

**MAINTENANCE STRATEGIES FOR
MANUFACTURING SYSTEMS USING
MARKOV MODELS**

by
Seung Chul Lee

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Mechanical Engineering)
in The University of Michigan
2010

Doctoral Committee:

Professor Jun Ni, Chair
Professor Xiuli Chao
Associate Professor Kazuhiro Saitou
Assistant Research Scientist Lin Li

© Seung Chul Lee

2010

To my parents

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Jun Ni, my graduate advisor and committee chair, for initiating my research at S. M. Wu Manufacturing Research Center. Without his continuous support and guidance, this work would not have been possible. I would also like to extend my sincere gratitude to Professor Xiuli Chao, Professor Kazuhiro Saitou, and Dr. Lin Li for serving on my doctoral committee.

I would also like to thank my friends and co-workers in the Wu Manufacturing Research Center for their help and support, especially Adam Brzezinski, Jaewook Oh, Dr. Kwanghyun Park, Dr. Bongsuk Kim, Christopher Jeon, Christopher Hajime Gerdes, Tian Jiang (TJ) Ye, Juil Yum, Dr. Masahiro Fujiki, Roberto Torres, Seng Keat Yeoh, Bruce Tai, Ahmad Almuhtady, Xiaoning Jin, Dr. Yang Liu, Dr. Jing Zhou, Professor Dragan Djurdjanovic, Professor Shihyoung Ryu, Professor Deokki Choi, Professor Aydin Goker, Dr. Omer Tsimhoni, Professor Dawn Tilbury, and Professor Jay Lee.

TABLE OF CONTENTS

DEDICATION	ii
ACKNOWLEDGEMENTS	iii
LIST OF FIGURES	vi
LIST OF TABLES	ix
LIST OF SYMBOLS	xi
CHAPTER	
I. Introduction	1
1.1 Motivation	1
1.2 Maintenance Strategies in Manufacturing	2
1.2.1 Reactive Maintenance (RM)	2
1.2.2 Preventive Maintenance (PM)	3
1.2.3 Condition-Based Maintenance (CBM)	3
1.3 Research Objectives	4
1.4 Outline of the Dissertation	5
II. Online Degradation Assessment and Adaptive Fault Detection Using Modified Hidden Markov Models	7
2.1 Introduction	7
2.2 The Proposed mHMM with Variable State Space	9
2.2.1 Hidden Markov Model and State Estimation	9
2.2.2 The Modified Hidden Markov Model	13
2.3 Case Studies	18
2.3.1 Inability of a Conventional HMM with Unknown States	18
2.3.2 Case Study on Tool Wear of Turning Process	25
2.4 Conclusion and Future work	33
III. Markov-based Preventive Maintenance Planning With Repair Time and Periodic Inspection	34

3.1	Introduction	34
3.2	Modeling of Maintenance Policies for a Single Machine	36
3.2.1	Modeling of Machine Degradation Processes with Markov Process	36
3.2.2	Approximation of Constant Time Delay in Markov Process	39
3.2.3	Maintenance Model for Single Unit System with Multiple Maintenance Tasks	40
3.2.4	Comparison with Conventional Markov Models	48
3.3	Modeling of Maintenance for a Two Unit System	52
3.3.1	Parallel Configuration	53
3.3.2	Serial Configuration	55
3.3.3	Optimal Inspection Interval	56
3.4	Case Study with Semiconductor Manufacturing Process Data	58
3.5	Conclusion	61
 IV. Decision Making for Simultaneous Maintenance Scheduling and Production Sequencing		63
4.1	Introduction	63
4.2	Model of A Manufacturing System	65
4.2.1	The Problem Statement	66
4.2.2	Model Assumptions	66
4.2.3	Degradation Model	67
4.3	Joint Job Sequencing and Maintenance Scheduling For a Multiple Product, Multiple Station System	68
4.3.1	Long-Term Planning	68
4.3.2	Short-Term Scheduling	72
4.4	Numerical Case Studies	76
4.4.1	Design	76
4.4.2	Sequencing and Maintenance Policies	78
4.4.3	Results	80
4.5	Case Study with Semiconductor Manufacturing Process Data	88
4.6	Conclusion	90
 V. Conclusions and Contributions		92
5.1	Conclusions	92
5.2	Contributions	93
5.3	Future Work	94
 BIBLIOGRAPHY		95

LIST OF FIGURES

Figure

2.1	Basic form of an HMM	10
2.2	Block diagram of the proposed modified HMM algorithm	16
2.3	Markov chain with an unknown state S_5	18
2.4	Original observable signals and the HMM with the four known states	19
2.5	HMM algorithm serving to estimate states	20
2.6	Posterior probability of $P\{q(n) = S_i O(1) \cdots O(n)\}$	21
2.7	Result of an incorrect state estimation	22
2.8	Posterior probability, but no jagged appearance is shown	23
2.9	Posterior probabilities using the mHMM	24
2.10	Test-bed of turning process with coolant supply (F_t : thrust force, F_c : cutting force)	25
2.11	Turning process with a new tool	27
2.12	Estimated states (vertical dashed lines indicate true states while solid lines represent estimated states)	29
2.13	Control chart with a conventional HMM	29
2.14	Control chart with a mHMM	30
2.15	The estimated states via various algorithms	32
3.1	A state transition diagram and its transition rate matrix	37
3.2	Simulation results for single unit degradation process	38

3.3	Simulation result from Erlang process with 200 intermediate states .	40
3.4	Illustration of the maintenance policy	41
3.5	Markov process for the abovementioned maintenance policy	42
3.6	State probability for the abovementioned maintenance policy	43
3.7	Sample path for the abovementioned maintenance policy	43
3.8	Availability as a function of inspection intervals	47
3.9	Maintenance policy comparison: not periodic and neglect repair time	48
3.10	Maintenance policy comparison: more realistic model with periodic inspection and non-negligible repair time	49
3.11	A traditional Markov model (left) and its simulation result (right) .	50
3.12	The proposed Markov model (a) and its simulation result (b)	51
3.13	Markov degradation model for a two-unit system without maintenance	52
3.14	Parallel (left) and serial (right) configurations	53
3.15	Maintenance policy in parallel configuration	53
3.16	Maintenance Markov model for a two unit parallel system	54
3.17	Maintenance policy in serial configuration	55
3.18	Maintenance Markov model for a two-unit serial system	55
3.19	Optimal intervals for PM with different connections	56
3.20	Optimal PM interval to maximize system productivity in parallel . .	57
3.21	5-state Markov chain with the corresponding transition probability matrix P	59
3.22	State probabilities and reliability distribution	59
3.23	Historical inspection intervals from the real Fab data	60

4.1	A multi-product, multi-station system in series	65
4.2	Illustration of a state transition diagram for a Makrov chain	68
4.3	Single machine, multiple product system	69
4.4	Improvement as a function of product demand ratio	81
4.5	Improvement as a function of degradation	83
4.6	Improvement as a function of rewards	84
4.7	Improvement as a function of maintenance costs	86
4.8	Impact of initial buffer contents	87
4.9	Impact of limited maintenance resource	88

LIST OF TABLES

Table

2.1	μ and Σ of Gaussian density distributions and P	18
2.2	The cutting conditions	26
2.3	Three states of HMM based on different tool flank wears	26
2.4	Correct estimation rate comparison	31
3.1	Event log table for the sample path in Figure 3.7	44
3.2	Benchmark results (r_p : Production per time unit, C_m : Maintenance cost per time unit	60
4.1	Product demand ratios	76
4.2	Transition probability matrices for equipment deterioration processes	77
4.3	Reward for products	78
4.4	Cleaning/repairing cost	78
4.5	Yield matrices	79
4.6	Benchmark policies	79
4.7	Simulation results with different product demand ratios	82
4.8	Transition probability matrix for degradation	82
4.9	Simulation results with different degradation processes	83
4.10	Simulation results with different rewards	85
4.11	Simulation results with different maintenance costs	85

4.12	5-state Markov chain with the corresponding transition probability matrix P	89
4.13	Benchmark results	89

LIST OF SYMBOLS

$\mathbf{S} = \{S_1, \dots, S_M\}$	the state space for a discrete degradation process
p_{ij}	the transition probability from state S_i to state S_j in a discrete time Markov chain (DTMC)
$P = \{p_{ij}\}$	the state transition probability matrix
$b_i(O(n))$	the observation symbol probability distribution, $= P\{O(n) q(n) = S_i\}$
$\pi = \{\pi_i\}$	the initial state probability distribution, $\pi_i = P\{q(1) = S_i\}$
$O_n = \{O(1), \dots, O(n)\}$	the sequence of all observation symbols up to time n
$q_n = \{q(1), \dots, q(n)\}$	the actual state sequence up to time n
$\hat{q}_n = \{\hat{q}(1), \dots, \hat{q}(n)\}$	the maximum likelihood state sequence up to time n
$\boldsymbol{\lambda} = (P, b, \pi)$	the HMM model parameters
$\alpha_n(i)$	$\alpha_n(i) = P\{O(1) \cdots O(n) \wedge q(n) = S_i \lambda\}$
$\xi_n(i, j)$	$\xi_n(i, j) = P(q(n) = S_i \wedge q(n+1) = S_j O_N \wedge \lambda)$
$\gamma_n(i)$	$\gamma_n(i) = P(q(n) = S_i O_N \wedge \lambda)$
Σ	the covariance matrix
$\bar{O}(n)$	the vector of the observation symbol mean of the n^{th} observation

$D^2(\bar{O}(n), \mu_i)$	the weighted distance of $\bar{O}(n)$ from the μ_i
UCL	Upper Control Limit
λ	the failure rate between states
m	the number of intermediate states
N	the number of different PM tasks
q_i	the probability of the i^{th} PM task being requested, and $\sum_{i=1}^N q_i = 1$
T	the time interval between consecutive inspections
T_{PM_i}	the time duration for the i^{th} PM task, $0 \leq i \leq N$
T_{PM}	Arithmetic average of T_{PM_i} , $T_{PM} = q_1 T_{PM_1} + \dots + q_N T_{PM_N}$
T_{RM}	the time duration for RM
μ	(= m/T) the transition rate for the periodic inspection
μ_{PM_i}	(= m/T_{PM_i}) the transition rate for the i^{th} PM task, $0 \leq i \leq N$
μ_{RM}	(= m/T_{RM}) the transition rate for the RM
P_{ij}	the steady state probability of state being in S_{ij}
$P_i(t)$	the state probability of state S_i at time t
Q	the transition rate matrix
r_p	the production per time unit
C_m	the maintenance cost per time unit
H	the number of stations in series
p_k	the product type k , where $k = 1, \dots, K$

ω_k	the long run proportion of product p_k , and $\sum_k \omega_k = 1$
$X(n)$	the degradation process of a machine at time n , represented by a discrete time Markov chain(DTMC)
$a(n)$	the action taken in time n
$Y(i, k)$	the yield of product k when the machine is in state S_i
$R(i, a)$	the immediate reward when action a is taken in state S_i
$P_a(i, j)$	$P\{X(n+1) = S_j X(n) = S_i, a(n) = a\}$
$x(i, a)$	the probability that a machine is in state S_i , and action a is taken
$\Pi(i, a)$	the decision rule that specifies action a when a machine is in state S_i
$V_\Pi(i)$	the long run expected average profit per unit time when a policy Π is employed and the initial state is S_i
t_p	the processing time
$T_F(i, n)$	the time when buffer i is full due to the failure of machine M_i
$T_E(i, n)$	the time when buffer i is empty due to the failure of machine M_i
$W(k, n)$	WIP of buffer k at time n
$W_i(k, n)$	WIP of buffer k at time n after machine M_i fails
$C(k)$	capacity of buffer k
$S(k, n)$	empty space (= slack) in buffer k at time n , and $S(k, n) =$

	$C(k) - W(k, n)$
$\mu(k)$	unit WIP inventory cost of buffer k
$\mathbf{Cost}_W(i, n)$	total WIP inventory costs at time n , ($= \sum_{k=1}^H W_i(k, n)\mu(k)$)
$\mathbf{Cost}_S(i, n)$	total slack WIP inventory costs at time n , ($= \sum_{k=1}^H S_i(k, n)\mu(k) = \sum_{k=1}^H [C(k) - W_i(k, n)]\mu(k)$)
C_{WIP}^*	the maximum WIP inventory costs
$d(i, n)$	the binary decision variables for repair on machine M_i in time n
$r(i, n)$	the repair tasks/requests on machine M_i in time n

CHAPTER I

Introduction

1.1 Motivation

Manufacturing systems have become highly automated and mechanized so that the impact of unplanned downtime caused by system failures becomes worse than ever. Unplanned downtime of equipment not only reduces line productivity but also negatively affects the quality control of the products. Another consequence of system failures is the escalation of maintenance expenses due to unpredictable maintenance. On the other hand, well planned maintenance cycles eventually decrease maintenance costs, increase productivity, reduce product variability, and ensure high quality goods and services.

Therefore, establishing a cost effective maintenance program is emerging as one of the key objectives in the production line. It has been recognized that the maintenance is not an isolated technical discipline but an integral part of the competitive plant operations [1]. To examine the trade-offs between maintenance costs and benefits, one needs an appropriate maintenance policy and relevant system performance measurements. These are typically brought together in a maintenance optimization model. This is a mathematical model in which both costs and benefits of maintenance are quantified and delivers an optimum balance between the two. However,

this mathematical model has not been well developed in a practical manner in spite of its importance. Insufficient representation of the degradation system under preventive maintenance has posed difficulties in the mathematical models, thus resulting in an inadequate maintenance policy. In addition, many of them have only taken into account a steady-state behavior of the system.

To overcome the aforementioned limitations this research will investigate the stochastic modeling techniques. Online degradation assessment and adaptive anomaly detection will be addressed for condition-based maintenance. Online machine health information can further be investigated for the relationship on the product quality and equipment deterioration. We investigate analytical and numerical examination of production lines within the Markov process framework, focusing on the more accurate dynamic behavior modeling and multiple maintenance tasks.

1.2 Maintenance Strategies in Manufacturing

Any systems used in manufacturing deteriorate with usage and age. System degradation causes more operating costs and decreases product quality. To keep production costs low while maintaining good quality, Reactive Maintenance (RM), Preventive Maintenance (PM), and Condition-Based Maintenance (CBM) are often performed on such deteriorating systems [2]. Here, maintenance can be defined as actions 1) to control the degradation process leading to failure of a system, and 2) to restore the system to its operational state.

1.2.1 Reactive Maintenance (RM)

Traditional maintenance responses is repair work to restore the system from its malfunctioning status. Such reactive actions would take place only when breakdowns are noticed from a system failure. Even today, reactive maintenance is still neces-

sary in all maintenance applications, because the complexity of a modern production system makes accurate prediction and monitoring difficult.

1.2.2 Preventive Maintenance (PM)

PM is defined as all actions performed in an attempt to retain an item in specified condition by providing systematic inspection, detection, and prevention of incipient failures [3]. Generally, PM involves lower downtime than RM due to availability of resources (spare parts, trained personnel, special tools, maintenance facilities), causing less logistic delay. Thus, the cost of PM is in general much less than that of the RM. Moreover, PM can prolong the useful life of the production equipment [4]. The concept of PM has been extensively studied. Since 1960s [5, 6, 7], researchers have recognized the importance of evaluating the effect of PM and scheduling it properly and efficiently. Summaries of these contributions can be found in the literature surveys of [2, 8, 9, 10, 11].

1.2.3 Condition-Based Maintenance (CBM)

With the rapid development of modern technology, products have become more complex while better quality and higher reliability are required. As the complexity of technology grows in a production system, the mechanisms of failure become more complicated, and a single pattern cannot manage to track all of them. Therefore, more efficient maintenance approaches such as CBM are implemented to handle the situation.

CBM recommends maintenance plans based on the information collected through numerous condition monitoring techniques [12, 13]. The basic principle of CBM is that defects which gradually develop in machines can be detected through suitable monitoring techniques at the early stages so that appropriate maintenance plans can

be scheduled accordingly. In other words, a CBM strategy can be used to dynamically determine system maintenance on the basis of the observed condition of the system.

A CBM program consists of three key steps [12]:

- 1) Data acquisition to obtain data relevant to system health [14, 15]
- 2) Data processing to handle and analyze data or signals [16, 17, 18, 19, 20]
- 3) Maintenance decision-making model to recommend efficient maintenance policies [21, 22, 23, 24, 25, 26, 27, 28]

1.3 Research Objectives

The purpose of this research is to develop methods of pursuing enhanced cost-effective maintenance policy for complex manufacturing systems by considering the effects of the degraded equipment condition. The fundamental challenges and objectives in this research can be summarized as follows:

- Most condition-based diagnosis methods mainly focus on online degradation assessment of a system, assuming that all possible system conditions are known a priori and that training data sets from associated conditions are available. These assumptions significantly impede machine diagnosis applications where it is difficult to identify and train all of the possible states of the system in advance. Note that training models from data is a necessary step to estimate system parameters for maintenance decision-making models. This problem may cause serious estimation errors in the occurrence of unknown or untrained faults that might generate catastrophic damages to systems. Therefore, this research introduces an anomaly detection algorithm to trigger a model to update its structure, and provides a more accurate model for the system.

- The second fundamental difficulty lies on how to create an appropriate model of both degradation processes and maintenance for multiple machine systems. Although such a model can be used to find the optimal maintenance policy, mathematical models for this purpose has not been well studied in a practical manner. Insufficient representation of the degradation process with maintenance actions leads to an inadequate maintenance decision-making policy. This research, therefore, presents an approach where stochastic models are used to represent equipment degradation and incorporated in various maintenance decision processes. To approximate non-negligible maintenance times and periodic inspections for more realistic maintenance activities, a fundamental understanding of the phase-type distribution will be investigated.
- Extending the models and generalizing the results from a single machine system to multiple product systems still remain as challenging tasks to the proposed research because the relationship between machine condition and product quality has not been successfully collaborated with maintenance decision-making. Therefore, a complex manufacturing system with multiple products has to be studied, focusing on the relationship between deteriorated equipment and associated product qualities. A new decision-making architecture will be developed to determine joint maintenance and product sequencing rules based on condition monitoring information.

1.4 Outline of the Dissertation

This research is organized as follows:

Chapter II presents how hidden Markov models can be applied to assess online degradation process and identify anomaly detection. A turning process will be demon-

strated to validate our proposed algorithm in this chapter. Chapter III studies the optimal preventive maintenance policy. Starting with a single component system, the problem of a two-unit system will be solved in this chapter. Chapter IV is devoted to a multiple product system with its application to a semiconductor manufacturing process. In this chapter, a policy for maintenance planning and product dispatching will be proposed in the presence of multiple products.

CHAPTER II

Online Degradation Assessment and Adaptive Fault Detection Using Modified Hidden Markov Models

2.1 Introduction

Condition-based maintenance (CBM) recommends maintenance plans based on the information collected through numerous condition monitoring techniques [12, 13]. The basic principle behind CBM is that defects that gradually develop in machines can be detected through suitable monitoring techniques at the early stages so that appropriate maintenance plans can be scheduled accordingly. Because of the complexity of modern plants, CBM has become widely accepted as one of the key drivers to reduce maintenance costs and machine downtime of manufacturing systems [29].

Condition monitoring techniques for machine diagnosis have been studied extensively [12]. Many signal processing techniques have been developed that involve the analysis of the acquired data in time domains [30], frequency domains [31], and time-frequency domains [32]. Paya et al. [33] developed a condition monitoring method, which relied on wavelet transformation and artificial neural networks. A similar work that uses principal component analysis was reported in Jin et al. [34]. These methods are feature-based methods using statistical features of the signal. However, these

methods require too much data and time to establish condition monitoring and diagnosis. In addition, they are data-based methods, which do not take the physical model of the system into consideration. On the other hand, model-based methods, under the assumption that measured information is stochastically correlated with the actual machine condition, take advantage of understanding the system structure [12].

This assumption leads to the application of a Hidden Markov Model (HMM) through a statistical approach in identifying the actual machine conditions from observable monitoring signals. Although HMMs were motivated by their successes in speech recognition [35], many applications of the HMM in machine process diagnosis have also been studied, demonstrating its effectiveness in online diagnosis. For example, Ertunc et al. [36] presented an HMM approach for tool wear detection and prediction in a drilling process. A similar approach was also described for a turning process by Wang et al. [37]. Li et al. used an HMM as a fault diagnosis tool in speed-up and speed-down processes for rotating machinery [38].

According to the literature review, most previous condition-based diagnosis models based on an HMM mainly focus on online degradation assessment of known faults [35, 36, 37, 38]. An HMM with known faults assumes that all possible system condition states are known a priori and that training data sets from associated states are available. In addition, training an HMM should be conducted offline. These assumptions significantly impede machine diagnosis applications when it is difficult to identify and train all of the possible states of the system in advance [39, 40]. For instance, if an HMM that has been trained to model gradual tool wear in a drilling process does not have a state to represent a tool breakage or shortage of coolant, it is impossible for the HMM to estimate the correct state when these untrained states occur. The state structure of a conventional HMM will not be updated after the train-

ing stage. This inflexibility may cause serious estimation errors in the emergence of unknown or untrained faults that might provoke catastrophic damages to machining processes.

Therefore, it is necessary to introduce an online anomaly detection algorithm into an HMM to trigger the HMM to adjust the number of hidden states or the hidden structure, and thus result in a more accurate model for the system. In this chapter, a modified Hidden Markov Model (mHMM) with variable state space is developed to estimate the current state of system degradation as well as to detect the emergence of unknown faults at an early stage. The Statistical Process Control (SPC) technique [41] is used in unknown fault detection and diagnostics in conjunction with the HMM. By measuring the deviation of the current signal from a reference signal representing prior known states, the mHMM is able to see whether the current signal is within the control limits.

