[23] Chung, S., Chan, F. T., and Chan, H. (2009). A modified genetic algorithm approach for scheduling of perfect maintenance in distributed production scheduling. *Engineering Applications of Artificial Intelligence*, 22(7):1005 – 1014. Distributed Control of Production Systems.

[24] Cua, K. O., Mckone, K. E., and Schroeder, R. G. (2001). Relationships between implementation of TQM , JIT , and TPM and manufacturing performance. *Journal of Operations Management*, 19(6):675–694.

[25] Cui, L., Kuo, W., Loh, H., and Xie, M. (2004). Optimal allocation of minimal amp;amp; perfect repairs under resource constraints. *IEEE Transactions on Reliability*, 53(2):193–199.

[26] Cuthbert, R., Pennesi, P., and Mcfarlane, D. (2008). A case study approach to examing service information requirements. In Barrett, M., Davidson, E., Middleton, C., and J, D., editors, *IFIP International Federation for Information Processing, Information Technology in the Service Economy: Challenges and Possibilities for the 21st Century*, volume 267, pages 383–385. Boston: Springer.

[27] Dalpiaz, G., Rivola, A., and Rubini, R. (2000). Effectiveness and Sensitivity of Vibration Processing Techniques for Local Fault Detection in Gears. *Mechanical Systems and Signal Processing*, 14(3):387–412.

[28] Decker, K. S. (1987). Distributed problem-solving techniques: A survey. *IEEE Transactions on Systems, Man, and Cybernetics*, 17(5):729–740.

[29] Deep, K., Singh, K. P., Kansal, M., and Mohan, C. (2009). A real coded genetic algorithm for solving integer and mixed integer optimization problems. *Applied Mathematics and Computation*, 212(2):505 – 518.

[30] Dieulle, L., BÃ©renguer, C., Grall, A., and Roussignol, M. (2003). Sequential condition-based maintenance scheduling for a deteriorating system. *European Journal of Operational Research*, 150(2):451 – 461.

[31] Ding, C. and Krajewski, L. (1994). A decision model for corrective maintenance management. *The International Journal of Production Research*, 32(6):1365–1382.

[32] Ding, S. H. and Kamaruddin, S. (2014). Maintenance policy optimization - literature review and directions. *International Journal of Advanced Manufacturing Technology*, 76(5-8):1263–1283.

[33] Djurdjanovic, D., Lee, J., and Ni, J. (2003). Watchdog agent - An infotronics-based prognostics approach for product performance degradation assessment and prediction. *Advanced Engineering Informatics*, 17(3-4):109–125.

[34] Do, P., Voisin, A., Levrat, E., and Iung, B. (2015). A proactive condition-based maintenance strategy with both perfect and imperfect maintenance actions. *Reliability Engineering & System Safety*, 133:22 – 32.

[35] Gan, S.-N. (1996). Storage hardening of natural rubber. *Journal of Macromolecular Science, Part A*, 33(12):1939–1948.

[36] Giordani, S., Lujak, M., and Martinelli, F. (2013). A distributed multi-agent production planning and scheduling framework for mobile robots. *Computers & Industrial Engineering*, 64(1):19 – 30.

[37] Godoy, D. R., Knights, P., and Pascual, R. (2018). Value-based optimisation of replacement intervals for critical spare components. *International Journal of Mining, Reclamation and Environment*, 32(4):264–272.

[38] Goldberg, D. E. and Deb, K. (1991). A comparative analysis of selection schemes used in genetic algorithms. volume 1 of *Foundations of Genetic Algorithms*, pages 69 – 93. Elsevier.

[39] Goode, K. B., Roylance, B. J., and Moore, J. (2000). Development of model to predict condition monitoring interval times. *Ironmaking & Steelmaking*, 27(1):63–68.

[40] Guo, C., Wang, W., Guo, B., and Si, X. (2013). A maintenance optimization model for mission-oriented systems based on wiener degradation. *Reliability Engineering & System Safety*, 111:183 – 194.

[41] Hadidi, L. a., Turki, U. M. A., and Rahim, A. (2012). Integrated models in production planning and scheduling, maintenance and quality: a review. *International Journal of Industrial and Systems Engineering*, 10(1):21.

[42] Hao, L., Liu, K., Gebraeel, N., and Shi, J. (2015). Controlling the Residual Life Distribution of Parallel Unit Systems Through Workload Adjustment. *IEEE Transactions on Automation Science and Engineering*, 14(2):1–11.

[43] Hartman, T. (2001). All-variable speed centrifugal chiller plants. *ASHRAE Journal*, 9(43):43 – 53.