The rest of this chapter is organized as follows: Section 2.2 introduces the principle of an HMM and the proposed mHMM for online degradation assessment and state update. In Section 2.3, case studies are performed to validate the effectiveness of the mHMM algorithm and to compare its performance with other methods using the example of a turning process. The conclusions and future research directions are given in Section 2.4.

2.2 The Proposed mHMM with Variable State Space

2.2.1 Hidden Markov Model and State Estimation

Before introducing the mHMM, we present the basic form of a traditional HMM with fixed state space as shown in Figure 2.1. The HMM, $\lambda = (P, b, \pi)$ under consideration consists of:

- a finite set of M states, $\mathbf{S} = \{S_1, \dots, S_M\}$
- a state transition probability matrix, $P = \{p_{ij}\}_{M \times M}$,
 where $p_{ij} = P\{q(n+1) = S_j | q(n) = S_i\}$, $1 \leq i, j \leq M$, $1 \leq n < N$
- an observation symbol probability distribution,
 $b_i(O(n)) = P\{O(n) | q(n) = S_i\}$, $1 \leq i \leq M$, $1 \leq n \leq N$
- an initial state probability distribution, $\pi = \{\pi_i\}_M$
 where $\pi_i = P\{q(1) = S_i\}$, $1 \leq i \leq M$

An HMM technique is applicable to a process that is assumed to possess homogeneous Markovian property [42] as follows:

$$p_{ij} = P\{q(n) = S_j | q(n-1) = S_i, \dots, q(1)\} = P\{q(n) = S_j | q(n-1) = S_i\} \quad (2.1)$$

Equation (2.1) implies that the conditional probability of the current state, given knowledge of all previous states, is the same as the conditional probability of the current state given knowledge of the system state of one previous time unit. In other words, the probability that a system will undergo a transition from one state to another state depends only on the current state of system.

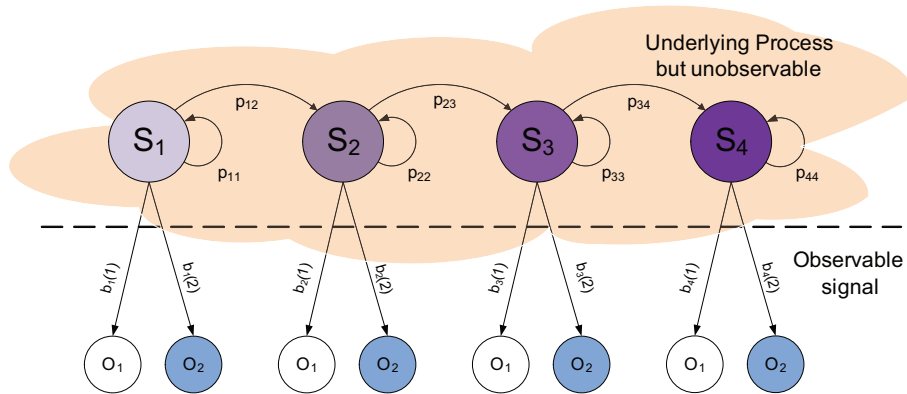


Figure 2.1: Basic form of an HMM

The state transition probability matrix P encodes the uncertainty in the true underlying state evolution of the stochastic process while each state emits observation symbols with the probability distribution $b_i(O(n))$, as shown in Figure 2.1. Let $O_n = \{O(1), \dots, O(n)\}$ denote a sequence of all observation symbols up to time n , where observed data points $O(n)$ are taken at time n . The actual state sequence up to time n can be represented as $q_n = \{q(1), \dots, q(n)\}$, where a state $q(k) \in \mathbf{S}$, $1 \leq k \leq n$. Then, we can find the maximum likelihood state sequence, $\hat{q}_t = \{\hat{q}(1), \dots, \hat{q}(t)\}$ associated with a given sequence of observations, $O_n = \{O(1), \dots, O(n)\}$ as well as an HMM model, $\boldsymbol{\lambda} = (P, b, \pi)$ through the Viterbi algorithm [43, 44]. Furthermore, it is possible to adjust the HMM model parameters, $\boldsymbol{\lambda} = (P, b, \pi)$ to maximize the probability of the observation sequence using an iterative procedure such as the Baum-Welch method [45] or the Expectation-Maximization (EM) algorithm [46].

Since the primary purpose of an HMM in this chapter is to estimate the system state as early as possible, the forward procedure [35] based on past and present measurements is employed. Consider the forward variable, $\alpha_n(i)$ defined as $\alpha_n(i) = P\{O(1) \cdots O(n) \wedge q(n) = S_i | \boldsymbol{\lambda}\}$, indicating the joint probability of a series of observed symbols $O_n = \{O(1), \dots, O(n)\}$ and state S_i at time n , given the model $\boldsymbol{\lambda}$. We can then calculate $\alpha_n(i)$ recursively, as follows:

1) Initialization

$$\alpha_1(i) = \pi_i b_i(O(1)), \quad 1 \leq i \leq M \quad (2.2)$$

2) Induction

$$\alpha_{n+1}(j) = \left[\sum_i p_{ij} \alpha_n(i) \right] b_j(O(n+1)), \quad 1 \leq n < N, \quad 1 \leq j \leq M \quad (2.3)$$

Once $\alpha_n(i)$'s are obtained, the posterior probabilities, $P\{q(n) = S_i | O(1) \cdots O(n) \wedge \boldsymbol{\lambda}\}$ that the current state $q(n)$ is in state S_i , given the observed symbols, $O_n = \{O(1), \dots, O(n)\}$ can be calculated by the Bayes' rule.

$$P\{q(n) = S_i | O(1) \cdots O(n) \wedge \boldsymbol{\lambda}\} = \frac{P\{q(n) = S_i \wedge O(1) \cdots O(n) | \boldsymbol{\lambda}\}}{P\{O(1) \cdots O(n) | \boldsymbol{\lambda}\}} = \frac{\alpha_n(i)}{\sum_j \alpha_n(j)}, \quad 1 \leq i \leq M \quad (2.4)$$

Hence, we can estimate the state $\hat{q}(n)$, which maximizes the posterior probability as:

$$\hat{q}(n) = \arg \max_i \{P\{q(n) = S_i | O(1) \cdots O(n) \wedge \boldsymbol{\lambda}\}\} \quad (2.5)$$

Furthermore, the EM algorithm is used to find the maximum likelihood HMM parameters, $\boldsymbol{\lambda} = (P, b, \pi)$ that could have produced the sequence of observations $O_N = \{O(1), \dots, O(N)\}$. Define $\xi_n(i, j)$ and $\gamma_n(j)$ as follows:

$$\xi_n(i, j) = P\{q(n) = S_i \wedge q(n+1) = S_j | O_N \wedge \boldsymbol{\lambda}\} \quad (2.6)$$

$$\gamma_n(i) = P\{q(n) = S_i | O_N \wedge \boldsymbol{\lambda}\} \quad (2.7)$$

$\xi_n(i, j)$ in Equation (2.6) is the probability of being in state S_i at time n and in state S_j at time $n+1$, given the model $\boldsymbol{\lambda}$ and the observation sequence O_N up to time N . Note that $\gamma_n(i)$ in Equation (2.7) is the probability of being in S_i at time n , given the model and the observation sequence O_N . Thus, a set of re-estimation for $\hat{\boldsymbol{\lambda}} = (\hat{P}, \hat{b}, \hat{\pi})$ would be expressed as:

$$\hat{\pi}_i = \gamma_1(i), \quad 1 \leq i \leq M \quad (2.8)$$

$$\hat{\alpha}_{ij} = \frac{\sum_{n=1}^{N-1} \xi_n(i, j)}{\sum_{n=1}^{N-1} \gamma_n(i)}, \quad 1 \leq i, j \leq M \quad (2.9)$$

$$\hat{b}_i(O(n)) \sim N(\mu_i, \Sigma_i), \quad 1 \leq i \leq M, 1 \leq n \leq N \quad (2.10)$$

We will use Equations (2.8), (2.9), and (2.10) to update an HMM [35]. It should be noted that the number of discrete states and the selection of training data sets have a great influence on HMM's performance for state estimation. Therefore, states have to be selected in such a way that maximizes the discrepancies among the states. In addition, the size of training data set has to be large enough to ensure observation symbol probability distributions to be statistically significant.

2.2.2 The Modified Hidden Markov Model

We propose to use the mHMM with variable state space to detect the emergence of anomalies at the early stages as well as to estimate the current state of system degradation. The technique of SPC is combined with an HMM to detect different failure modes and diagnostics. The mHMM can check whether the current signals are emitted from unknown failure modes that have not been observed by calculating the deviation of the current signal from reference signals representing prior known states. Therefore, the mHMM is equivalent to an HMM equipped with the reinforcement learning technique.

Suppose that there are m observations available from the process, each of size b , and the observation symbol probability distributions, $b_i(O(n)) = P\{O(n)|q(n) = S_i\}$ follow a p -jointly Gaussian density distribution. This assumption is reasonable in many applications because of the central limit theorem, which states that the sum of independently distributed random variables is approximately Gaussian-distributed regardless of the distributions of the individual variables as the number of samples becomes large [42]. Then $b_i(O(n))$ can be expressed as:

$$b_i(O(n)) = P\{O(n)|q(n) = S_i\} = \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(O(n)-\mu_i)^T \Sigma_i^{-1} (O(n)-\mu_i)\right) \quad (2.11)$$

where μ is the mean vector and Σ is covariance matrix of the distribution.

Then, the weighted distance of $\bar{O}(n)$ from the μ_i , known as Mahalanobis distance [47], can be calculated as:

$$D^2(\bar{O}(n), \mu_i) = b(\bar{O}(n) - \mu_i)^T \Sigma_i^{-1} (\bar{O}(n) - \mu_i) \quad (2.12)$$

where $\bar{O}(n)$ is the vector of the observation symbol mean of the n^{th} observation, $\bar{O}(n) = \frac{1}{b} \sum_{k=1}^b O_k(n)$.

We use this statistic to detect an unknown state in the mHMM because $D^2(\bar{O}(n), \mu_i)$ can represent a dissimilarity distance when the number of monitoring signals is more than one [48]. The most familiar multivariate process monitoring technique is the Hotelling multivariate control chart [49]. We use the Hotelling multivariate control chart technique for an anomaly detection algorithm in the mHMM because this method can deal with multiple monitoring signals, make an online decision based on current monitoring signals, and has shown effectiveness especially in a manufacturing process [48]. The Hotelling multivariate control chart signals that a statistically significant shift in the mean has occurred when

$$D^2(\bar{O}(n), \mu_i) > \text{UCL} \quad (2.13)$$

where $\text{UCL} > 0$ is a specified Upper Control Limit (UCL).

The calculation of UCL depends on whether the values of μ and Σ are known or not in advance. If μ and Σ are known, the D^2 statistic follows χ^2 -distribution with p degrees of freedom [50]. Thus, UCL can be obtained as

$$\text{UCL} = \chi_{\alpha, p}^2 \quad (2.14)$$

where α is the risk level.

If μ and Σ are not known, the m observation subgroups of each size b must be used to estimate μ with $\bar{\bar{O}}$, the overall mean vector, and Σ with $\bar{\bar{S}}$, the covariance matrix. $\bar{\bar{O}}$ and $\bar{\bar{S}}$ can be calculated:

$$\bar{\bar{O}} = \frac{1}{m} \sum_{n=1}^m \bar{O}(n) \quad (2.15)$$

$$\bar{\bar{S}} = \frac{1}{m} \sum_{n=1}^m (\bar{O}(n) - \bar{\bar{O}})^T (\bar{O}(n) - \bar{\bar{O}}) \quad (2.16)$$

It has been shown that $\bar{\bar{O}}$ and $\bar{\bar{S}}$ are the maximum likelihood estimates of μ and Σ , respectively [51]. In this case, the D^2 statistics and the UCL for the Hotelling multivariate control chart are defined as follows:

$$D^2(\bar{O}(n), \bar{\bar{O}}) = b(\bar{O}(n) - \bar{\bar{O}})^T \bar{\bar{S}}^{-1} (\bar{O}(n) - \bar{\bar{O}}) \quad (2.17)$$

$$\text{UCL} = \frac{p(m-1)(b-1)}{mb-m-p+1} F_{\alpha, p, mb-m-p+1} \quad (2.18)$$

Equation (2.18) is based on the fact that the $D^2(\bar{O}(n), \bar{\bar{O}})$ statistic follows an F -distribution with p and $(mb - m - p + 1)$ degrees of freedom when its mean and covariance are not known [52].

Therefore, we can claim that the process of interest is experiencing a statistically significant shift in the mean if Mahalanobis distance D^2 becomes larger than UCL. The mHMM makes use of this characteristic of SPC for the purpose of detecting unknown states. The mHMM with variable state space will adjust the number of hidden states or the hidden structure based on the result of the Hotelling multivariate control chart. A summary of the mHMM algorithm is shown in Figure 2.2.

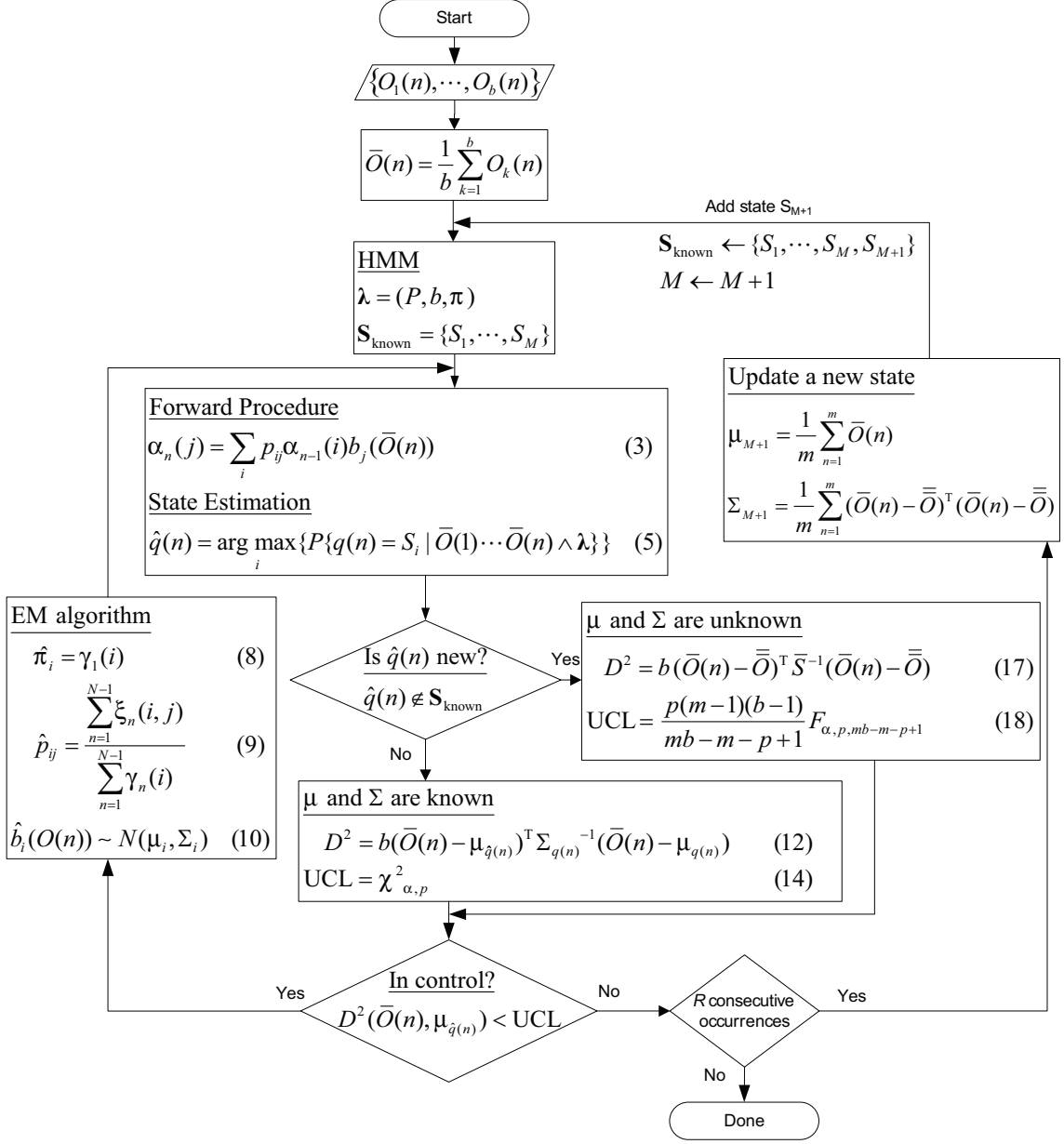


Figure 2.2: Block diagram of the proposed modified HMM algorithm

Suppose the initial mHMM is trained only with prior known states, $\mathbf{S}_{\text{known}} = \{S_1, \dots, S_M\}$ and associated training data sets. This mHMM receives a set of data $\{O_1(n), \dots, O_b(n)\}$ with a batch size b at time n . The sample mean of each set, $\bar{O}(n)$ is calculated (i.e., $\bar{O}(n) = \frac{1}{b} \sum_{k=1}^b O_k(n)$) and fed to the HMM state estimation algorithm, shown in Equations (2.3) and (2.5), to estimate the current state $\hat{q}(n)$ from the sequence of observation symbols $\bar{O}_n = \{\bar{O}(1), \dots, \bar{O}(n)\}$.

If $\hat{q}(n)$ belongs to the prior known state set $\mathbf{S}_{\text{known}}$, then the distance $D^2(\bar{O}(n), \mu_{\hat{q}(n)})$ and UCL are obtained by means of Equations (2.12) and (2.14), respectively. This is possible because the corresponding μ and Σ of $\hat{q}(n)$ are known. If $\hat{q}(n)$ does not belong to the prior known state space $\mathbf{S}_{\text{known}}$, Equations (2.17) and (2.18) can be used instead. If any anomalous behavior has not been detected via the control chart (i.e., $D^2 < \text{UCL}$), the sequence of observation symbols $\bar{O}_n = \{\bar{O}(1), \dots, \bar{O}(n)\}$ will be used to update the mHMM through the EM algorithm (i.e., re-learning or reinforcement learning). On the other hand, if $D^2 > \text{UCL}$ occurs R consecutive times, a new state S_{M+1} needs to be introduced to the mHMM to model an unknown state of the system with μ_{M+1} and Σ_{M+1} , as shown in Equations (2.15) and (2.16). The number R can be used to control the sensitivity of the unknown detection algorithm. However, there is a tradeoff between robustness and a detection speed. For instance, if R is increased, the unknown detection algorithm may become more robust against false detections caused by process randomness itself while a detection speed might be slow (i.e., the mHMM may respond more slowly to the unknown state). Thus, the number R needs to be tuned with considerable caution according to its purpose [53].

2.3 Case Studies

2.3.1 Inability of a Conventional HMM with Unknown States

In this section, we illustrate the effectiveness and outperformance of the mHMM with comparison to a conventional HMM using numerically generated case studies. To study the numerical cases where some of the hidden states of an HMM are not known, we consider the HMM which is trained initially with the four states, $\mathbf{S}_{\text{known}} = \{S_1, S_2, S_3, S_4\}$, while the true system actually contains another unknown state S_5 , as shown in Figure 2.3.

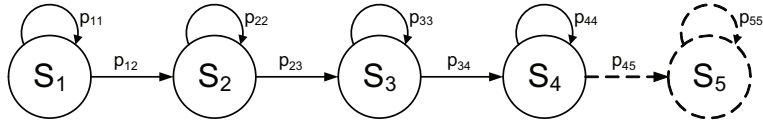


Figure 2.3: Markov chain with an unknown state S_5

Table 2.1: μ and Σ of Gaussian density distributions and P

$\mu_1 = \begin{bmatrix} 20 \\ 20 \end{bmatrix}$	$\mu_2 = \begin{bmatrix} 20 \\ 35 \end{bmatrix}$	$\mu_3 = \begin{bmatrix} 35 \\ 35 \end{bmatrix}$	$\mu_4 = \begin{bmatrix} 35 \\ 20 \end{bmatrix}$
$\Sigma_1 = \begin{bmatrix} 20 & 0 \\ 0 & 20 \end{bmatrix}$	$\Sigma_2 = \begin{bmatrix} 15 & 0 \\ 0 & 15 \end{bmatrix}$	$\Sigma_3 = \begin{bmatrix} 15 & -2 \\ -2 & 15 \end{bmatrix}$	$\Sigma_4 = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$
$P = \begin{bmatrix} 0.99 & 0.11 & 0 & 0 \\ 0 & 0.99 & 0.01 & 0 \\ 0 & 0 & 0.99 & 0.01 \\ 0.01 & 0 & 0 & 0.99 \end{bmatrix}$			

Suppose that two signals (X_1, X_2) are monitored. The observation symbol probabilities from each state have two-jointly Gaussian density distributions and the HMM has the transition probability matrix P , as summarized in Table 2.1.