[44] Howard, I., Jia, S., and Wang, J. (2001). The Dynamic Modelling of a Spur Gear in Mesh Including Friction and a Crack. *Mechanical Systems and Signal Processing*, 15(5):831–853.

[45] Howard, R. A. (1971). *Dynamic Probabilistic Systems, Vol 1: Markov Models, Vol 2: Semi-Markov and Decision Processes*. Wiley, New York.

[46] Huynh, K., Castro, I., Barros, A., and Bérenguer, C. (2012). Modeling age-based maintenance strategies with minimal repairs for systems subject to competing failure modes due to degradation and shocks. *European Journal of Operational Research*, 218(1):140 – 151.

[47] Iakovou, E., Ip, C. M., and Koulamas, C. (1996). Machining economics with phase-type distributed tool lives and periodic maintenance control. *Computers and Operations Research*, 23(1):53–62.

[48] Ilgin, M. A. and Tunali, S. (2007). Joint optimization of spare parts inventory and maintenance policies using genetic algorithms. *International Journal of Advanced Manufacturing Technology*, 34(5-6):594–604.

[49] Jafari, L. and Makis, V. (2016). Optimal lot-sizing and maintenance policy for a partially observable production system. *Computers & Industrial Engineering*, 93:88 – 98.

[50] Jain, M., Kumar, A., and Sharma, G. (2002). Maintenance cost analysis for replacement model with perfect/minimal repair. *International Journal of Engineering. Transactions A: Basics*, 2.

[51] Jalan, A. K. and Mohanty, A. R. (2009). Model based fault diagnosis of a rotor-bearing system for misalignment and unbalance under steady-state condition. *Journal of Sound and Vibration*, 327(3-5):604–622.

[52] Jardine, A., Banjevic, D., Wiseman, M., Buck, S., and Joseph, T. (2001). Optimizing a mine haul truck wheel motors' condition monitoring program Use of proportional hazards modeling. *Journal of Quality in Maintenance Engineering*, 7(4):286–302.

[53] Jardine, A. K. S., Lin, D., and Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, 20(7):1483–1510.

[54] Jaturonnatee, J., Murthy, D., and Boondiskulchok, R. (2006). Optimal preventive maintenance of leased equipment. *European Journal of Operational Research*, 174(1):201–215.

[55] Jin, L. (2015). Optimal decision procedure for an operation-dependent deteriorating system. *Applied Stochastic Models in Business and Industry*, 31(3):394–404.

[56] Keizer, M. C. O., Flapper, S. D. P., and Teunter, R. H. (2017). Condition-based maintenance policies for systems with multiple dependent components: A review. *European Journal of Operational Research*, 261(2):405 – 420.

[57] Klutke, G.-A. and Yang, Y. (2002). The availability of inspected systems subject to shocks and graceful degradation. *IEEE Transactions on Reliability*, 51(3):371–374.

[58] Le, M. D. and Tan, C. M. (2013). Optimal maintenance strategy of deteriorating system under imperfect maintenance and inspection using mixed inspectionscheduling. *Reliability Engineering & System Safety*, 113:21 – 29.

[59] Liang, W. K., Balcıoğlu, B., and Svaluto, R. (2013). Scheduling policies for a repair shop problem. *Annals of Operations Research*, 211(1):273–288.

[60] Liao, H. and Elsayed, E. A. (2006). Reliability inference for field conditions from accelerated degradation testing. *Naval Research Logistics (NRL)*, 53(6):576–587.

[61] Liao, H., Elsayed, E. A., and Chan, L.-Y. (2006). Maintenance of continuously monitored degrading systems. *European Journal of Operational Research*, 175(2):821 – 835.

[62] Liu, B., Zhenglin, L., Parlikad, A. K., Xie, M., and Kuo, W. (2017). Condition-based maintenance for systems with aging and cumulative damage based on proportional hazards model. *Reliability Engineering & System Safety*, 168:200–209.

[63] Liu, Z., Ma, X., Yang, J., and Zhao, Y. (2014). Reliability modeling for systems with multiple degradation processes using inverse gaussian process and copulas. *Mathematical Problems in Engineering*, 2014:1–10.

[64] Lu, B. (2009). A Review of Recent Advances in Wind Turbine Condition Monitoring and Fault Diagnosis. *Power Electronics and Machines in Wind Applications, IEEE*, pages 1 – 7.

[65] M. Xie, C. D. L. (1995). Reliability analysis using an additive Weibull model with bathtub-shaped failure rate function. *Reliability Engineering and System Safety*, 52(1995):87–93.