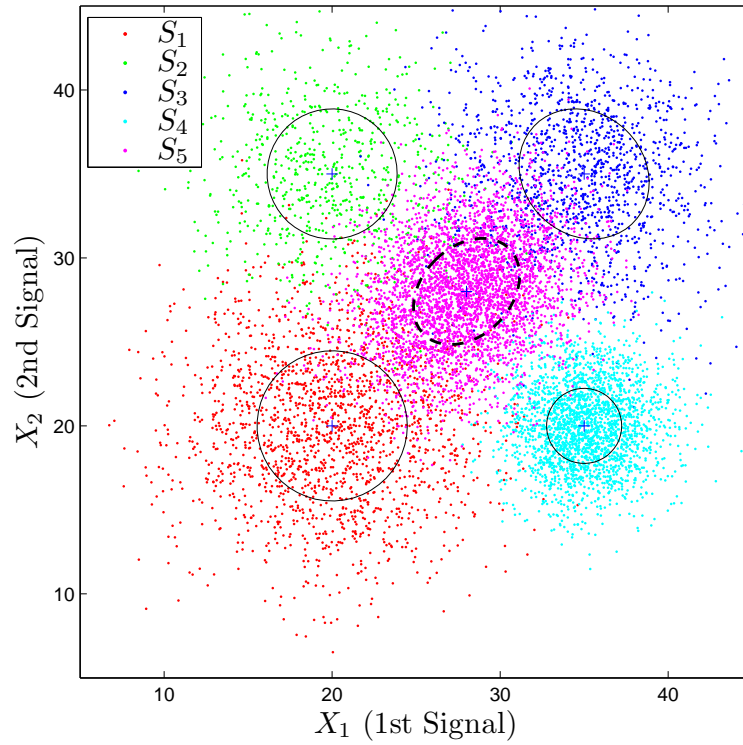


Figure 2.4: Original observable signals and the HMM with the four known states

One possible result of the observable signals is illustrated in Figure 2.4 if samples of size $b = 10$ (i.e., one subgroup consists of 10 samples) are taken. Note that these signals are abstract and not linked to any specific physical meaning.

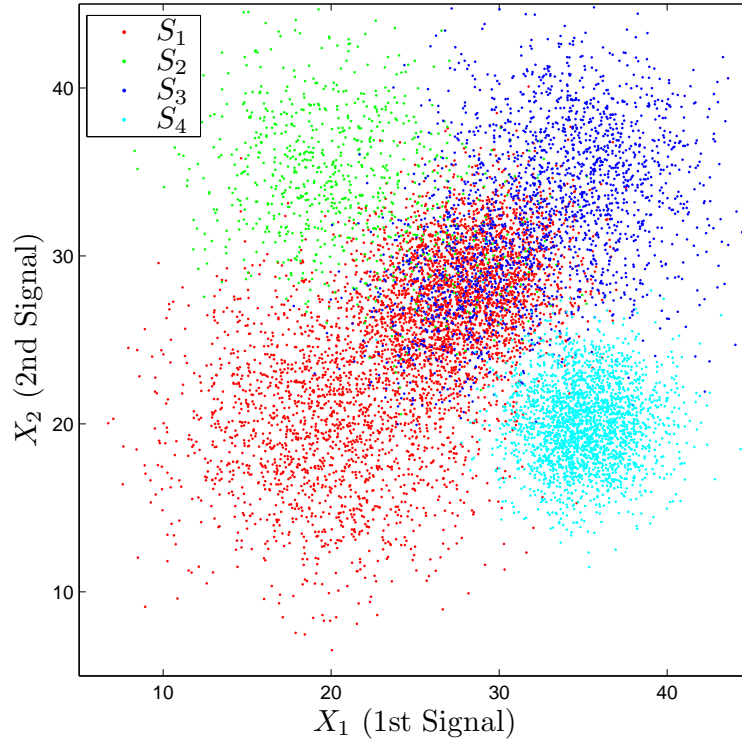


Figure 2.5: HMM algorithm serving to estimate states

However, the estimated states obtained from the sequence of observable signals by means of the conventional HMM algorithm are different from the true states of the system as shown in Figure 2.5. This is because the conventional HMM has to assign each observation to one of the known states, $\mathbf{S}_{\text{known}} = \{S_1, S_2, S_3, S_4\}$ according to the posterior probability calculation via Equation (2.4) even when an observation signal is emitted from the unknown state S_5 , (where $\mu_5 = [\frac{28}{28}]$, $\Sigma_5 = [\frac{10}{3} \frac{3}{10}]$).

The posterior probabilities of being in each state given the sequence of observation symbols up to the current time are obtained and illustrated in Figure 2.6.

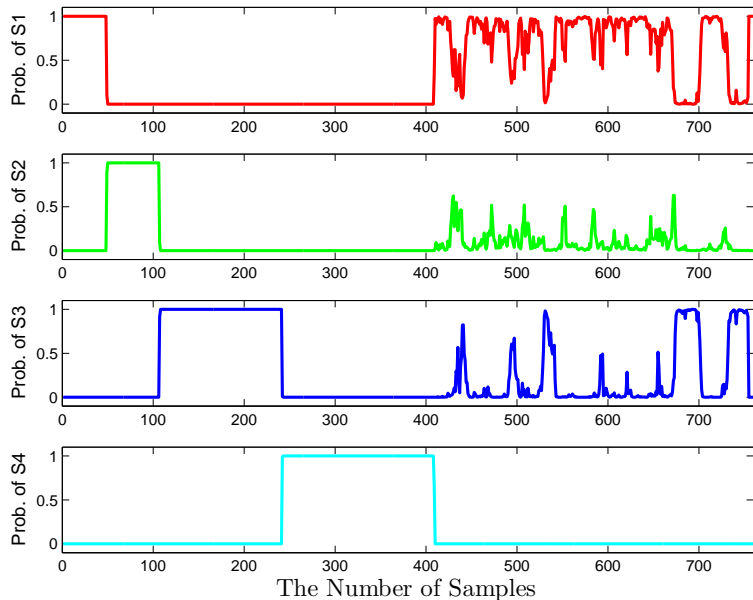


Figure 2.6: Posterior probability of $P\{q(n) = S_i | O(1) \cdots O(n)\}$

The jagged appearance in the posterior probabilities happens after approximately the 400th sample, since the conventional HMM does not account for the emergence of the unknown state. In this case, the conventional HMM is unable to estimate the correct states. On the other hand, the jagged appearance in the posterior probabilities represents the presence of an unknown state from the observation symbols. From the result shown in Figure 2.6, we might conclude that the posterior probability is the key criterion in determining the detection of unknown states, as explained in [39]. However, the following case study shows that this conclusion may not always hold.

The trained HMM, $\lambda = (P, b, \pi)$ is the same as the previous example. However, in this case it turns out that an unknown state S_5 has the following Gaussian observation symbol density distribution:

$$\mu_5 = \begin{bmatrix} 50 \\ 10 \end{bmatrix}, \quad \Sigma_5 = \begin{bmatrix} 10 & 3 \\ 3 & 10 \end{bmatrix}$$

Instead of being in the middle of the other states, the unknown state is far away from other four known states, as shown in Figure 2.7.

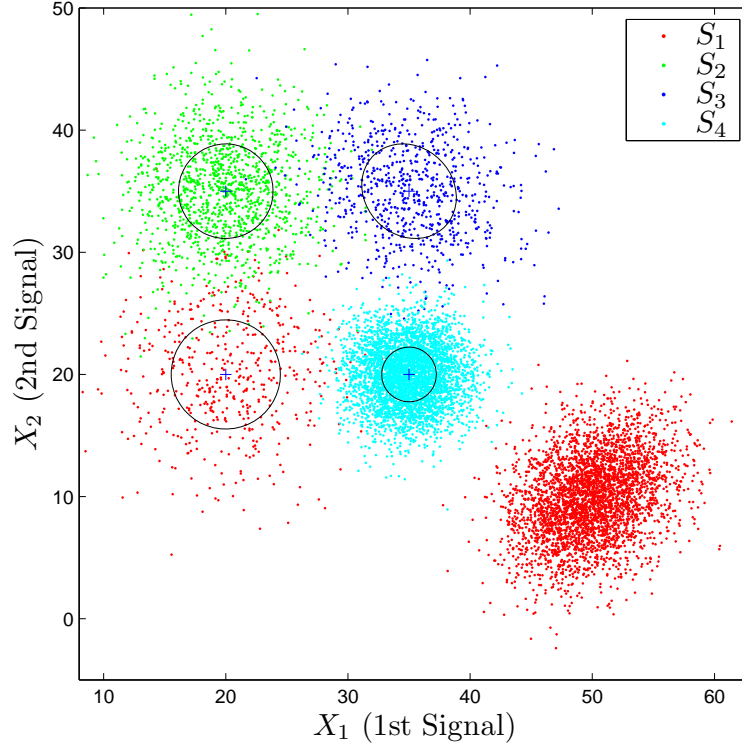


Figure 2.7: Result of an incorrect state estimation

As shown in Figure 2.8, we cannot see the jagged appearance in the posterior probabilities even with the presence of the unknown state in Figure 2.7. In this case, the conventional HMM algorithm disguises the unknown state by calculating $P\{q(n) = S_1 | O(1) \cdots O(n)\} = 0.9972$ after around the 650th sample. The conventional HMM misinterprets an unknown state S_5 as the first state S_1 with a high probability even though the unknown state is located far from the first state S_1 . The fourth row of the transition probability matrix, $P(4, :) = [0.01 \ 0.00 \ 0.00 \ 0.99]$ is defined in such a way that a state will move to either state S_1 or state S_4 after being in state S_4 . The conventional HMM, however, excludes the chance of being in state S_4 after observing that an observation symbol is far away from state S_4 . Thus, the conventional HMM misjudges that $P\{q(n) = S_1 | O(1) \cdots O(n)\} = 0.9972$.

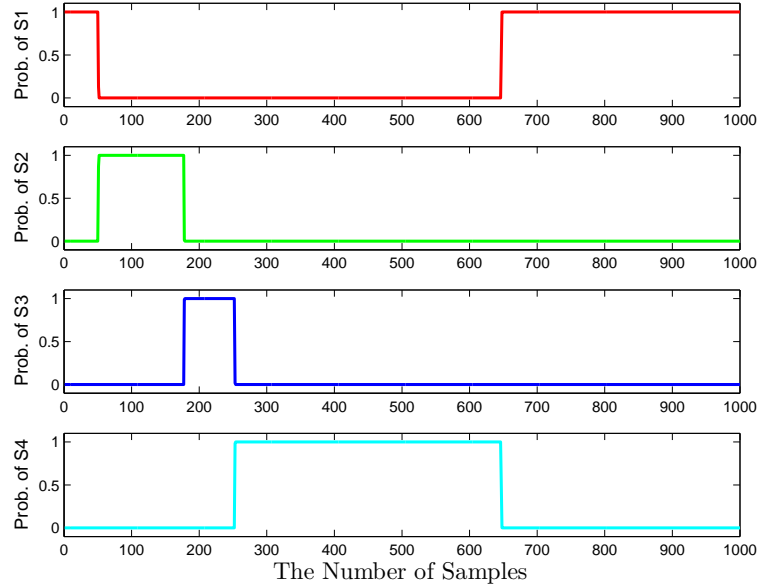
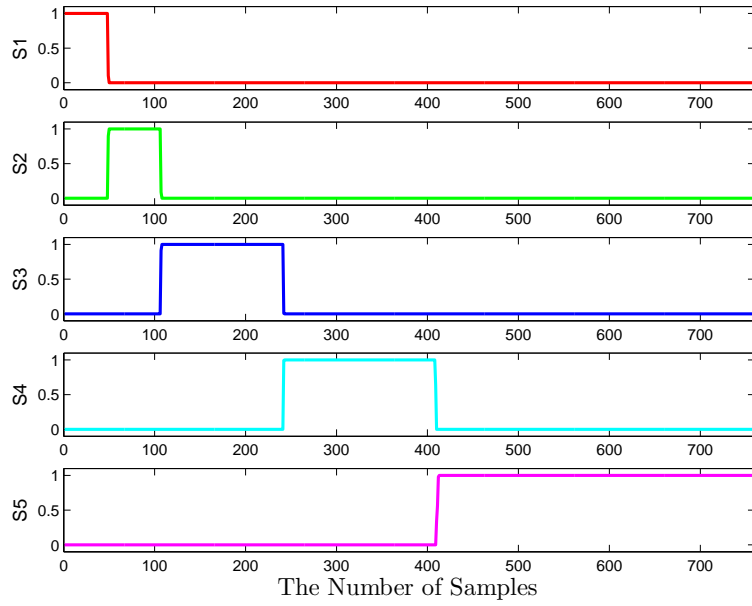
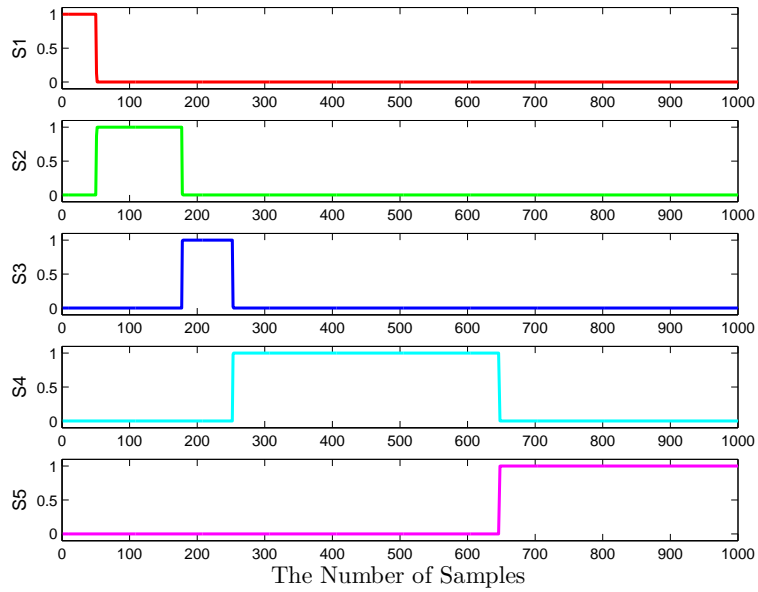


Figure 2.8: Posterior probability, but no jagged appearance is shown

These two examples lead us to conclude that considering only the posterior probability in the identification of the unknown state is not sufficient based on the conventional HMM. This is why we propose the modified HMM (mHMM) algorithm to deal with challenges related to unknown states using the Hotelling multivariate control chart. The simulation results using the mHMM are illustrated in Figures 2.9(a) and 2.9(b). Both cases show that unknown states are successfully detected and new states are added into a conventional Markov chain.



(a) Posterior probability with the first example



(b) Posterior probability with the second example

Figure 2.9: Posterior probabilities using the mHMM

2.3.2 Case Study on Tool Wear of Turning Process

We illustrate how the mHMM operates with the example of a tool degradation process. The proposed mHMM has been tested with a turning process and is shown to be able to perform an adaptive diagnosis of different failure modes as well as on-line degradation assessment. A ceramic tool is used to turn an Inconel718 workpiece with coolant supplied, as shown in Figure 2.10. During the turning process, two orthogonal forces (the cutting force and thrust force) are measured by the dynamometer.

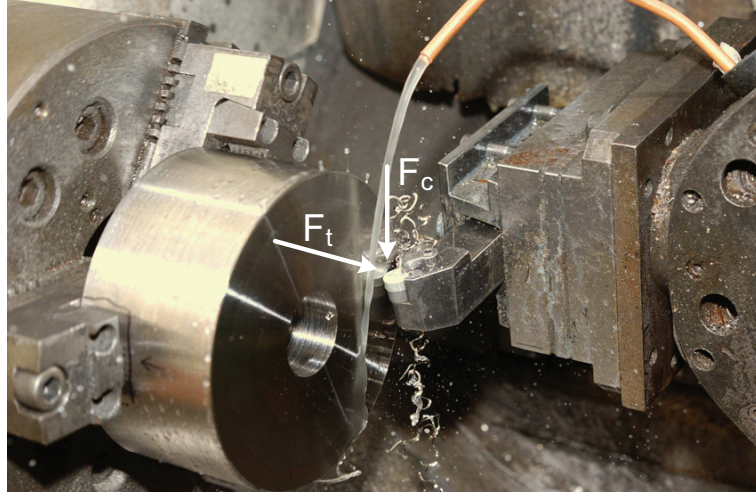


Figure 2.10: Test-bed of turning process with coolant supply (F_t : thrust force, F_c : cutting force)

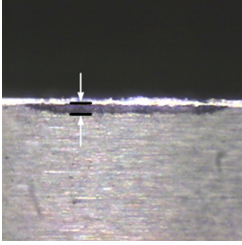
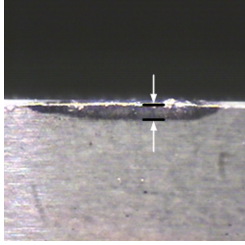
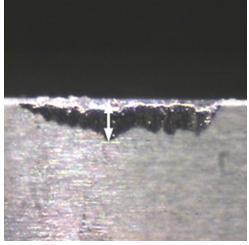
The first step is to train the mHMM using training data sets associated with each state. The states are defined as degree of tool flank wear. Three different degrees of tool flank wears, $\mathbf{S} = \{S_1, S_2, S_3\}$, are used to train the mHMM. The cutting and thrust forces are measured under the same turning process conditions such as the depth of cut, the feed rate, and the cutting speed (see Table 2.2). Note that enough coolant was supplied during this training stage.

Table 2.2: The cutting conditions

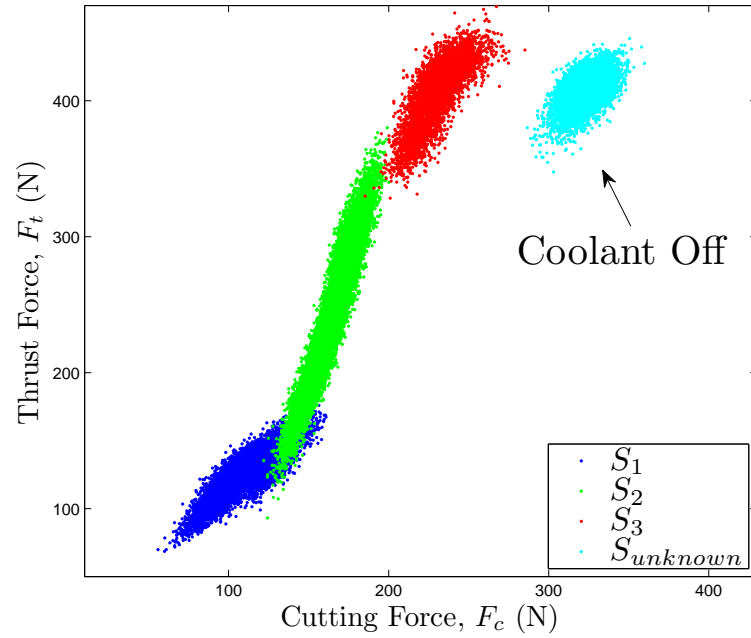
	Depth of Cut	Feed Rate	Cutting Speed
Condition	228.6 (μm)	228.6 (μm per rev)	152.4 (m per min)

Observation symbol probability distributions for each state are then calculated from two force signals in the form of the joint Gaussian density functions. The resultant mean and covariance matrix with corresponding tool wears are displayed in Table 2.3.

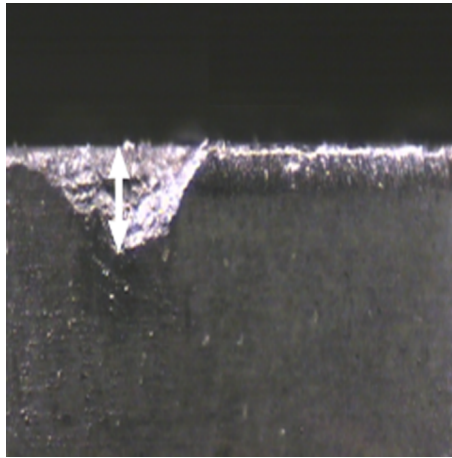
Table 2.3: Three states of HMM based on different tool flank wears

State	S_1	S_2	S_3
Pictures			
Tool flank wear(μm)	79.05 ± 0.005	103.70 ± 0.005	151.80 ± 0.005
Mean of two forces(N)	[108.5 124.6]	[166.7 251.1]	[230.4 404.8]
Covariance matrix	$\begin{bmatrix} 226.0 & 199.1 \\ 199.1 & 242.0 \end{bmatrix}$	$\begin{bmatrix} 151.1 & 547.7 \\ 547.7 & 2198.7 \end{bmatrix}$	$\begin{bmatrix} 159.8 & 234.2 \\ 234.2 & 538.4 \end{bmatrix}$

We then restart the turning process with a new tool while measuring the cutting and thrust forces. As shown in Figure 2.11, both cutting force and thrust force increase with process duration as a cutting tool loses its sharpness. After a tool wear status reaches state S_3 , the coolant supply is removed to introduce a different tool wear mode. The cutting force seems to increase when the coolant is not supplied. This dry machining condition generates non-experienced forces from an unknown state, S_{unknown} , that has not been seen during the training stage.



(a) Measured cutting and thrust forces



(b) Tool breakage due to dry machining

Figure 2.11: Turning process with a new tool

Figures 2.12 and 2.13 demonstrate the problem or drawback of the conventional HMM, showing that a conventional HMM fails to estimate S_{unknown} with high Mahalanobis distances. The distance statistic D^2 becomes larger than UCL after around the 2800th sample, which corresponds to the moment when the coolant is shut off. The estimation failure in Figure 2.12 causes higher Mahalanobis distance in Figure 2.13. On the other hand, the mHMM is able to update its structure to add new states successfully by calculating a statistical distance between the current forces and known states. Since the mHMM has a new state to represent an unknown condition, the Mahalanobis distance between the incoming data and the new state is less than UCL, as shown in Figure 2.14.

Although the estimation delay from state S_1 to state S_2 causes some non-consecutive data points to be out of control, these points are not statistically significant to add another state in the mHMM. However, the appearance of unknown states after the 2800th sample does trigger the mHMM to add another state, resulting in Figure 2.14. It is critical to diagnose coolant shortage as early as possible to avoid excessive tool wear, as shown in Figure 2.11(b). The appearance of unknown states can be identified through the emergence of a new state in the mHMM. Figures 2.13 and 2.14 illustrate that the mHMM is not only a stochastic modeling technique but also an adaptive fault detector.

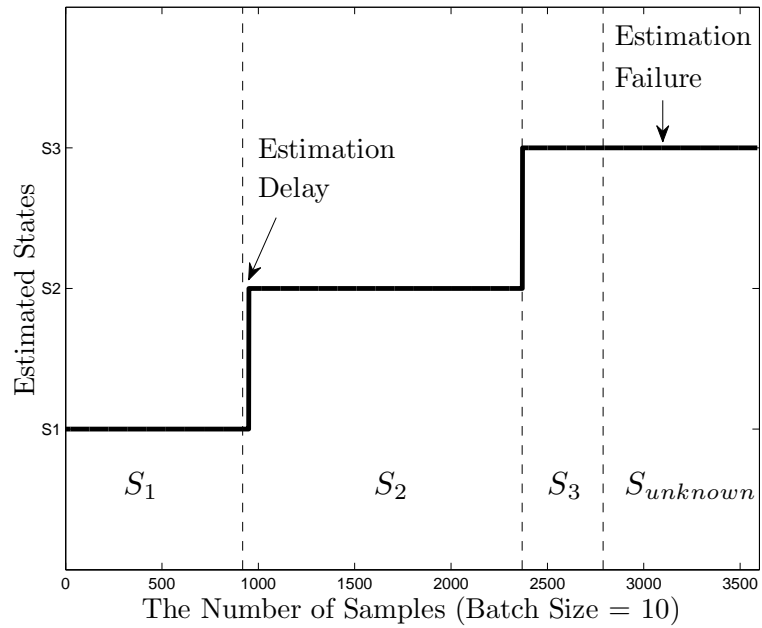


Figure 2.12: Estimated states (vertical dashed lines indicate true states while solid lines represent estimated states)

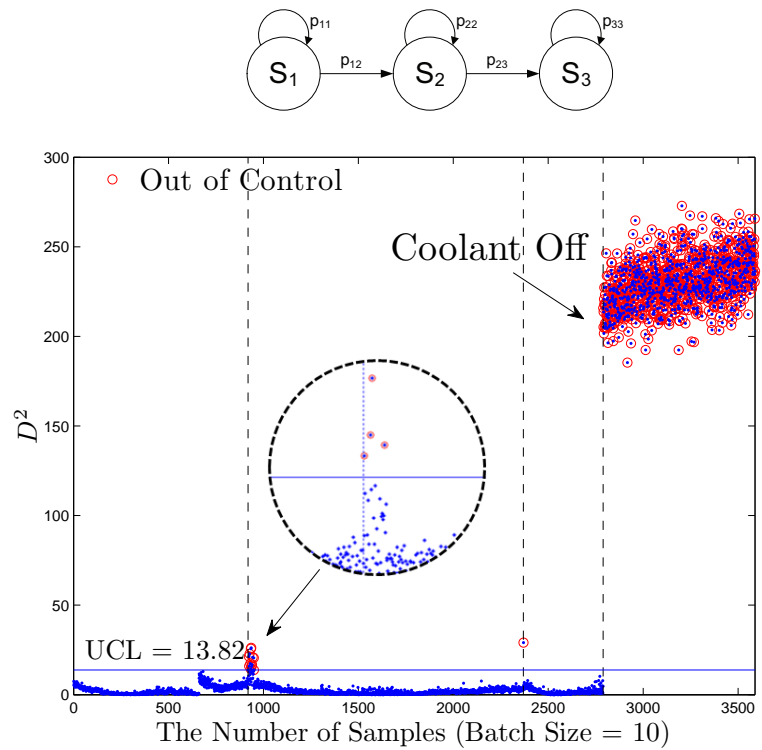


Figure 2.13: Control chart with a conventional HMM

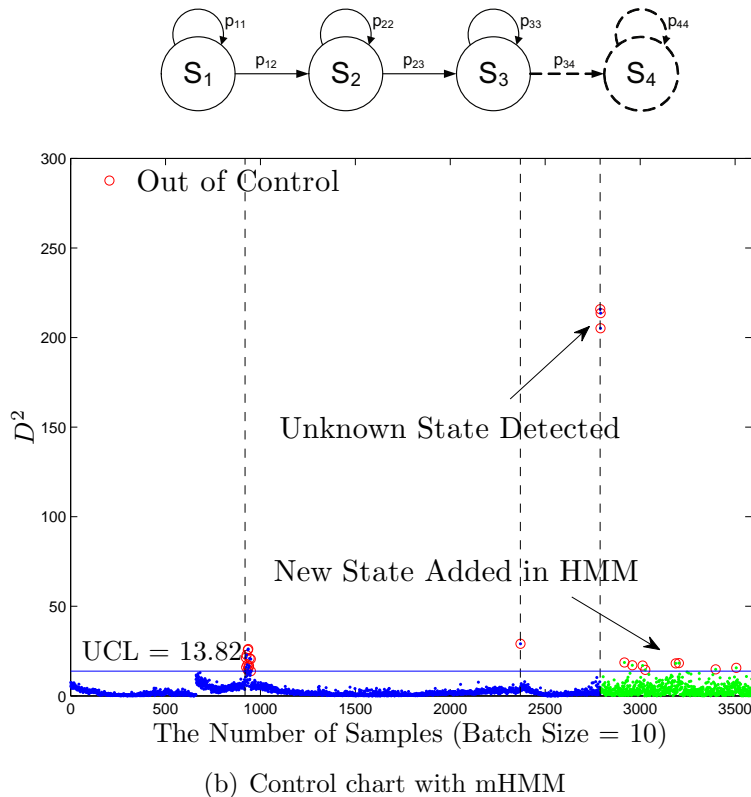


Figure 2.14: Control chart with a mHMM

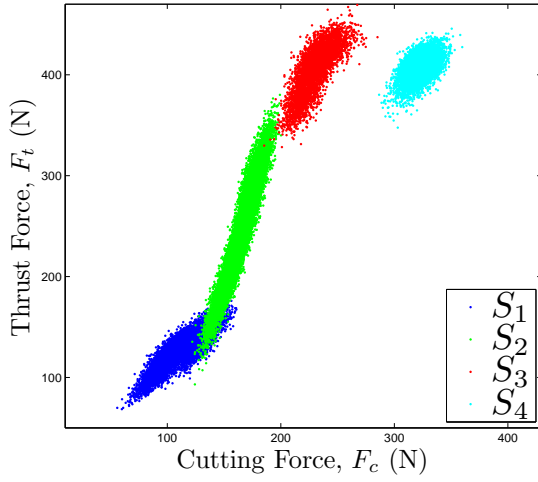
We have also compared the proposed mHMM with other clustering algorithms such as neural networks, Gaussian Mixture Model (GMM), and K-means clustering [21, 25]. Artificial neural networks are motivated by biological neural networks and have been used extensively over the past three decades for both classification and clustering [54]. GMM is based on the idea that the data can be clustered using a mixture of multivariate Gaussian distributions. On the other hand, K-means is the simplest and most commonly used algorithm. K-means starts with a random initial partition and keeps re-assigning the patterns to clusters based on the similarity between the patterns and the cluster centers until a convergence criterion is met [55]. The mHMM algorithm is based on online data streaming, which is more applicable in online equipment condition diagnosis, while GMM and K-means clustering approaches are based on offline but unsupervised machine learning. Note that we do not compare

a mHMM with reinforcement learning versions of neural networks, GMM, and K-means.

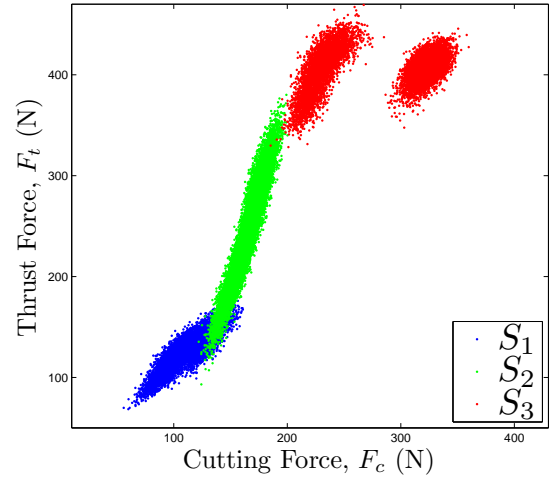
The classification results of the five different algorithms are illustrated in Figure 2.15 and summarized in Table 2.4. The accuracies are calculated by counting errors between the true state and the state estimated via the clustering algorithms. The mHMM clearly outperforms the others in terms of estimation accuracy because the mHMM makes use of information regarding the transition probability as well as observation symbol distributions.

Table 2.4: Correct estimation rate comparison

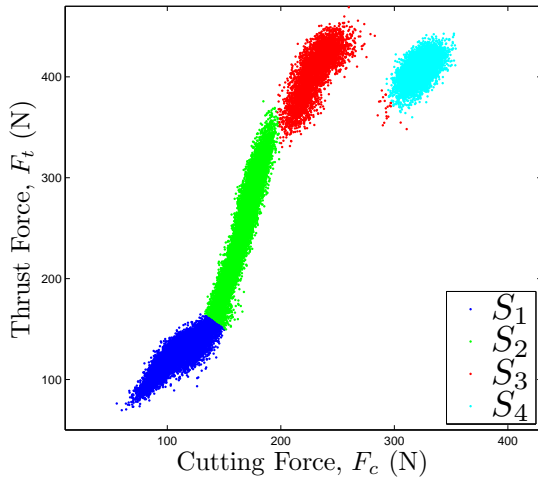
Estimation methods	Accuracy (%)
mHMM	99.06
HMM	77.69
Neural Networks	97.71
GMM	96.49
K-means	92.81



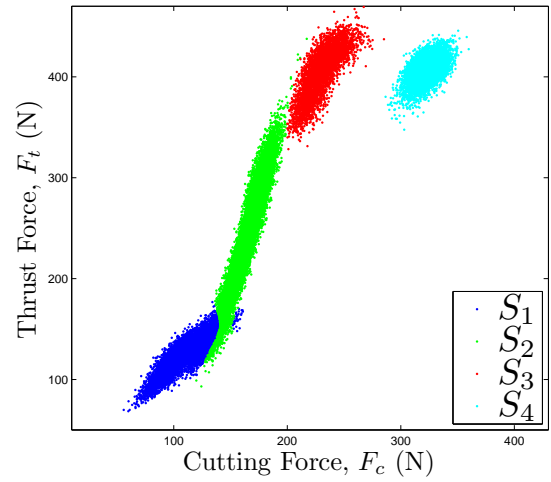
(a) mHMM



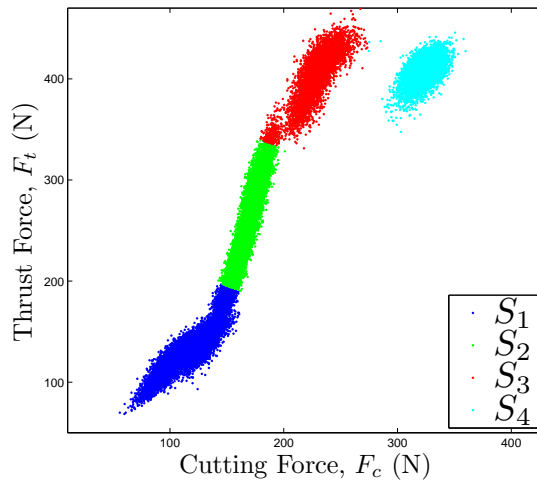
(b) HMM



(c) Neural Network



(d) GMM



(e) K-means

Figure 2.15: The estimated states via various algorithms

The mHMM enables the identification of anomalous behavior of a system by measuring Mahalanobis distance. We have shown that the proposed mHMM algorithm is successfully able to modify its structure by increasing the number of states and estimate the state of a system even in the existence of an unknown state.

2.4 Conclusion and Future work

In this chapter, the modified Hidden Markov Model (mHMM) algorithm is developed to deal with variable state space. A method in SPC has been combined into the mHMM for unknown state detection and diagnosis. The results illustrate that the proposed mHMM can 1) estimate current tool conditions more effectively than other classification algorithms such as GMM, K-means, and neural network; 2) detect anomalous behavior or an unknown state at an early stage by using the Hotelling multivariate control chart; and 3) change its structure to represent degradation processes more accurately in the presence of unknown faults.

Future work will involve further experimental validations of the mHMM algorithm. Furthermore, the mHMM needs to be modified to handle general distributions. The assumption that the monitoring signals follow the Gaussian density distribution has to be released to solve the systems which have different observation symbol probability distributions.

CHAPTER III

Markov-based Preventive Maintenance Planning With Repair Time and Periodic Inspection

3.1 Introduction

Maintenance affects many aspects of manufacturing: productivity, product quality, maintenance cost, etc. Unplanned downtime of equipment might not only reduce line productivity but also affect the quality of products. In addition, system failures will increase maintenance expenses due to unpredictable maintenance. Therefore, determining when to maintain a system before its failure is one of the critical problems in manufacturing plant floors [10, 56, 57, 58]. For preventive maintenance (PM), we should find an appropriate frequency of PM to reduce unnecessary maintenance costs and increase system reliabilities by developing the mathematical models which can represent both degradation processes and maintenance actions.

Among a number of mathematical modeling techniques, Markov process models are widely used to describe the dynamic and stochastic behaviors of equipment degradation processes [59, 60, 61, 62, 63, 64]. Chan et al. [59] found the optimal maintenance policy that maximized the availability of a component subject to random failure and degradation through Markov processes. Chen et al. [60] further developed the Markov model from [59] to perform minimal and major maintenance with re-

spect to equipment degradation conditions. Maillart [61] examined the maintenance-related imperfect observation information problem using Partially Observed Markov Decision Processes (POMDP). However, these Markov models found in the previous literature [59, 63, 64] are not accurate enough to represent a variety of maintenance activities with appropriate random distributions. The non-exponential sojourn time distributions between discrete states cannot be modeled due to the memory-less property of a Markov chain. Instead, the exponential distributions with the same mean values have been used to approximate the effect of non-exponential holding time distributions. The inspection duration, maintenance duration, and time interval between inspections are, for instance, assumed to follow the exponential distributions in most of the previous Markov models [59, 62, 63, 64]. Even though semi-Markov processes have been employed to model degraded systems by allowing the holding time distributions to be non-exponential, it is generally assumed that the mathematical formulations of semi-Markov models [60, 65, 66] are so complicated that they are not analytically tractable.

In general, a system consists of more than a single unit. If all units in the system are stochastically independent of one another, a maintenance policy for the single unit model [63, 64] may be applied to multi-unit maintenance problems. On the other hand, if any units in the system are stochastically dependent on each other, then an optimal decision for maintenance of one unit is not necessarily the optimum for the entire system [67, 68]. A decision must be made to improve the entire system rather than only a single subsystem. Therefore, we must also investigate optimal maintenance policies for a multi-unit system, where units may or may not depend on each other. Although the complexity of a multi-unit system poses challenges in finding optimal maintenance policies, the development of such a model may introduce

an opportunity for group replacement of several components provided that a joint replacement cost is less than that of the separate replacements of the components [69, 70, 71].

Therefore, in this chapter, we examine the problem of maintenance decision-making for single and two-unit systems subject to random degradation processes. We establish a novel modeling technique to approximate non-negligible repair times and periodic inspection within a Markov chain framework for more realistic maintenance activities. Then, the Markov model is used to find the optimal preventive maintenance intervals which maximize the system performances such as availability, productivity or profit. In addition, dependencies among different units in the two-unit system have been considered in the decision model in the presence of the multiple preventive maintenance tasks.

3.2 Modeling of Maintenance Policies for a Single Machine

3.2.1 Modeling of Machine Degradation Processes with Markov Process

We will first use a single unit system to analyze the degradation process, and then broaden our scope to a two-unit system in Section 3.3. A Markov process with three discrete states (see Figure 3.1) is used to model the degradation process under the following assumptions:

- Three discrete states (S_1 : fully operational, S_2 : degraded but still operational, and S_3 : failed) are used to represent the equipment degradation status. However, the number of states can be easily changed, depending on the degree of model specificity.
- A system degrades gradually so that the state transition diagram is “linear,” as

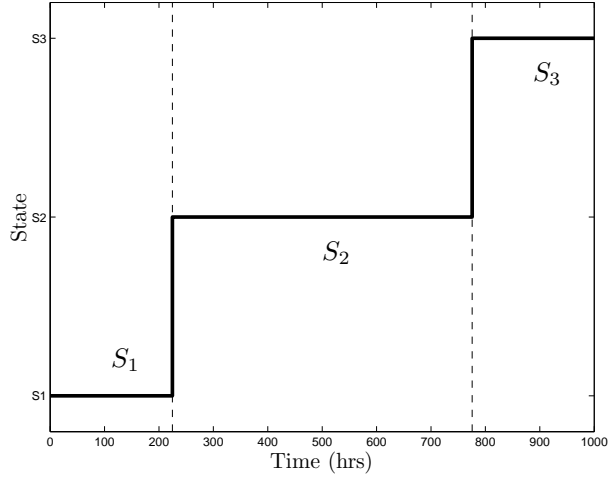
shown in Figure 3.1.

- The failure rates λ_1 and λ_2 are constant between states.

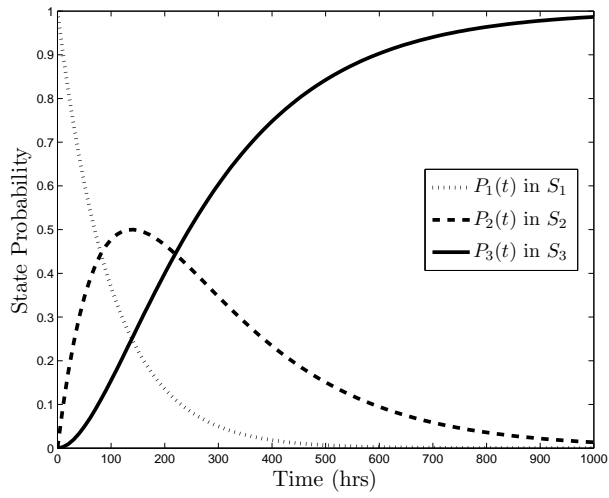


Figure 3.1: A state transition diagram and its transition rate matrix

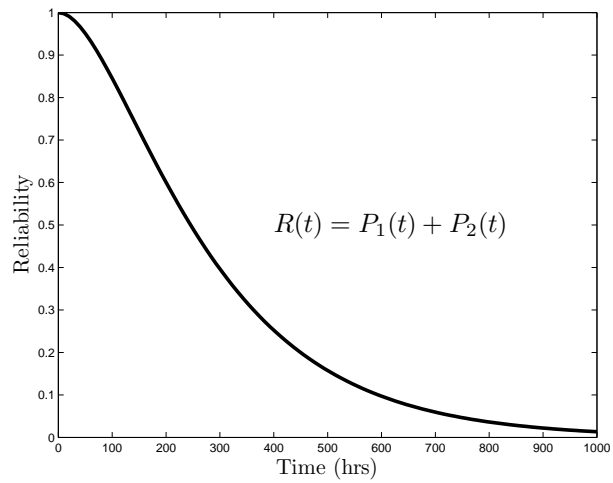
As displayed in Figure 3.2 in the case of $\lambda_1 = 0.010$ and $\lambda_2 = 0.005$ the corresponding sample path (a), state probabilities (b) and reliability function (c) of the above Markov degradation model are then generated. The reliability function is calculated by summing up all the state probabilities except the failed state, S_3 . Reversely, the Markov process can also be derived to approximate a given reliability function. Note that although we use only three states to model degradation processes via a Markov chain, the number of states can be changed to better approximate degradation processes if necessary.



(a) Sample path



(b) State probability



(c) Equivalent reliability function

Figure 3.2: Simulation results for single unit degradation process

3.2.2 Approximation of Constant Time Delay in Markov Process

A degraded system will eventually fail, requiring repair or replacement. Hence, it is important to have a model that can represent maintenance effects on machine condition as well as the degradation process. However, many actual degraded systems with maintenance may involve state transitions, which depend explicitly on time, or occur discretely. For these reasons, maintenance actions cannot generally be modeled by a simple exponential distribution within the Markov process. For example, a non-negligible repair time or periodic inspection, which is necessary for modeling of more realistic maintenance activities, does not generally follow the exponential distribution.

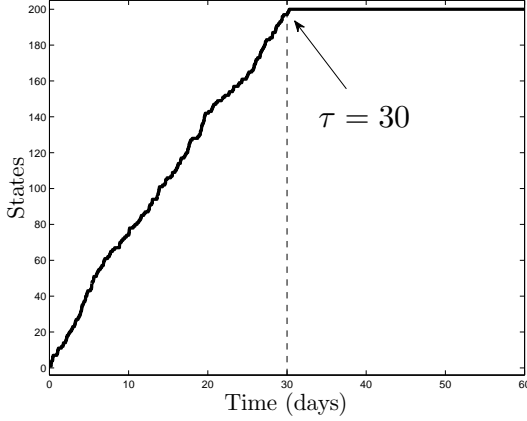
Therefore, we have to develop the approximation methodology to allow the Markov processes to model non-negligible holding times. Since a Markov model provides analytical ways to calculate any state probabilities of interest in a closed form, we can use this Markov model to find the optimal PM intervals with respect to various objectives such as availability, productivity, and profit. The concept of a phase-type distribution [72, 73, 74] can be used to approximate a time delay until absorption to one of the states in the Markov chain. It is also known that the Erlang process (i.e., summation of identical exponential distributions as displayed in Figure 3.3) minimizes the variance among any phase type distributions [73]. In other words, non-exponential holding time distributions can be approximated by inserting multiple intermediate states between the two degradation states. This Erlang process approximation of the constant time delay in the Markov process enables the incorporation of various maintenance activities into the equipment degradation.