[66] Manupati, V. K., Panigrahi, S., Ahsan, M., Lahiri, S., Chandra, A., Thakkar, J. J., Putnik, G., and R., V. M. L. (2019). Estimation of manufacturing systems degradation rate for residual life prediction through dynamic workload adjustment. *Sādhanā*, 30(44):1 – 23.

[67] Marais, K. B. and Saleh, J. H. (2009). Beyond its cost, the value of maintenance: An analytical framework for capturing its net present value. *Reliability Engineering and System Safety*, 94(2):644–657.

[68] Marseguerra, M., Zio, E., and Podofillini, N. (2002). Condition-based maintenance optimization by means of genetic algorithms and Monte Carlo simulation. *Reliability Engineering and System Safety*, 77:151–176.

[69] Mechefske, C. K. and Wang, Z. (2001). Using fuzzy linguistics to select optimum maintenance and condition monitoring strategies. *Mechanical Systems and Signal Processing*, 15(6):1129 – 1140.

[70] Moubray, J. (1995). Reliability-centred maintenance. *Fuel and Energy Abstracts*, 36(4):304.

[71] Moudani, W. E. and Mora-Camino, F. (2000). A dynamic approach for aircraft assignment and maintenance scheduling by airlines. *Journal of Air Transport Management*, 6(4):233 – 237.

[72] Nguyen, K.-A., Do, P., and Grall, A. (2015). Multi-level predictive maintenance for multi-component systems. *Reliability Engineering & System Safety*, 144:83–94.

[73] Oppenheimer, C. H. and Loparo, K. A. (2002). Physically based diagnosis and prognosis of cracked rotor shafts. In *Poceedings of SPIE*, 4733, pages 122–132, Orlando. FL.

[74] Palau, A. S., Bakliwal, K., Dhada, M. H., Pearce, T., and Parlikad, A. K. (2018). Recurrent neural networks for real-time distributed collaborative prognostics. In *2018 IEEE International Conference on Prognostics and Health Management (ICPHM)*, pages 1–8.

[75] Park, D. H., Jung, G. M., and Yum, J. K. (2000). Cost minimization for periodic maintenance policy of a system subject to slow degradation. *Reliability Engineering and System Safety*, 68(2):105–112.

[76] Peng, H. and van Houtum, G.-J. (2016). Joint optimization of condition-based maintenance and production lot-sizing. *European Journal of Operational Research*, 253(1):94 – 107.

[77] Petchrompo, S. and Parlikad, A. K. (2019). A review of asset management literature on multi-asset systems. *Reliability Engineering and System Safety*, 181:181–201.

[78] Pham, H. and Wang, H. (1996). Imperfect maintenance. *European Journal of Operational Research*, 94(3):425 – 438.

[79] Radhoui, M., Rezg, N., and Chelbi, A. (2010). Joint quality control and preventive maintenance strategy for imperfect production processes. *Journal of Intelligent Manufacturing*, 21(2):205–212.

[80] Rasmekomen, N. and Parlikad, A. K. (2013). Maintenance optimization for asset systems with dependent performance degradation. *IEEE Transactions on Reliability*, 62(2):362–367.

[81] Rausch, M., Member, S., and Liao, H. (2010). Joint production and spare part inventory control strategy driven by condition based maintenance. *IEEE Transactions on Reliability*, 59(3):507–516.

[82] Rehorn, A. G., Jiang, J., and Orban, P. E. (2005). State-of-the-art methods and results in tool condition monitoring: A review. *International Journal of Advanced Manufacturing Technology*, 26(7-8):693–710.

[83] Safaei, N., Banjevic, D., and Jardine, A. K. S. (2011). Workforce-constrained maintenance scheduling for military aircraft fleet: a case study. *Annals of Operations Research*, 186(1):295–316.

[84] Sharma, A., Yadava, G., and Deshmukh, S. (2011). A literature review and future perspectives on maintenance optimization. *Journal of Quality in Maintenance Engineering*, 17(1):5–25.

[85] Siddique, A., Yadava, G. S., and Singh, B. (2003). Applications of artificial intelligence techniques for induction machine stator fault diagnostics: review. *Diagnostics for Electric Machines, Power Electronics and Drives, 2003. SDEMPED 2003. 4th IEEE International Symposium on*, pages 29–34.

[86] Smith, D. J. (2013). Power-by-the-hour : the role of technology in reshaping business strategy at Rolls-Royce. *Technology Analysis & Strategic Management*, 25(8):987–1007.

[87] Smith, R. (1980). The contract net protocol: High-level communication and control in a distributed problem solver. *IEEE Transactions on Computers*, 29(12):1104–1113.