Figure 3.3 illustrates that the Markov process with $m = 200$ realizes a constant time delay ($\tau = 30$ days), and its probability density function is similar to a step function of $\tau = 30$, respectively. The application of a constant time delay has been

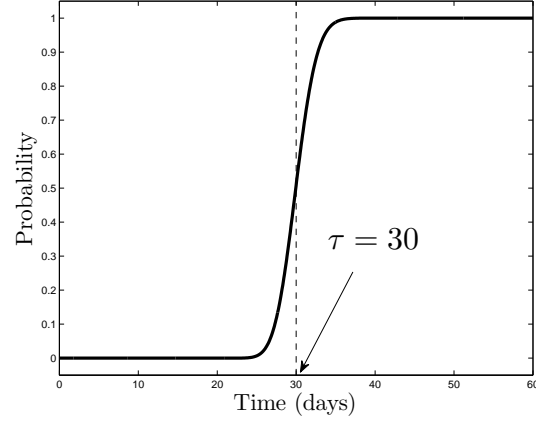
implemented for modeling of repair and periodic inspections in the following maintenance models while preserving the Markov property.



(a) Markov chain for Erlang process



(b) Sample path



(c) Probability of delay time

Figure 3.3: Simulation result from Erlang process with 200 intermediate states

3.2.3 Maintenance Model for Single Unit System with Multiple Maintenance Tasks

We consider a discrete multi-stage degradation, where the first state is an “as good as new” state and the last state is the failed state. The system is subject to periodic inspection that identifies the degree of degradation. After an inspection, based on the degree of degradation, an appropriate PM task is determined and performed with the corresponding repair time, T_{PM} . Failure of the system can be identified immediately. If the system fails before the next planned inspection, the system is restored to the initial condition through RM for the time, T_{RM} . Then, the inspection will be rescheduled. It is assumed that the system is as good as new after any type of maintenance is conducted. This maintenance policy is illustrated in Figure 3.4.

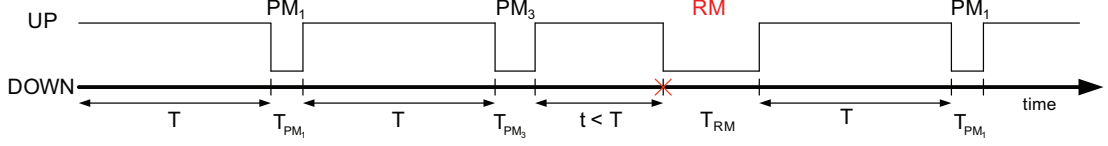


Figure 3.4: Illustration of the maintenance policy

The system and maintenance policy model assumes that the unit will be inspected after T time units of operation, the i^{th} PM task is performed for the time T_{PM_i} , and the i^{th} PM task is requested with the probability of q_i after the inspection. Then, the Markov process for the above maintenance policy can be modeled as illustrated in Figure 3.5. The states here are representing:

$$S_{1i} : \text{Fully operational} \quad (1 \leq i \leq m)$$

$$S_{2i} : \text{Degraded but still operational} \quad (1 \leq i \leq m)$$

$$S_{3i} : \text{PM}_1 \quad (1 \leq i \leq m)$$

$$S_{4i} : \text{PM}_2 \quad (1 \leq i \leq m)$$

$$S_{5i} : \text{PM}_3 \quad (1 \leq i \leq m)$$

$$S_{6i} : \text{RM} \quad (1 \leq i \leq m)$$

The Markov process is created by stacking the Erlang processes (shown in Figure 3.3(a)) on top of the degradation model shown in Figure 3.1. Since the machine failure is self-announcing, all the failure states with different times are combined into a single state, S_f . The associated transition rates of $\mu, \mu_{PM_1}, \mu_{PM_2}, \mu_{PM_3}$, and μ_{RM} are also displayed in Figure 3.5.

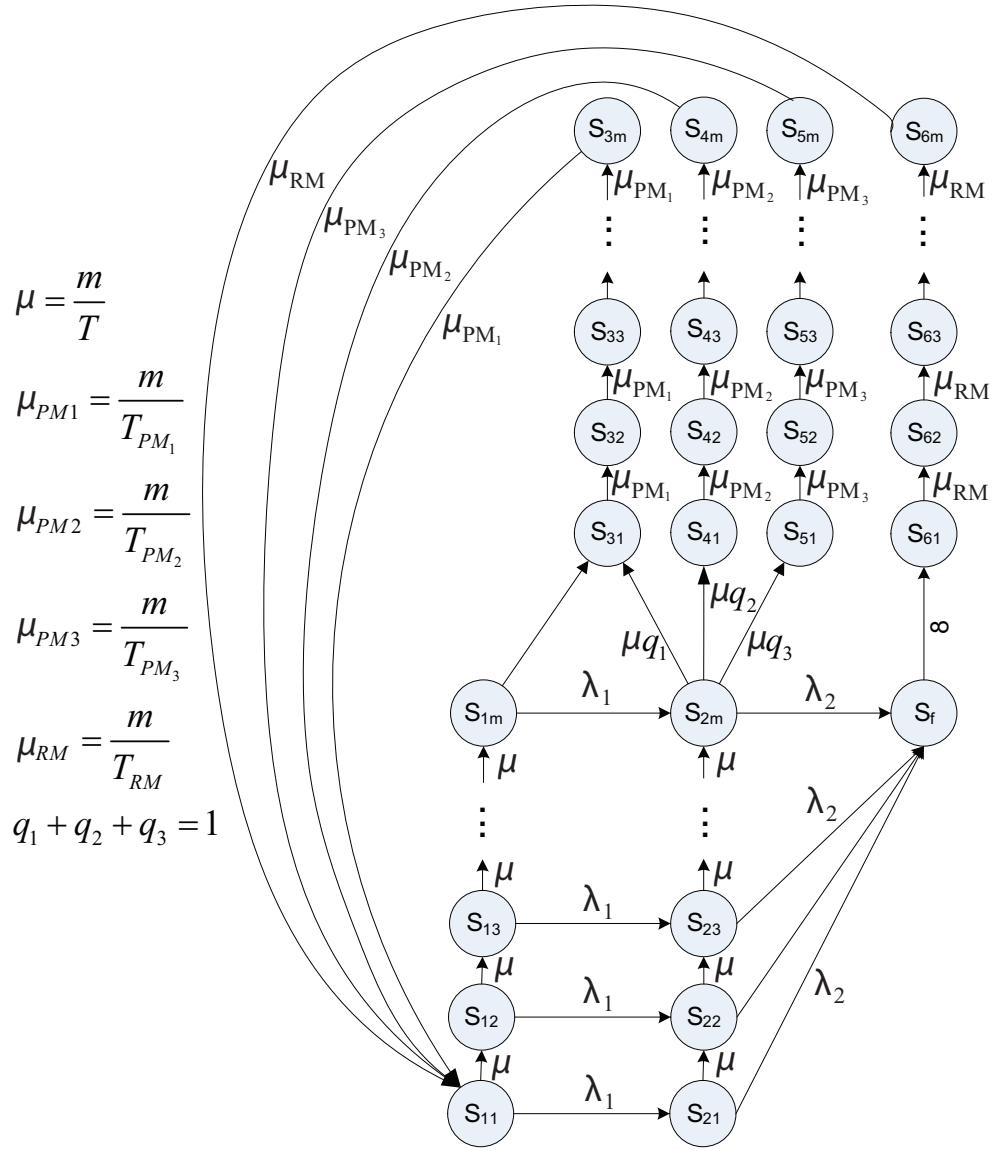


Figure 3.5: Markov process for the abovementioned maintenance policy

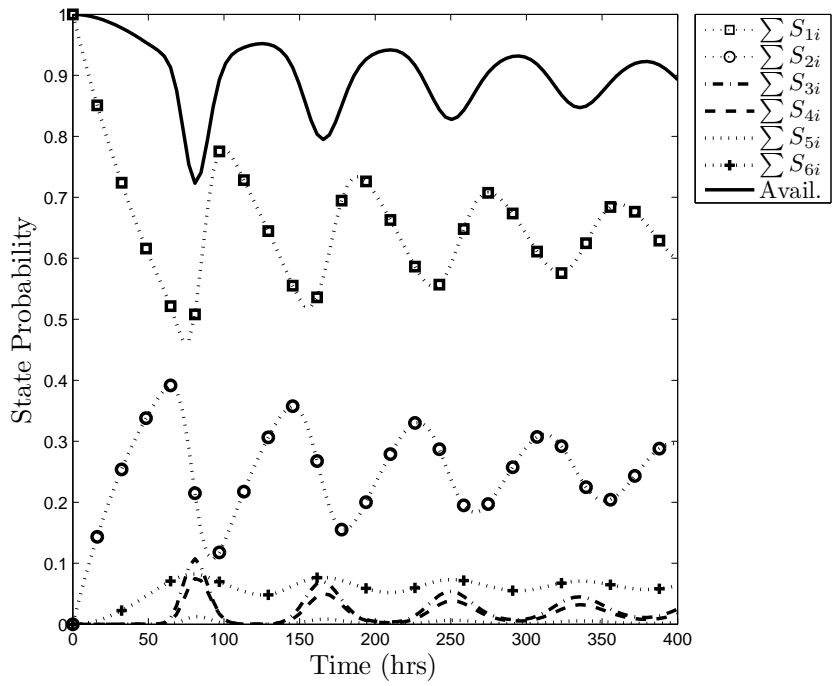


Figure 3.6: State probability for the abovementioned maintenance policy

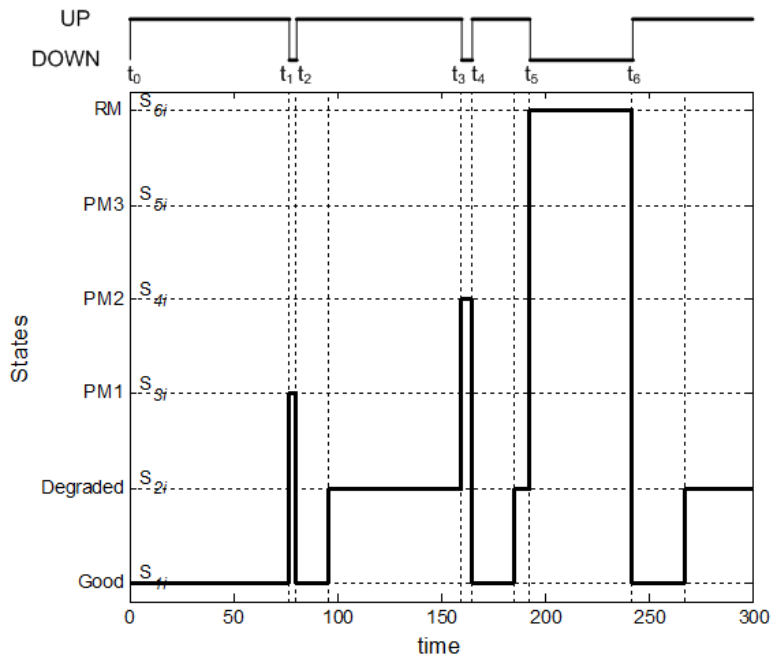


Figure 3.7: Sample path for the abovementioned maintenance policy

Table 3.1: Event log table for the sample path in Figure 3.7

	Time	Event	Event Duration	Ideal Event Duration
t_0	0			
t_1	76.68	Inspection	76.68	80
t_2	80.33	PM ₁	3.65	4
t_3	159.74	Inspection	79.42	80
t_4	164.61	PM ₂	4.86	5
t_5	192.33	Failure	27.73	Less than 80
t_6	242.07	RM	49.73	50

The simulation results of state probabilities and a sample path are illustrated in Figures 3.6, 3.7, and Table 3.1, respectively. We set $T = 80$, $T_{PM_1} = 4$, $T_{PM_2} = 5$, $T_{PM_3} = 6$, T_{RM} , $q_1 = 0.2$, $q_2 = 0.7$, $q_3 = 0.1$, $\lambda_1 = 0.010$, and $\lambda_2 = 0.005$ in this numerical case.

For this example of sample paths, PM₁ has performed for 3.65 time units after the first inspection at 76.68 time units because the machine is in good condition. At the second inspection, PM₂ is conducted for 4.86 time units because the machine is in degraded condition. The machine is then found to have failed before the next scheduled inspection so that RM is immediately executed for 49.73 time units. These simulation results show us that the proposed Markov chain well models the designed maintenance policy.

Since this maintenance model assumes that the current state of equipment subject to stochastic failure is unknown unless an inspection is carried out, the maintenance optimization problem of finding an optimal inspection interval based on system performance measurements needs to be examined. One advantage of using Markov processes is that we are able to calculate any state probabilities of interest in a closed form [42]. Let us assume that the only performance criterion of interest is the availability of the system $A(t)$, defined as the probability that the system is functioning at

time t . Steady-state system availability is then equal to $A(\infty) = \lim_{m \rightarrow \infty} \sum_{i=1}^m (P_{1i} + P_{2i})$ in the Markov process in Figure 3.5. The availability of the system will depend on the value of inspection interval T , given other system parameters such as λ_1 , λ_2 , μ , μ_{PM_1} , μ_{PM_2} , μ_{PM_3} , and μ_{RM} . In other words, the controllable variable T can change the system availability. Then, the following Equation (3.1) gives us the optimal inspection interval, T^* that maximizes the steady-state availability of a given system:

$$\begin{aligned} & [(\lambda_1 + \lambda_2)(T_{PM} - T_{RM}) + \lambda_1 T_{RM}] e^{\lambda_1 T} - \\ & [(\lambda_1 + \lambda_2)(T_{PM} - T_{RM}) + \lambda_2 T_{RM}] e^{\lambda_2 T} + (\lambda_1 - \lambda_2)(T_{PM} - T_{RM}) = 0, \end{aligned} \quad (3.1)$$

where $T_{PM} = q_1 T_{PM_1} + q_2 T_{PM_2} + q_3 T_{PM_3}$

Proof. Equation (3.1) is derived by solving the balance equations of the Markov process and setting the derivative of $A(\infty)$ equal to zero. The steady-state system balance equations are:

$$(\mu + \lambda_1)P_{1i} = \mu P_{1i-1}, \quad 2 \leq i \leq m \quad (3.2)$$

$$(\mu + \lambda_1)P_{11} = \mu P_{1m} + \mu_{PM_1} P_{3m} + \mu_{PM_2} P_{4m} + \mu_{PM_3} P_{5m} + \mu_{RM} P_{6m} \quad (3.3)$$

$$(\mu + \lambda_2)P_{21} = \lambda_1 P_{11} \quad (3.4)$$

$$(\mu + \lambda_2)P_{2i} = \lambda_1 P_{1i} + \mu P_{2i-1}, \quad 2 \leq i \leq m \quad (3.5)$$

$$\mu_{\text{PM}_1} P_{31} = \mu P_{1m} + \mu q_1 P_{2m} \quad (3.6)$$

$$\mu_{\text{PM}_1} P_{3i} = \mu_{\text{PM}_1} P_{3i-1}, \quad 2 \leq i \leq m \quad (3.7)$$

$$\mu_{\text{PM}_2} P_{41} = \mu q_2 P_{2m} \quad (3.8)$$

$$\mu_{\text{PM}_2} P_{4i} = \mu_{\text{PM}_2} P_{4i-1}, \quad 2 \leq i \leq m \quad (3.9)$$

$$\mu_{\text{PM}_3} P_{51} = \mu q_3 P_{2m} \quad (3.10)$$

$$\mu_{\text{PM}_3} P_{5i} = \mu_{\text{PM}_4} P_{5i-1}, \quad 2 \leq i \leq m \quad (3.11)$$

$$\mu_{\text{RM}} P_{61} = \lambda_2 (P_{21} + \cdots + P_{2m}) \quad (3.12)$$

$$\mu_{\text{RM}} P_{6i} = \mu_{\text{RM}} P_{6i-1}, \quad 2 \leq i \leq m \quad (3.13)$$

Then, the steady-state availability in this Markov process is equivalent to

$$A(\infty) = \lim_{m \rightarrow \infty} \sum_{i=1}^m (P_{1i} + P_{2i}) = \frac{\lambda_2^2 - \lambda_1^2 - \lambda_2^2 e^{-\lambda_1 T} + \lambda_1^2 e^{-\lambda_2 T}}{\lambda_2^2 - \lambda_1^2 - \lambda_2^2 e^{-\lambda_1 T} + \lambda_1^2 e^{-\lambda_2 T} + \lambda_1 \lambda_2 (\lambda_2 - \lambda_1) T_{\text{RM}} + \lambda_1 \lambda_2 (T_{\text{PM}} - T_{\text{RM}}) (\lambda_2 e^{-\lambda_1 T} - \lambda_1 e^{-\lambda_2 T})} \quad (3.14)$$

By definition, an optimal PM Policy is one that maximizes the steady-state availability of a given system. Hence, we wish to maximize Equation (3.14) by appropriate selection of T , which is the time interval between PMs. Taking the derivative with respect to T , and setting it equal to zero, we obtain

$$\begin{aligned} 0 &= \frac{dA(\infty)}{dT} = \\ & \frac{d}{dT} \left\{ \frac{\lambda_2^2 - \lambda_1^2 - \lambda_2^2 e^{-\lambda_1 T} + \lambda_1^2 e^{-\lambda_2 T}}{\lambda_2^2 - \lambda_1^2 - \lambda_2^2 e^{-\lambda_1 T} + \lambda_1^2 e^{-\lambda_2 T} + \lambda_1 \lambda_2 (\lambda_2 - \lambda_1) T_{\text{RM}} + \lambda_1 \lambda_2 (T_{\text{PM}} - T_{\text{RM}}) (\lambda_2 e^{-\lambda_1 T} - \lambda_1 e^{-\lambda_2 T})} \right\} \\ \therefore & [(\lambda_1 + \lambda_2)(T_{\text{PM}} - T_{\text{RM}}) + \lambda_1 T_{\text{RM}}] e^{\lambda_1 T} - \\ & [(\lambda_1 + \lambda_2)(T_{\text{PM}} - T_{\text{RM}}) + \lambda_2 T_{\text{RM}}] e^{\lambda_2 T} + (\lambda_1 - \lambda_2)(T_{\text{PM}} - T_{\text{RM}}) = 0 \end{aligned} \quad (3.15)$$

□

Figure 3.8, for instance, shows that the optimal inspection interval is equal to 1815 time units if system and maintenance parameters are as shown in the figure.

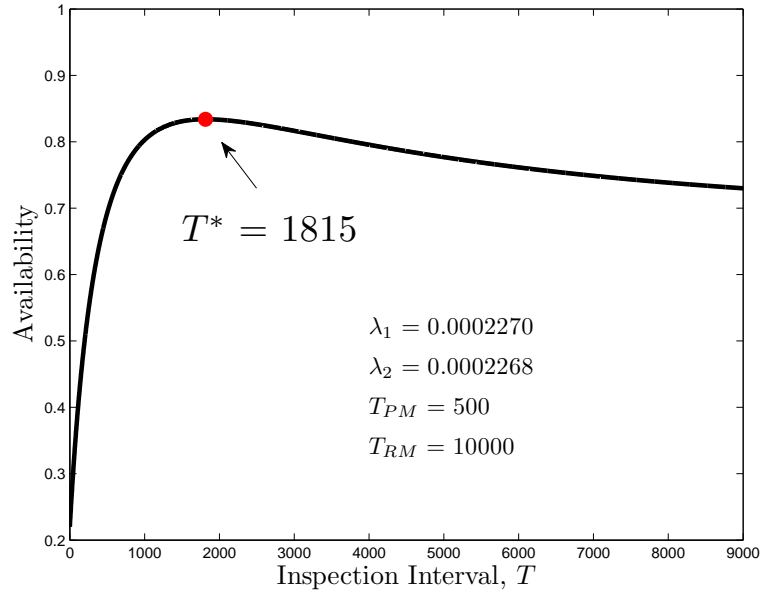


Figure 3.8: Availability as a function of inspection intervals

We are interested in not only finding the optimal PM interval but also investigating how much a PM interval T is sensitive to the system parameters (for instance, λ_1, λ_2 for degradation processes and T_{PM}, T_{RM} for repair times). The sensitivity can be analyzed from Equation (3.1) by taking a partial derivative of T with respect to system parameters of interest. The analytical sensitivities in a closed form are given in Equations (3.16), (3.17), (3.18), and (3.19):

$$\frac{\partial T}{\partial \lambda_1} = \frac{-[T_{PM} + (\lambda_1 + \lambda_2)(T_{PM} - T_{RM})T + \lambda_1 T_{RM}T]e^{\lambda_1 T} + (T_{PM} - T_{RM})(e^{\lambda_2 T} - 1)}{[\lambda_1 T_{PM} + \lambda_2(T_{PM} - T_{RM})]\lambda_1 e^{\lambda_1 T} - [\lambda_1(T_{PM} - T_{RM}) + \lambda_2 T_{PM}]\lambda_2 e^{\lambda_2 T}} \quad (3.16)$$

$$\frac{\partial T}{\partial \lambda_2} = \frac{(T_{PM} - T_{RM})(1 - e^{\lambda_1 T}) + [T_{PM} + (\lambda_1 + \lambda_2)(T_{PM} - T_{RM})T + \lambda_2 T_{RM}T]e^{\lambda_2 T}}{[\lambda_1 T_{PM} + \lambda_2(T_{PM} - T_{RM})]\lambda_1 e^{\lambda_1 T} - [\lambda_1(T_{PM} - T_{RM}) + \lambda_2 T_{PM}]\lambda_2 e^{\lambda_2 T}} \quad (3.17)$$

$$\frac{\partial T}{\partial T_{PM}} = \frac{(\lambda_1 + \lambda_2)(-e^{\lambda_1 T} + e^{\lambda_2 T}) - (\lambda_1 - \lambda_2)}{[\lambda_1 T_{PM} + \lambda_2(T_{PM} - T_{RM})]\lambda_1 e^{\lambda_1 T} - [\lambda_1(T_{PM} - T_{RM}) + \lambda_2 T_{PM}]\lambda_2 e^{\lambda_2 T}} \quad (3.18)$$

$$\frac{\partial T}{\partial T_{RM}} = \frac{\lambda_2 e^{\lambda_1 T} - \lambda_1 e^{\lambda_2 T} + (\lambda_1 - \lambda_2)}{[\lambda_1 T_{PM} + \lambda_2(T_{PM} - T_{RM})]\lambda_1 e^{\lambda_1 T} - [\lambda_1(T_{PM} - T_{RM}) + \lambda_2 T_{PM}]\lambda_2 e^{\lambda_2 T}} \quad (3.19)$$

3.2.4 Comparison with Conventional Markov Models

All of the previous models discussed in the literature [59, 60, 61] have assumed either negligible or exponentially distributed inspection periods, replacement and repair times, as shown in Figures 3.9 and 3.10. In addition, those systems are not periodically inspected in the strict sense. Instead, they are inspected according to an interval that follows an exponential distribution with a mean value T .

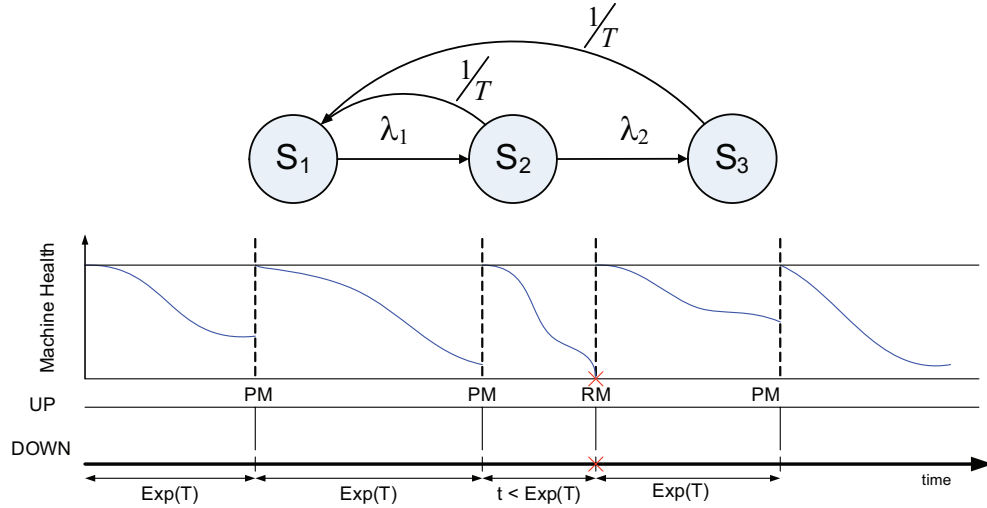


Figure 3.9: Maintenance policy comparison: not periodic and neglect repair time

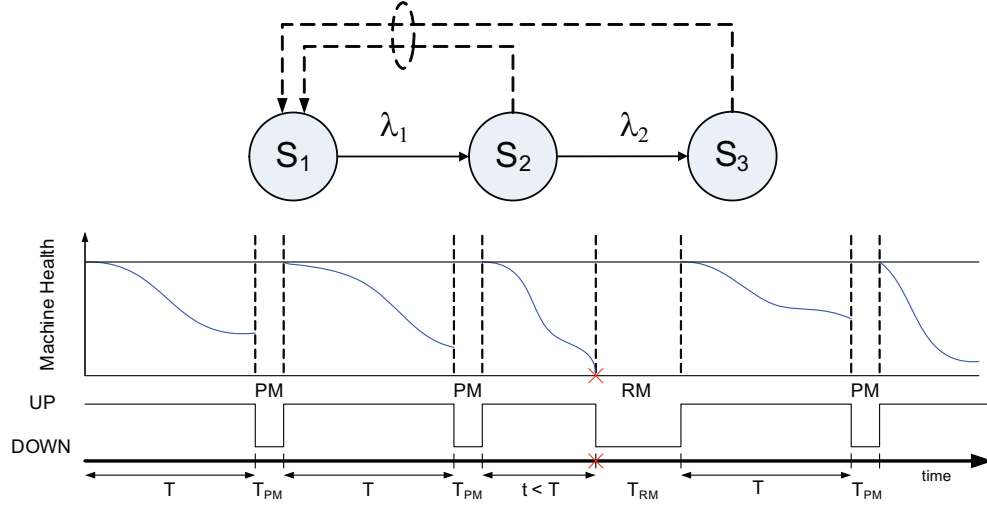


Figure 3.10: Maintenance policy comparison: more realistic model with periodic inspection and non-negligible repair time

In order to see the improvement of the proposed Markov model, two Markov models have been compared in Figures 3.11 and 3.12. To simplify our discussion, we assume that there is only one type of the preventive maintenance task in the model shown in Figures 3.11 and 3.12. The states here are representing:

S_1, S_{1i} : Fully operational ($1 \leq i \leq m$)

S_2, S_{2i} : Degraded but still operational ($1 \leq i \leq m$)

S_3, S_{3i} : PM($1 \leq i \leq m$)

S_4, S_{4i} : RM($1 \leq i \leq m$)

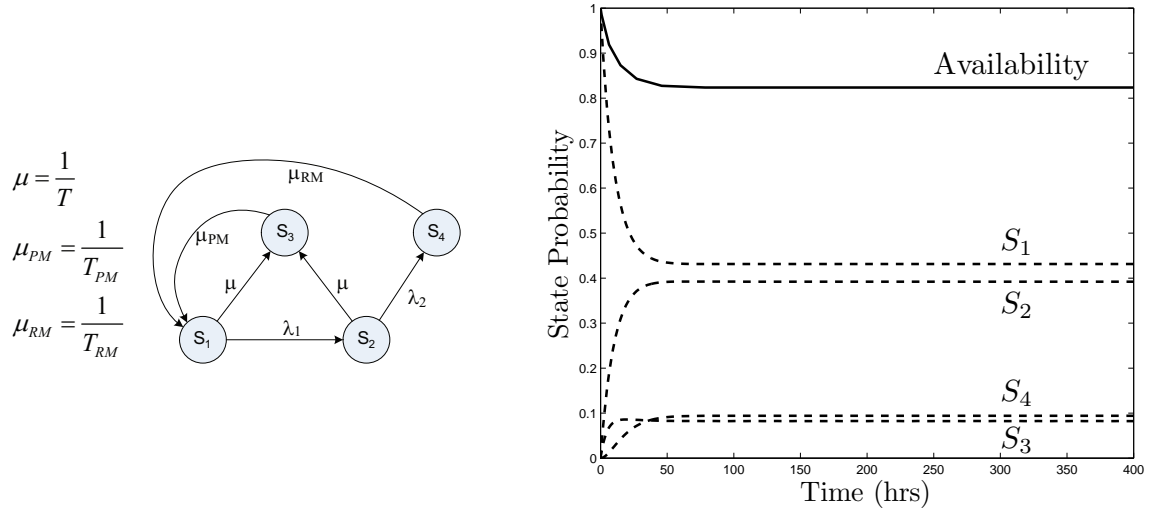


Figure 3.11: A traditional Markov model (left) and its simulation result (right)

The clear discrepancy between the traditional Markov model (Figure 3.11) and the proposed Markov model (Figure 3.12) can be seen, evidenced in the dynamic fluctuations. Although the steady state probabilities between two models are similar, we cannot observe any effects of maintenance activities in Figure 3.11. On the other hand, the proposed Markov model in Figure 3.12 is able to provide detailed information about the system availability decrease during maintenance and availability improvement after maintenance. Since the proposed Markov model can be detailed enough to represent more realistic maintenance characteristics, it can lead to more accurate optimal maintenance decisions. These characteristics may have significant manufacturing implications. For example, if the steady state is reached only after a relatively long settling time, the production system may lose some of its throughput, thus leading to a lower efficiency [75].

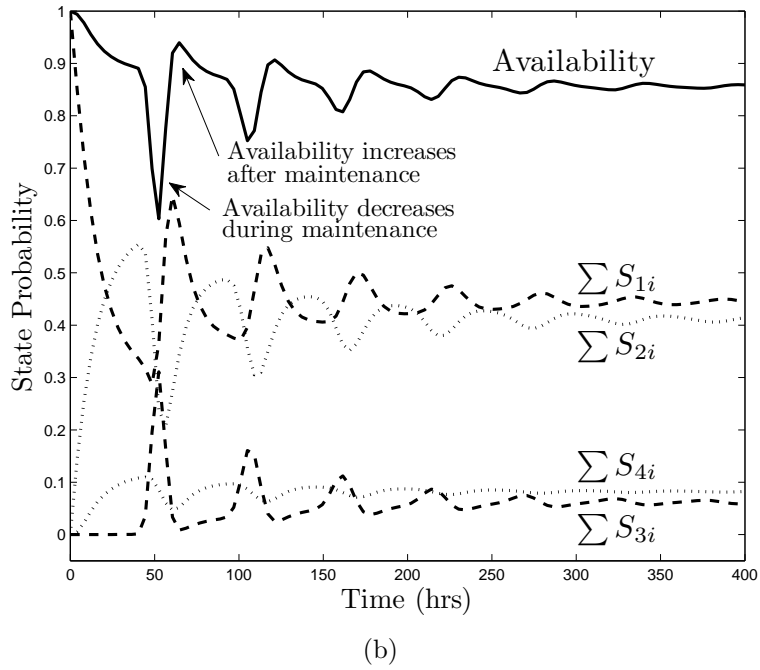
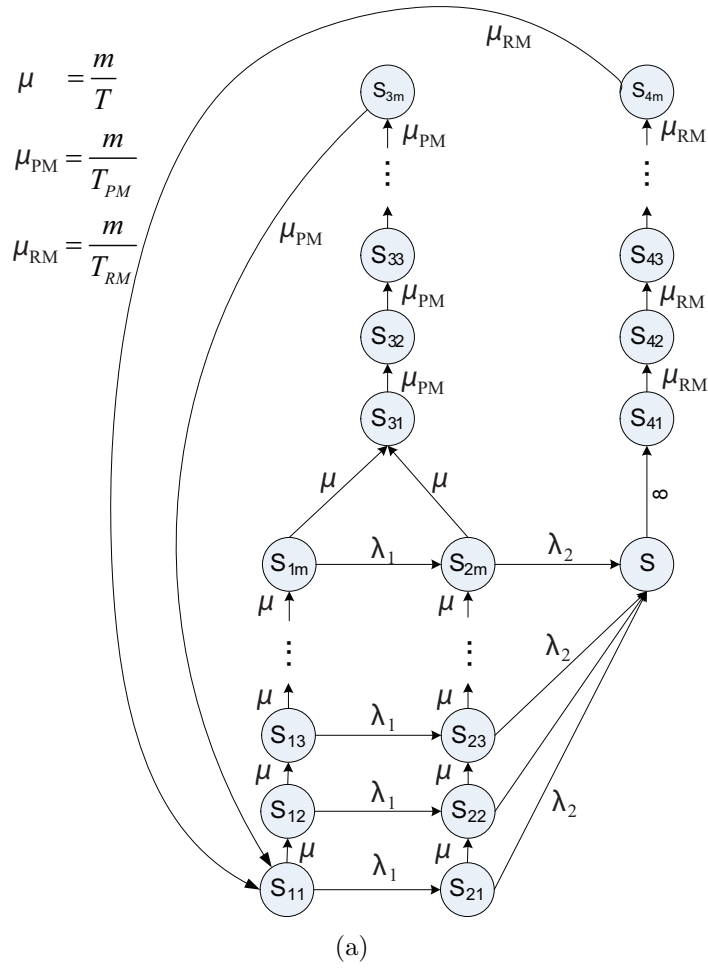


Figure 3.12: The proposed Markov model (a) and its simulation result (b)

3.3 Modeling of Maintenance for a Two Unit System

In this section, we investigate optimal maintenance policies for a two-unit system, where units may or may not depend on each other. Without maintenance, the Markov process of the degradation process for two identical unit systems can be modeled as shown in Figure 3.13 with the nine states. For example, S_{11} represents both M_1 and M_2 as fully operational while S_{33} denotes both M_1 and M_2 as failing. However, the Markov model of a two-unit system will be much more complicated if we consider a maintenance policy.

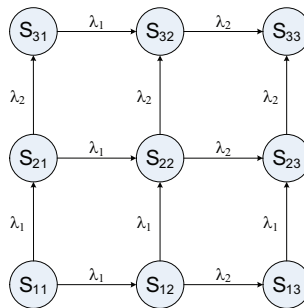


Figure 3.13: Markov degradation model for a two-unit system without maintenance

It is assumed that the time spent on maintenance depends on the machines' condition at the moment of inspection. For instance, the time T_{PM} is required for repair if one of the units is degraded. On the other hand, it will take the time $2 \times T_{PM}$ if both are degraded. Since two different configurations (parallel and serial) are possible with two components as shown in Figure 3.14, both configurations are examined.



Figure 3.14: Parallel (left) and serial (right) configurations

3.3.1 Parallel Configuration

The parallel system in Figure 3.14 can run a production line unless both units fail. Therefore, RM will be performed only when both of the components are down (S_{33}). The units will be inspected after T time units of operation and group repair of two units will then be performed if necessary. The maintenance policy and corresponding model are illustrated in Figures 3.15 and 3.16. The system with two units in parallel configuration requires a different Markov model.

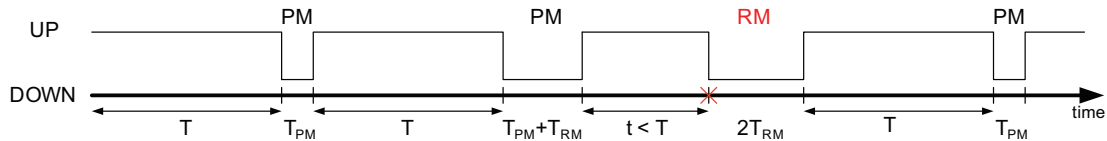


Figure 3.15: Maintenance policy in parallel configuration

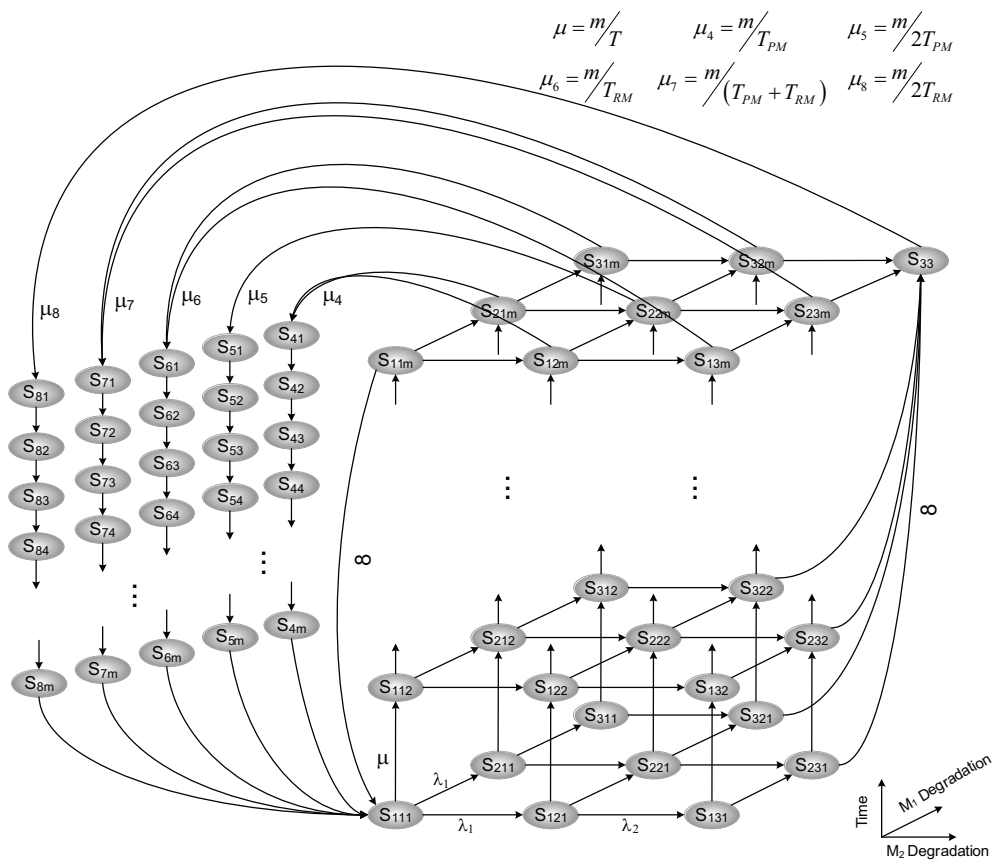


Figure 3.16: Maintenance Markov model for a two unit parallel system

3.3.2 Serial Configuration

In serial configuration, failure of one component will stop the entire production line. Therefore, RM will be conducted whenever one of the components is down ($S_{31}, S_{32}, S_{13}, S_{23}, S_{33}$). The corresponding Markov model of maintenance policy in Figure 3.17 is illustrated in Figure 3.18.

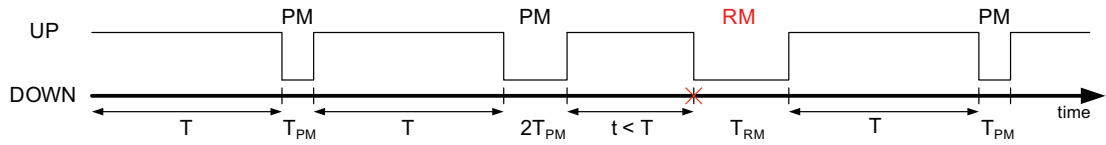


Figure 3.17: Maintenance policy in serial configuration

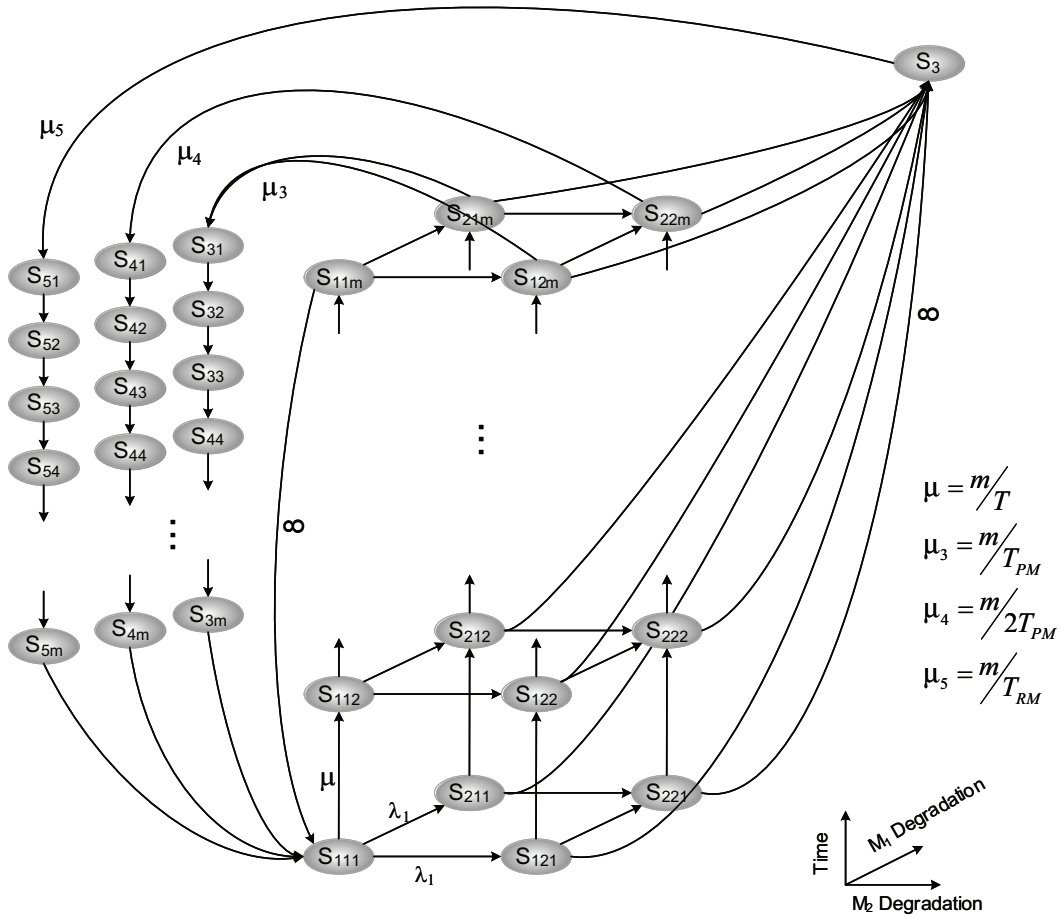


Figure 3.18: Maintenance Markov model for a two-unit serial system

3.3.3 Optimal Inspection Interval

Maximizing the availability of the system can also be the objective of finding the optimal inspection interval in a two-unit system. The availabilities of the system are given by the following equations:

$$\text{Parallel :} \quad A_p(\infty) = \lim_{m \rightarrow \infty} \sum_{(i,j) \neq (3,3)} \sum_{k=1}^m P_{ijk} \quad (3.20)$$

$$\text{Serial :} \quad A_s(\infty) = \lim_{m \rightarrow \infty} \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^m P_{ijk} \quad (3.21)$$

Instead of finding a closed expression for steady-state probabilities P_{ijk} , a numerical method can be used to solve linear equations. By maximizing $A(\infty)$ with respect to T , the optimal time interval between consecutive inspections is determined. Let r be the ratio T_{RM}/T_{PM} . Figure 3.19 shows that optimal maintenance policies depend on the system configuration.