[88] Sobczyk, K. (1989). Stochastic models for fatigue damage of materials. *Mathematical and Computer Modelling*, 12(8):1046.

[89] Sobczyk, K. and Spencer, J. B. (1992). *Random Fatigue from Data to Theory*. Academic Press, San Diego.

[90] Starr, A. G. (1997). A structured approach in the selection of condition based maintenance. In *Proceedings - 5th International Conference on FACTORY 2000*, 435, pages 131–138, Cambridge, UK. IET.

[91] Tian, Z. and Liao, H. (2011). Condition based maintenance optimization for multi-component systems using proportional hazards model. *Reliability Engineering and System Safety*, 96(5):581–589.

[92] Upasani, K., Bakshi, M., Pandhare, V., and Lad, B. K. (2017). Distributed maintenance planning in manufacturing industries. *Computers & Industrial Engineering*, 108:1 – 14.

[93] Van Horenbeek, A., Scarf, P. A., Cavalcante, C. A. V., and Pintelon, L. (2013). The Effect of Maintenance Quality on Spare Parts Inventory for a Fleet of Assets. *IEEE Transactions on Reliability*, 62(3):596–607.

[94] Van Noortwijk, J. M. (2009). A survey of the application of gamma processes in maintenance. *Reliability Engineering and System Safety*, 94(1):2–21.

[95] Wang, L. and Hong, T. (2012). Modeling and simulation of hvac faulty operations and performance degradation due to maintenance issues. In *ASim 2012 - 1st Asia conference of International Building Performance Simulation Association*.

[96] Wang, X. and Xu, D. (2010). An inverse gaussian process model for degradation data. *Technometrics*, 52(2):188–197.

[97] Wooldridge, M. (20009). *An Introduction to Multi-agent Systems*. John Wiley & Sons, second edition.

[98] Yan, J., Koç, M., and Lee, J. (2004). A prognostic algorithm for machine performance assessment and its application. *Production Planning & Control*, 15(8):796–801.

[99] Yang, F., Kwan, C. M., and Chang, C. S. (2008). Multiobjective evolutionary optimization of substation maintenance using decision-varying markov model. *IEEE Transactions on Power Systems*, 23(3):1328–1335.

[100] Yang, Z., Djurdjanovic, D., and Ni, J. (2007). Maintenance scheduling for a manufacturing system of machines with adjustable throughput. *IIE Transactions*, 39(12):1111–1125.

[101] Yao, X., Xie, X., Fu, M. C., and Marcus, S. I. (2005). Optimal joint preventive maintenance and production policies. *Naval Research Logistics*, 52(7):668–681.

[102] Ye, Z., Revie, M., and Walls, L. (2014). A load sharing system reliability model with managed component degradation. *IEEE Transactions on Reliability*, 63(3):721–730.

[103] Ye, Z.-S. and Chen, N. (2014). The inverse gaussian process as a degradation model. *Technometrics*, 56(3):302–311.

[104] Yin, R. K. (2002). *Case Study Research: Design and Methods (Applied Social Research Methods)*. Sage Publications Inc, third edition.

[105] Yu, F. and Chan, K. (2007). Optimum load sharing strategy for multiple-chiller systems serving air-conditioned buildings. *Building and Environment*, 42(4):1581 – 1593.

[106] Zhang, M., Gaudoin, O., and Xie, M. (2015). Degradation-based maintenance decision using stochastic filtering for systems under imperfect maintenance. *European Journal of Operational Research*, 245(2):531 – 541.

[107] Zhou, J., Djurdjanovic, D., Ivy, J., and Ni, J. (2007a). Integrated reconfiguration and age-based preventive maintenance decision making. *IIE Transactions*, 39(12):1085–1102.

[108] Zhou, J., Djurdjanovic, D., Ivy, J., and Ni, J. (2007b). Intergrated load-allocation and condition-based maintenance in a multi-unit load-sharing deteriorating system. In *Machine Failure Prevention Technologies Conference, Proceedings of*, pages 215–228.

[109] Zhou, R., Fox, B., Lee, H. P., and Nee, A. Y. C. (2004). Bus maintenance scheduling using multi-agent systems. *Engineering Applications of Artificial Intelligence*, 17(6):623–630.

[110] Zhou, X., Xi, L., and Lee, J. (2007c). Reliability-centered predictive maintenance scheduling for a continuously monitored system subject to degradation. *Reliability Engineering and System Safety*, 92(4):530–534.

[111] Zhu, Y., Elsayed, E., Liao, H., and Chan, L. (2010). Availability optimization of systems subject to competing risk. *European Journal of Operational Research*, 202(3):781 – 788.