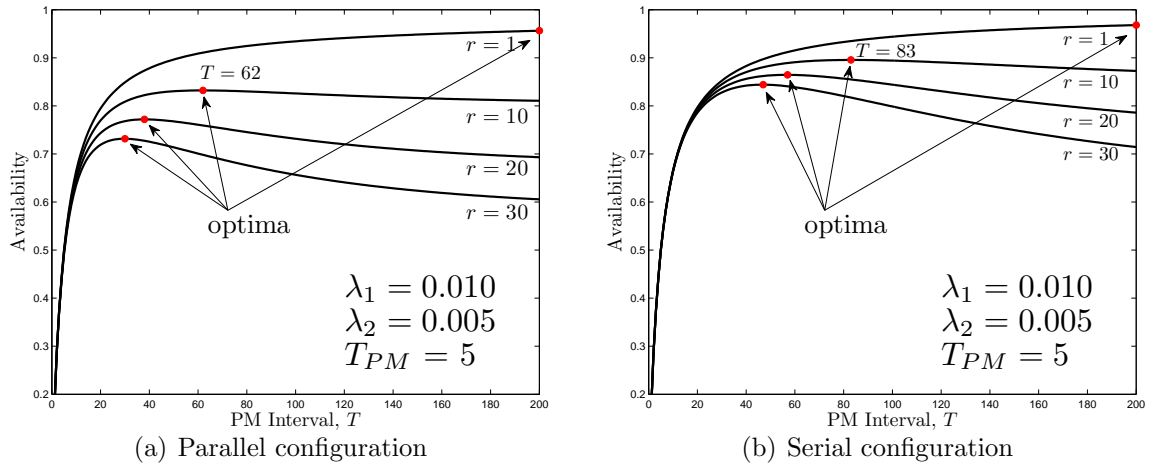


Figure 3.19: Optimal intervals for PM with different connections

These simulation results also show the effect of r on the availability and corresponding optimal inspection intervals. For example, the optimal interval value for the parallel configuration is 62 time units for $r = 10$. The availability declines only

slowly as T exceeds its optimal value; the decrease is much faster if T is less than its optimal value. If the duration for RM is not penalized enough (i.e., $r \approx 1$), the optimal interval value between consecutive inspections will go to infinity. In other words, for this case, running up to failure is the best policy.

In the parallel case, maximization of productivity, $N_p(t)$, rather than the availability of the system is of interest because the system is twice as productive when both components are functional. The productivity of the system can be calculated by Equation (3.22).

$$N_p(\infty) = \lim_{m \rightarrow \infty} \left\{ \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^m 2P_{ijk} + \sum_{i=1}^2 \sum_{k=1}^m P_{i3k} + \sum_{j=1}^2 \sum_{k=1}^m P_{3jk} \right\} \quad (3.22)$$

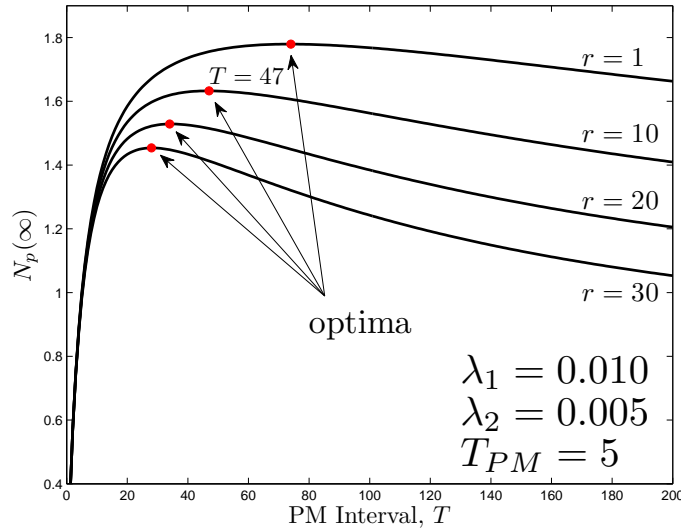


Figure 3.20: Optimal PM interval to maximize system productivity in parallel

Figure 3.19(a) and Figure 3.20 suggest that the optimal interval values for inspection are strongly dependent on criteria of interest. For $r = 10$, the optimal inspection interval for maximizing productivity is 47 time units, which is less than the 62 time units that corresponds to the case of maximizing availability. In other words, main-

tenance will be conducted more frequently. This suggests that we want to keep both of the units out of the failed state in order to achieve maximum productivity.

In this section, we demonstrate that if machines in the system are stochastically dependent on each other, then an optimal decision (i.e., 88.22 time units) on maintenance of single unit system is not necessarily the optimum (i.e., 62 or 83 time units) for the two-unit system. These results show that maintenance policy has to rely on the machine configuration and system performance measure for maintenance decision. This proposed approach can be generalized to analyze the multiple machine system although there are some challenges such as computational complexity due to the inflated system dimensions.

3.4 Case Study with Semiconductor Manufacturing Process Data

In this section, we illustrate and validate our proposed modeling and optimization techniques with a set of industrial data from the semiconductor manufacturing process. Data have been collected from a chamber tool during the chemical vapor deposition process. Nine process parameters and the succeeding metrology measurement are periodically monitored in conjunction with associated process events.

The degradation process has been modeled using a 5-state discrete time Markov chain shown in Figure 3.21. The corresponding probability transition matrix P is obtained from a set of manufacturing process data using a Hidden Markov Model (HMM) [35, 76]. An HMM is selected because it enables us to estimate machine condition from a sequence of measurements (on-wafer particle counts, temperature, pressure, etc.).

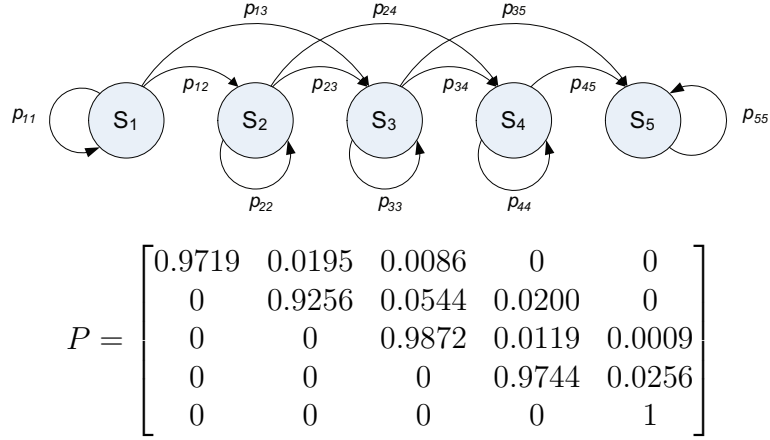


Figure 3.21: 5-state Markov chain with the corresponding transition probability matrix P

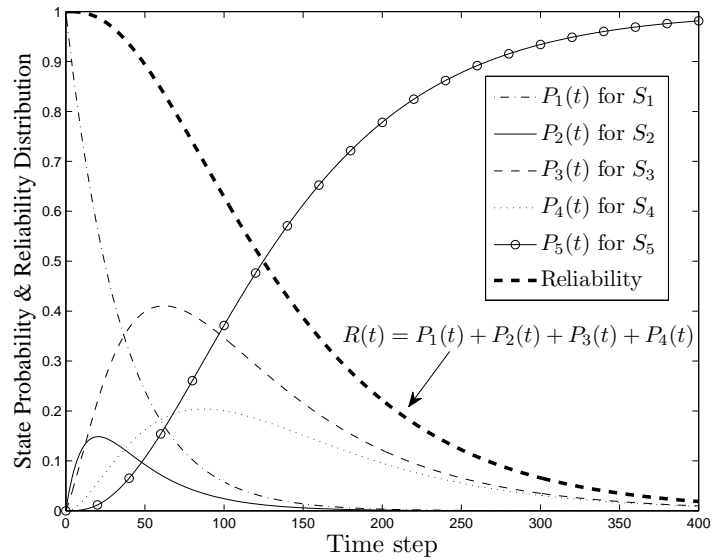


Figure 3.22: State probabilities and reliability distribution

Then, we calculate the reliability of this chamber tool from the Markov chain using the method in Section 3.2.1. The resulting reliability distribution is given in Figure 3.22.

We construct a Markov model for maintenance using the proposed method illustrated in Section 3.2.3 to obtain the inspection interval that maximizes the chamber tool availability. We recommend the inspection interval ($T = 70$ time units), summa-

rized in Table 3.2.

Table 3.2: Benchmark results (r_p : Production per time unit, C_m : Maintenance cost per time unit)

	Inspection Interval, T		Improvement
	60 (current practice)	70 (proposed practice)	
Availability	0.8967	0.9247	3% increase
Productivity Rate	$0.8967 \times r_p$	$0.9247 \times r_p$	3% increase
Maint. Cost Rate	$0.1033 \times C_m$	$0.0753 \times C_m$	27% decrease

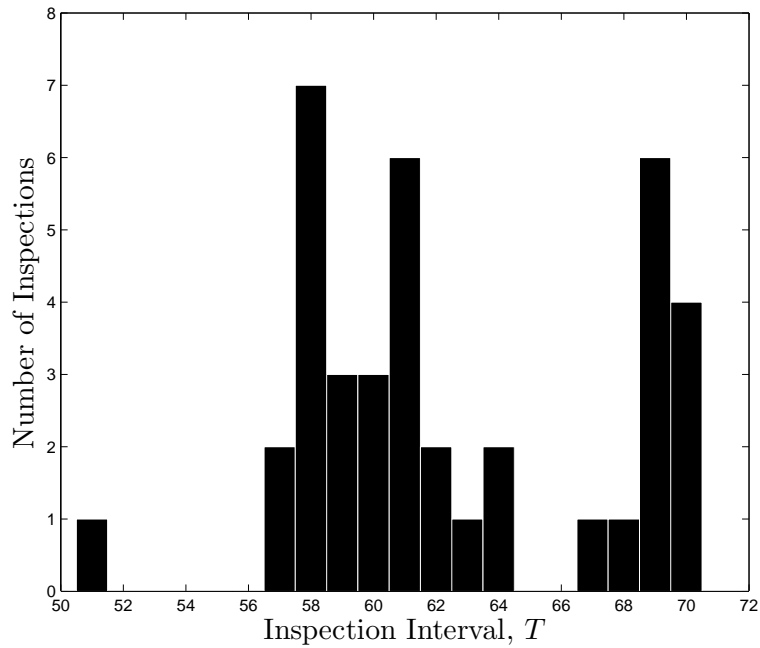


Figure 3.23: Historical inspection intervals from the real Fab data

Manufacturers are generally conservative for production and maintenance planning. We learned that 60 time units has been set as a current PM interval suggested by chamber tool manufacturers, and an actual PM in a production line has been performed at every 62.63 ± 4.97 time units, as displayed in Figure 3.23. However, we recommend $T = 70$, which is longer than what the manufacturer originally recommended ($T = 60$). The proposed inspection interval $T = 70$ results in reduction of the maintenance cost by 27% through avoiding excessive maintenance, while increasing

system availability by 3%. In addition, equipment can be devoted to producing more wafers with good quality.

Furthermore, we have to examine how sensitive the PM interval T is with respect to the maintenance cost, $M(T)$, at the optimal value of $T = 70$.

$$\frac{\partial M}{\partial T} \approx \left. \frac{\Delta M}{\Delta T} \right|_{T=70} \approx \frac{1}{2} \left\{ \frac{|M(70) - M(75)|}{5} + \frac{|M(70) - M(65)|}{5} \right\} = 0.0105 \quad (3.23)$$

Since we learned from historical inspection intervals shown in Figure 3.23 that there is the variation of 4.97 time units in executing PMs with the current interval $T = 60$, we vary 5 time units for the sensitivity analysis. The result from Equation (3.23) shows that the uncertainties near the recommended PM interval $T = 70$ change the maintenance cost by about 1%.

3.5 Conclusion

We have presented a method of obtaining an optimal maintenance inspection policy in a single unit system as well as in a two-unit system. The difference from the previous works lies in the fact that we consider non-negligible repair time and periodic inspections for preventive maintenance. The dynamic system behavior can be monitored and recognized. A constant time repair model will be useful if the mean time to repair information is available or time for repair is almost constant. This is modeled via the Erlang process. With more realistic maintenance characteristics, we have demonstrated the optimal interval for inspection in terms of availability and productivity of the system.

For future work, multiple maintenance tasks will be considered within the multi-unit system. Additionally, the required number of states needed to indicate multiple

degraded systems increases so rapidly that there is a computational limit. Having more than two units in a system might be difficult to solve analytically as the complexity of the proposed method grows exponentially according to the number of components.

CHAPTER IV

Decision Making for Simultaneous Maintenance Scheduling and Production Sequencing

4.1 Introduction

Yield, the percentage of working devices that emerge from a manufacturing process, is an important performance metric for most processes [77]. It is also well known that maintenance is correlated to yield for many processes in a crucial manner [78]. Previous studies [79, 80] discussed the extensive usage of condition monitoring to increase yield in a variety of manufacturing processes. In the semiconductor industry, for example, particulate contamination in equipment is one of major sources of yield loss [81]. Therefore, a significant effort has been delivered to develop condition monitoring techniques that control particulate contamination for enhanced product quality.

Although machine condition information enables us to track and predict levels of equipment degradation, little attention has been paid to modeling an appropriate decision-making to proactively maintain a level of machine condition, unlike the extensive literature on condition monitoring techniques and their applications [29]. For instance, an either time-based or usage-based maintenance scheduling strategy is still dominant for many manufacturing industries [82]. In semiconductor fabrica-

tion processes, a recent survey conducted on current maintenance practices reveals that simple heuristic rules for preventive maintenance are widely employed [83]. As a result, equipment has experienced unnecessary cleanings, which usually generate substantial costs to ensure the product quality.

In order to solve both equipment maintenance and product sequencing problems, researchers have begun to explore the interaction between machine condition and product quality. Yano et al. [84] presented a comprehensive review of production models with variable yield. However, most of their work focused on single product system and did not treat product quality as a function of process condition. In other words, the product sequencing problem has not been considered in the models although the level of degradation may have a larger impact on some products than on others. Various extensions of similar work have been performed: allowing process inspections during a production cycle [85], studying different cost structures [86], considering the effects of machine failures [87], and allowing process improvements to be made [88]. All of these models did examine how much to produce. However, they did not address the question of which product to process next.

Sloan et al. [89] and Zhou [90] set a milestone for a multi-stage, multi-product system by developing models that employ combined decision-making on maintenance and product sequencing. Although both models used an explicit link between equipment condition and product quality for maintenance and product sequencing decisions, only the steady state (long run) condition is studied for the decision-making. The short-term effects such as work-in-process (WIP) levels and the maintenance resource limitation were not considered. Therefore, these models are not able to precisely respond to dynamic changes and process variations in production lines.

To incorporate short-term decisions into long-term decisions, we propose an in-

tegrated job sequencing and maintenance scheduling policy for a multi-stage, multi-product system often found in manufacturing processes. Tool degradation, equipment condition monitoring and product sequencing are simultaneously considered for the long-term decision. Transient process variations can be mitigated by dynamically rerouting material to the stations with less variation or to non-bottleneck stations for the short-term decision. A Markov decision process for the long-term and integer programming for the short-term are used.

The remainder of this chapter is organized in the following manner. Section 4.2 explicates assumptions and system characteristics, followed by the methodologies for joint decision making developed in Section 4.3. Then, numerical examples are provided in Section 4.4 to benchmark our integrated policy with reference policies. Section 4.5 also demonstrates the effectiveness of the proposed methodology applied to semiconductor manufacturing processes. Finally, Section 4.6 concludes this chapter with some remarks and suggestions on future work.

4.2 Model of A Manufacturing System

We consider a multi-product, multi-station system. The system that manufactures K products consists of H stations connected in series with intermediate buffers, as displayed in Figure 4.1.

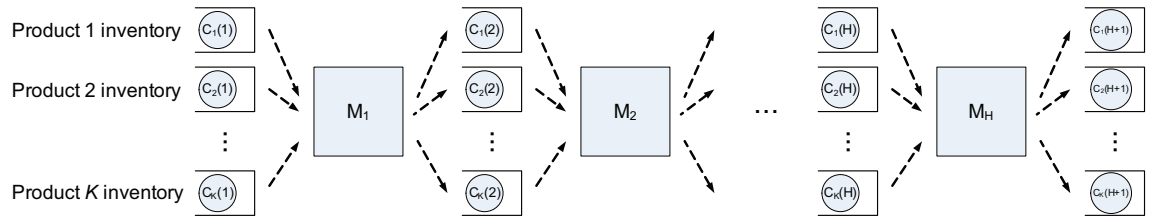


Figure 4.1: A multi-product, multi-station system in series

4.2.1 The Problem Statement

In order to make machine condition back to acceptable status and produce high yield, maintenance should be considered cautiously. Based on the observed machine conditions, we must decide whether to clean or repair the machine before the next inspection. Then, given repair tasks generated from all machines, a detailed maintenance schedule needs to be determined according to repair resource availability and WIP dynamics. At the same time, we must determine a product type for the subsequent production cycle, based on how sensitive certain products are as equipment deteriorates.

4.2.2 Model Assumptions

Suppose that we need to manufacture K types of products (p_1, \dots, p_K) . The processing times for these products on each station are assumed equal although products have different prices. We have to meet some pre-specified production ratios in the long-term production demand. Let ω_k be the long run proportion of product p_k required, where $k = 1, \dots, K$, and $\sum_k \omega_k = 1$. Setup times are assumed negligible or included in the processing times. Otherwise, we assume that we are able to reduce changeover times using many techniques [91, 92]. We assume that inspections are instantaneous and occur only at discrete times and these inspections can perfectly reveal the condition of machine. Assumptions related to the system are summarized as follows:

- The system consists of H stations connected in series with intermediate buffers.
- There are K types of products (p_1, \dots, p_K) .
- A long-term production demand must be met (ω_k for $k = 1, \dots, K$).

- The processing times for these products on each station are assumed equal.
- Setup times between different product types are included in the processing times.
- The degradation process can be modeled by a discrete-time Markov chain.
- Inspections are instantaneous and occur only at discrete times.
- The inspection can perfectly reveal the condition of a machine.
- Buffer capacities $C(i)$ are finite.
- Available maintenance resources are limited.

4.2.3 Degradation Model

The degradation process of each machine can be modeled by a discrete-time Markov chain (DTMC) $\{X(n), n \geq 0\}$ with a discrete state space $\mathbf{S} = \{S_1, \dots, S_M\}$. A Markov chain is a stochastic process with a Markovian property, namely, that the future and past states are independent given the present state [42]. Figure 4.2 provides an illustration of the Markov unidirectional equipment degradation process. For instance, p_{12} means the probability that state S_1 will transition to state S_2 . In our case, the transition probabilities p_{ij} of a DTMC are zero whenever $i > j$, since the machine condition is assumed to become only worse with time unless maintenance is performed. Furthermore, appropriate Markov models for deterioration processes can be estimated from a set of measurement data through a Hidden Markov Model (HMM), which is a method that enables us to stochastically relate available measurements to the machine conditions [93, 76].

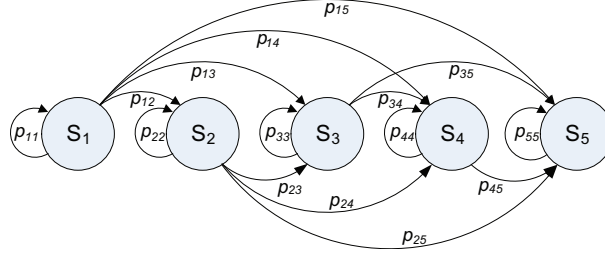


Figure 4.2: Illustration of a state transition diagram for a Markov chain

4.3 Joint Job Sequencing and Maintenance Scheduling For a Multiple Product, Multiple Station System

To deal with highly dynamic and complex problems in a manufacturing area, we propose a new decision making structure, given that tool conditions are available through an HMM. First, we aim to make a long-term plan on product sequencing and equipment cleaning on the system level. In this phase of decision-making, we focus on the stochastic nature of the problem, while intentionally ignoring the interdependencies between stations, which are left to the short-term decision-making. Given repair requests generated from all machines, we will modify our decisions if necessary. We prioritize these repair tasks and determine detailed maintenance schedules, considering the complex interdependencies among repairs on all stations with respect to maintenance availability and WIP inventory costs.

4.3.1 Long-Term Planning

In this section, we derive the optimal policy for job sequencing, preventive cleaning, and repair planning. This planning model is formulated as a Markov Decision Process (MDP), which seeks to keep a balance between preventive cleaning cost and yield loss in the long-term. We consider the problem of scheduling production and maintenance for a single machine, multiple product system, as shown in Figure 4.3. The analytical approach for simple cases provides some insight for more complicated

systems.

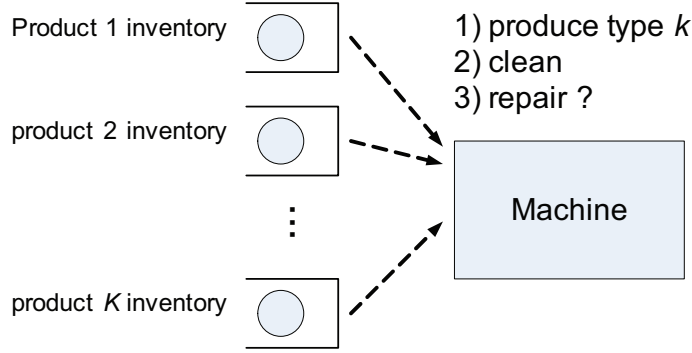


Figure 4.3: Single machine, multiple product system

Most operations in manufacturing heavily depend not only on the condition of a machine but also on its cleanliness, especially in semiconductor manufacturing process. For example, particles accumulated in the chambers increase the risk of yield loss. Therefore, it is important to determine when to perform cleaning between consecutive production cycles. Note that cleaning does not require any maintenance crew to get involved. Less frequent cleaning will result in higher yield loss, but too many cleanings will induce unnecessary maintenance costs and productivity losses. Thus, we intend to find the tradeoff between over- and under- cleaning. Moreover, we also need to find economic rules for job sequencing since it is beneficial to produce more sensitive products right after cleaning operation and to produce less sensitive products as a tool deteriorates.

Let $a(n)$ denote the action taken in time n . Possible actions include:

$$a(n) = \begin{cases} k & \text{if producing the product } p_k \text{ } (k = 1, \dots, K) \\ K + 1 & \text{if cleaning the machine} \\ K + 2 & \text{if repairing the machine} \end{cases} \quad (4.1)$$

We assume that changes in the machine state depend only on the current state and the action taken. Then this process can be expressed as the following transition

probability:

$$\begin{aligned}
P_a(i, j) &= P\{X(n+1) = S_j | X(n) = S_i, X(n-1), \dots, a(n) = a, a(n-1), \dots\} \\
&= P\{X(n+1) = S_j | X(n) = S_i, a(n) = a\}
\end{aligned} \tag{4.2}$$

The condition, or state, of machine deteriorates over time, and ultimately affects the probability of successfully producing the various products in a negative manner. We, therefore, refer to $Y(i, k)$ as the yield of product p_k when the machine is in state S_i . We assume that the yield is non-increasing as machine condition deteriorates, and the yield for all products is zero when the machine is in the worst state S_M . Cleaning can be initiated from any state and will return the equipment to a better condition. On the other hand, repairing can be initiated only from state S_M and return equipment condition to state S_1 with probability of 1. Once in state S_M , the machine cannot escape this state unless it is repaired. The machine degradation process is affected by the choice of which product to manufacture. An immediate reward $R(i, a)$, a function of product prices and maintenance costs, is received when action a is taken in state S_i . These assumptions about the long-term planning are summarized as follows:

$$Y(i, a) \geq Y(j, a) \quad \text{if } i < j \text{ for all } a \tag{4.3}$$

$$Y(M, a) = 0 \quad \text{for } a = 1, \dots, K \tag{4.4}$$

$$P_a(i, j) = \begin{cases} P_k(i, j) & \text{for } a = 1, \dots, K \\ P_{K+1}(i, j) & \text{for } a = K + 1 \\ P_{K+2}(i, j) & \text{for } a = K + 2 \end{cases} \tag{4.5}$$

A policy $\Pi(i)$ is a decision rule that prescribes an action for each state S_i , and our objective is to determine a policy that maximizes long-run expected average profit

$V(i)$. The optimal value of $V(i)$ will satisfy the following Bellman equations [94]:

$$\Pi(i) = \arg \max_a \left\{ R(i, a) + \sum_j P_a(i, j) V(j) \right\} \quad (4.6)$$

$$V(i) = R(i, \Pi(i)) + \sum_j P_{\Pi(i)}(i, j) V(j) \quad (4.7)$$

To solve the Bellman Equations (4.6) and (4.7), we consider randomized policies, in which actions are chosen according to some probability distributions. Let $x(i, a)$ denote the probability that the machine is in state S_i , and action a is taken. Then, the optimal policy can be found by solving the following linear programming [95].

$$\text{objective} \quad \max \sum_i \sum_a R(i, a) Y(i, a) x(i, a) \quad (4.8)$$

$$\text{subject to} \quad \sum_a x(j, a) - \sum_i \sum_a P_a(i, j) x(i, a) = 0 \quad \text{for all } j \quad (4.9)$$

$$\sum_i \sum_a x(i, a) = 1 \quad (4.10)$$

$$x(i, a) \geq 0 \quad \text{for all } i, a \quad (4.11)$$

$$\sum_i Y(i, k) x(i, k) - \omega_k \sum_k \sum_i Y(i, k) x(i, k) = 0 \quad \text{for } k = 1, \dots, K \quad (4.12)$$

Equation (4.9) shows that the state balance equations for the Markov chain that governs the machine state transitions. Equations (4.10) and (4.11) guarantee that all probabilities sum to 1, and are non-negative, respectively. Since we can interpret $x(i, a)$ as the long run proportion of time that the process is in state S_i and action a is taken, Equation (4.12) can serve another constraint to ensure that the long run average production requirement are met. Then, $x(i, a)$ gives the steady state probabilities for the policy $\Pi(i)$ that chooses action a in state S_i with probability,

$$\frac{x(i, a)}{\sum_a x(i, a)} \text{ (if } \sum_a x(i, a) > 0\text{)}.$$

4.3.2 Short-Term Scheduling

Since multiple stations may demand the same resource for their operations, we need to direct repair task to work on the more critical stations in case of a conflict [96, 97]. Although we derive the long-term policy in the previous decision phase, dynamic process variations have to be managed based on the real time information. This phase deals with determining which maintenance requests have higher priorities than other requests with constraints of maintenance resource availability and the dynamic changes in WIP cost induced by production and maintenance.

There are two factors to consider for the repair prioritization: 1) the urgency and 2) the WIP inventory cost. $T(i, n)$, the urgency of the broken machine M_i , is first given by Equation (4.13).

$$T(i, n) = \min(T_F(i, n), T_E(i + 1, n)) \quad (4.13)$$

where $T_F(i, n) = n + \{C(i) - W(i, n)\} t_p$ is the time when buffer i is full due to the failure of M_i at time n and $T_E(i + 1, n) = n + W(i + 1, n) t_p$ is the time when buffer $i + 1$ is empty due to the failure of M_i at time n .

Physically, $(T(i, n) - n)$ represents the minimum amount of time that the system can run without experiencing either blockage or starvation due to the broken machine M_i . For instance, if machine M_i is down, blockage will occur at buffer i after $T_F(i, n)$. Conversely, starvation will occur at buffer $i + 1$ after $T_E(i + 1, n)$. Hence, $(T(i, n) - n)$ is the minimum time that allows the production line to run without overall utilization loss. Therefore, we need to select the machine with a smaller value of $T(i, n)$ to

prevent the unnecessary propagation of machine idle (i.e., starvation or blockage).

Second, different repair sequences may introduce less WIP inventory cost to the production system. We define two relevant WIP inventory costs (total WIP inventory costs and total slack WIP inventory cost at time n):

$$\mathbf{Cost}_{\mathbf{W}}(i, n) \triangleq \sum_{k=1}^H W_i(k, n)\mu(k) \quad (4.14)$$

$$\mathbf{Cost}_{\mathbf{S}}(i, n) \triangleq \sum_{k=1}^H S_i(k, n)\mu(k) = \sum_{k=1}^H [C(k) - W_i(k, n)]\mu(k) \quad (4.15)$$

where $W_i(k, n)$ is WIPs in buffer k at time n after machine M_i fails, $S(k, n)$ is empty space (= slack) in buffer k , $C(k)$ is capacity of buffer k , and $\mu(k)$ is an unit WIP inventory cost of buffer k .

$W(k, n)\mu(k)$ is the WIP inventory cost of buffer k at time n while $S(k, n)\mu(k)$ is the slack WIP inventory cost of buffer k at time n . We introduce the slack WIP $S(k, n)$ in order to define the repair rank index later. Furthermore,

$$\mathbf{Cost}_{\mathbf{W}}(i, n) \geq \mathbf{Cost}_{\mathbf{W}}(j, n) \iff \mathbf{Cost}_{\mathbf{S}}(i, n) \leq \mathbf{Cost}_{\mathbf{S}}(j, n) \quad \text{for } i \neq j \quad (4.16)$$

Proof. Since $C(i)$ does not change with time,

$$\begin{aligned} & \mathbf{Cost}_{\mathbf{W}}(i, n) \geq \mathbf{Cost}_{\mathbf{W}}(j, n) \\ \iff & \sum_{k=1}^H W_i(k, n)\mu(k) \geq \sum_{k=1}^H W_j(k, n)\mu(k) \\ \iff & \sum_{k=1}^H C(k)\mu(k) - \sum_{k=1}^H W_i(k, n)\mu(k) \leq \sum_{k=1}^H C(k)\mu(k) - \sum_{k=1}^H W_j(k, n)\mu(k) \\ \iff & \sum_{k=1}^H [C(k) - W_i(k, n)]\mu(k) \leq \sum_{k=1}^H [C(k) - W_j(k, n)]\mu(k) \\ \iff & \sum_{k=1}^H S_i(k, n)\mu(k) \leq \sum_{k=1}^H S_j(k, n)\mu(k) \\ \iff & \mathbf{Cost}_{\mathbf{S}}(i, n) \leq \mathbf{Cost}_{\mathbf{S}}(j, n) \quad \square \end{aligned}$$

Equation (4.16) shows that if and only if machine M_i induces higher WIP inventory cost than machine M_j , then the slack WIP inventory cost induced by machine M_i is smaller than that of machine M_j . Therefore, assigning higher repairing priorities on the machine that has higher WIP inventory cost is equivalent to selecting the machine that has smaller slack WIP inventory cost with respect to reduction of the WIP inventory cost.

Then, we define the repair rank index of machine M_i at time n as:

$$U(i, n) \triangleq \sum_{m=n}^{T(i,n)} \sum_{k=1}^H S_i(k, m) \mu(k) \quad (4.17)$$

Note that $U(i, n)$ provides a measure of both the urgency and WIP inventory cost for machines since it adds the total slack WIP costs until $T(i, n)$. Our objective is, then, to select the repair request with the lowest rank index. We can formulate this problem as binary integer programming:

$$\text{objective} \quad \max \sum_i \frac{1}{U(i, n)} r(i, n) d(i, n) \quad (4.18)$$

$$\text{subject to} \quad \sum_i d(i, n) \leq N_r, \quad \forall n \quad (4.19)$$

$$\sum_i d(i, n) \leq \sum_i r(i, n), \quad \forall n \quad (4.20)$$

$$\sum_i W(i, n) \mu(i) \leq C_{WIP}^*, \quad \forall n \quad (4.21)$$

$$d(i, n) \in \{0, 1\}, \quad \forall i, n \quad (4.22)$$

where

- $d(i, n)$: the binary decision variables in the integer programming

$$d(i, n) = \begin{cases} 1 & \text{if conducting repair job on machine } M_i \text{ at time } n \\ 0 & \text{otherwise} \end{cases} \quad (4.23)$$

- $r(i, n)$: the repair tasks/requests on station M_i

$$r(i, n) = \begin{cases} 1 & \text{if repair at time } n \text{ is requested from station } M_i \\ 0 & \text{otherwise} \end{cases} \quad (4.24)$$

- N_r : the total number of maintenance personnel

Hence, Equation (4.18) assigns higher repair priorities to those machines that have lower rank index $U(i, n)$. Instead of minimizing $U(i, n)$, we maximize $\frac{1}{U(i, n)}$ in Equation (4.18) because of the characteristics of binary variables, $r(i, n)$ and $d(i, n)$. The constraint in Equation (4.19) ensures that no more than the available amount of resource for maintenance is committed. Equation (4.20) represents that the number of repairs cannot exceed its requests. The value of $W(i, n)$ is determined according to the job sequencing rule and maintenance decision obtained in the long-term decision. Equation (4.21) ensures that the WIP inventory cost $\sum_i W(i, n)\mu(i)$ does not exceed its limit, C_{WIP}^* .

When the machines require repairs, rank indexes $U(i, n)$ are computed for the broken machines considering both the urgency and WIP inventory cost. Then, the repair scheduling is determined in descending order of the rank index through the above integer programming. The short-term decision works toward avoiding either: 1) the unnecessary propagation of machine idle (i.e., starvation or blockage) caused by fixing less urgent machine, or 2) the accumulation of excess inventory by incorrectly selected repairs.

4.4 Numerical Case Studies

In this section, numerical experiments are carried out to demonstrate how much improvement the proposed method can achieve, compared with other policies.

4.4.1 Design

To compare the proposed policy with other policies, we present a number of numerical examples with the system that consists of 10 stations in series and buffers between stations (see Figure 4.1).

4.4.1.1 Product demand ratio

Product demand ratio ω represents how much of each product has to be manufactured as a fraction of total production. We assume that there are four different product types (p_1, p_2, p_3, p_4) . Different levels of their product demand ratios are displayed in Table 4.1. For example, the scenario Ω_1 only requires to produce a type p_1 while the scenario Ω_{12} requires to make equal amount of products among different types.

Table 4.1: Product demand ratios

Scenario	p_1	p_2	p_3	p_4
Ω_1	1.00	0.00	0.00	0.00
Ω_2	0.00	1.00	0.00	0.00
Ω_3	0.00	0.00	1.00	0.00
Ω_4	0.00	0.00	0.00	1.00
Ω_5	0.70	0.10	0.10	0.10
Ω_6	0.10	0.70	0.10	0.10
Ω_7	0.10	0.10	0.70	0.10
Ω_8	0.00	0.10	0.10	0.70
Ω_9	0.40	0.40	0.10	0.10
Ω_{10}	0.10	0.40	0.40	0.10
Ω_{11}	0.10	0.10	0.40	0.40
Ω_{12}	0.25	0.25	0.25	0.25

4.4.1.2 Deterioration processes

Machine degradation is represented by transition probability matrices (P_1, P_2, P_3, P_4) of Markov chains. As mentioned earlier, different products require different operating conditions, resulting in different equipment deterioration processes. We assume that product p_i requires a less severe operating environment than that of product p_j when $i < j$. Cleaning action (P_5) can make equipment less degraded but cannot repair a broken machine. Only repairing action (P_6) can fix broken equipment and return its condition to state S_1 with a probability of 1. All transition probability matrices of the corresponding Markov chains are listed in Table 4.2.

Table 4.2: Transition probability matrices for equipment deterioration processes

$P_1 = \begin{bmatrix} 0.9 & 0.09 & 0.005 & 0.003 & 0.002 \\ 0 & 0.90 & 0.090 & 0.005 & 0.005 \\ 0 & 0 & 0.900 & 0.090 & 0.010 \\ 0 & 0 & 0 & 0.900 & 0.100 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$	$P_2 = \begin{bmatrix} 0.8 & 0.1 & 0.05 & 0.03 & 0.02 \\ 0 & 0.8 & 0.10 & 0.05 & 0.05 \\ 0 & 0 & 0.80 & 0.10 & 0.10 \\ 0 & 0 & 0 & 0.80 & 0.20 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$
$P_3 = \begin{bmatrix} 0.7 & 0.2 & 0.05 & 0.03 & 0.02 \\ 0 & 0.7 & 0.20 & 0.05 & 0.05 \\ 0 & 0 & 0.70 & 0.20 & 0.10 \\ 0 & 0 & 0 & 0.70 & 0.30 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$	$P_4 = \begin{bmatrix} 0.6 & 0.3 & 0.05 & 0.03 & 0.02 \\ 0 & 0.6 & 0.30 & 0.05 & 0.05 \\ 0 & 0 & 0.60 & 0.30 & 0.10 \\ 0 & 0 & 0 & 0.60 & 0.40 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$
$P_5 = \begin{bmatrix} 1.0 & 0 & 0 & 0 & 0 \\ 0.9 & 0.1 & 0 & 0 & 0 \\ 0.8 & 0.1 & 0.1 & 0 & 0 \\ 0.7 & 0.1 & 0.1 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$	$P_6 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$

4.4.1.3 Product prices and maintenance costs

An immediate reward is received when action a is taken in state S_i . The reward values, which are assumed product prices, and cleaning/repairing cost are shown in Tables 4.3 and 4.4. We test five different levels of product prices (R_1, \dots, R_5), while holding the cleaning and repairing cost constant to investigate effects of different

product prices, as displayed in Table 4.3. On the other hand, we have another 8 combinations of cleaning and repair costs (R_6, \dots, R_{13}) in Table 4.4 to explore the effects of cleaning and repairing costs, while keeping product prices constant.

Table 4.3: Reward for products

Reward (\$)	p_1	p_2	p_3	p_4	cleaning	repairing
R_1	340	270	130	60	-200	-400
R_2	300	250	150	100	-200	-400
R_3	200	200	200	200	-200	-400
R_4	100	150	250	300	-200	-400
R_5	60	130	270	340	-200	-400

Table 4.4: Cleaning/repairing cost

Reward (\$)	p_1	p_2	p_3	p_4	cleaning	repairing
R_6	250	200	150	100	-25	-50
R_7	250	200	150	100	-50	-100
R_8	250	200	150	100	-75	-150
R_9	250	200	150	100	-100	-200
R_{10}	250	200	150	100	-125	-250
R_{11}	250	200	150	100	-150	-300
R_{12}	250	200	150	100	-175	-350
R_{13}	250	200	150	100	-200	-400

4.4.1.4 Yield

The yield is represented by the yield matrices. The element $Y(i, k)$ of the yield matrices defines the expected yield values when product p_k is produced in state S_i . Four different yield matrices, shown in Table 4.5, are obtained from [13]. Note that Y_i has a lower variance than Y_j when $i < j$.

4.4.2 Sequencing and Maintenance Policies

Our objective is to find job sequencing and maintenance rules that maximize the expected average profit per time unit while maintaining a certain level of product

Table 4.5: Yield matrices

$Y_1 = \begin{bmatrix} 0.9689 & 0.9598 & 0.9996 & 0.9536 \\ 0.9446 & 0.7211 & 0.9976 & 0.6809 \\ 0.7926 & 0.4685 & 0.9875 & 0.3935 \\ 0.4392 & 0.2927 & 0.0762 & 0.2620 \\ 0 & 0 & 0 & 0 \end{bmatrix}$	$Y_2 = \begin{bmatrix} 0.9974 & 0.9790 & 0.5313 & 0.8859 \\ 0.8124 & 0.7077 & 0.4818 & 0.8744 \\ 0.7750 & 0.3460 & 0.2375 & 0.3668 \\ 0.3819 & 0.0003 & 0.0219 & 0.2581 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
$Y_3 = \begin{bmatrix} 1.0000 & 0.9997 & 0.8655 & 0.9998 \\ 0.9030 & 0.9973 & 0.6498 & 0.9548 \\ 0.8854 & 0.7204 & 0.4969 & 0.1772 \\ 0.5990 & 0.1316 & 0.2564 & 0.0793 \\ 0 & 0 & 0 & 0 \end{bmatrix}$	$Y_4 = \begin{bmatrix} 0.9359 & 1.0000 & 0.9845 & 1.0000 \\ 0.4848 & 1.0000 & 0.5718 & 0.7049 \\ 0.2394 & 0.9883 & 0.3911 & 0.2700 \\ 0.0402 & 0.6110 & 0.1274 & 0.0628 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

proportions. Traditionally, job sequencing and maintenance scheduling are treated sequentially and locally [78]. Based on the approaches often used in industrial practice, we define three other policies for the purpose of a benchmark as shown in Table 4.6. Note that we denote Policy 4 as the proposed policy in Section 4.3.

Table 4.6: Benchmark policies

Polices	Job sequencing	Cleaning	Repairing
Policy 1	FCFS	Preventive	FCFS
Policy 2	FCFS	Condition-based	FCFS
Policy 3	Long-term planning		FCFS
Policy 4	Long-term planning	Short-term scheduling	

- FCFS product sequencing: the product is simply dispatched in a first come first serve (FCFS) basis.
- FCFS repairing: if repair tasks are requested more than maintenance resource available, this policy will select the repair tasks that requested first.
- Preventive cleaning: we clean machine after N products have been produced, regardless of product type.
- Condition-based cleaning: cleaning is performed whenever machine condition reaches a pre-defined state, regardless of product type.

4.4.3 Results

To quantify the improvement of the proposed policy on the system performance we have compared the simulation results by varying one of four parameters (product demand ratio, deterioration processes, product prices, and cleaning/repairing costs). Each simulation result is explained in the following sections. Policy 4 performs better than Policies 1, 2, and 3 with respect to the expected average reward throughout the simulation results. Systems are over-maintained (i.e., too frequent preventive cleaning) with Policy 1 (the fixed preventive cleaning). Policy 2 does not consider that different products might require different conditions to trigger cleaning although it has considered the condition of machine. Policy 3 undergoes the FCFS basis while Policy 4 considers the real time information on dynamic manufacturing variations for the short-term decision.

In addition to the policy benchmarks, the simulation results about buffer capacities are demonstrated to find the tradeoff between production smoothness and WIP holding costs. The impacts of maintenance staffing (N_r) are also simulated to find an appropriate level of maintenance resources [98].

4.4.3.1 The effect of product demand ratio

To examine the effect of different product demand ratios, we assume that degradation processes are independent of product types and Y_1 is arbitrarily selected from Table 4.5 as the yield model. The average reward values for each policy have been achieved through 50 replications of simulation runs and reported in Figure 4.4 and Table 4.7.

Policy 4 provides more improvement over Policy 1 and Policy 2 when its product demand ratios among products are high. This phenomenon can be observed in Fig-

ure 4.4, having increased improvement with the scenarios of Ω_9 , Ω_{10} , Ω_{11} , and Ω_{12} . The reason of increased improvement lies on the higher product mix rates that can provide more opportunities for job sequencing to maximize the profits. However, both Policy 1 and 2 do not take different product prices and yield loss into account when it comes to job sequencing. Policy 2 performs better than Policy 1 because Policy 2 makes use of condition monitoring information for decision-making. Condition-based maintenance (CBM) can be a more cost-efficient maintenance policy over preventive maintenance (PM) if the system is able to provide information of machine condition. On the other hand, Policy 3 is as good as Policy 4 regardless of product demand ratio, because the short-term decision in Policy 4 is not dependent of product mix rates but of the degradation.

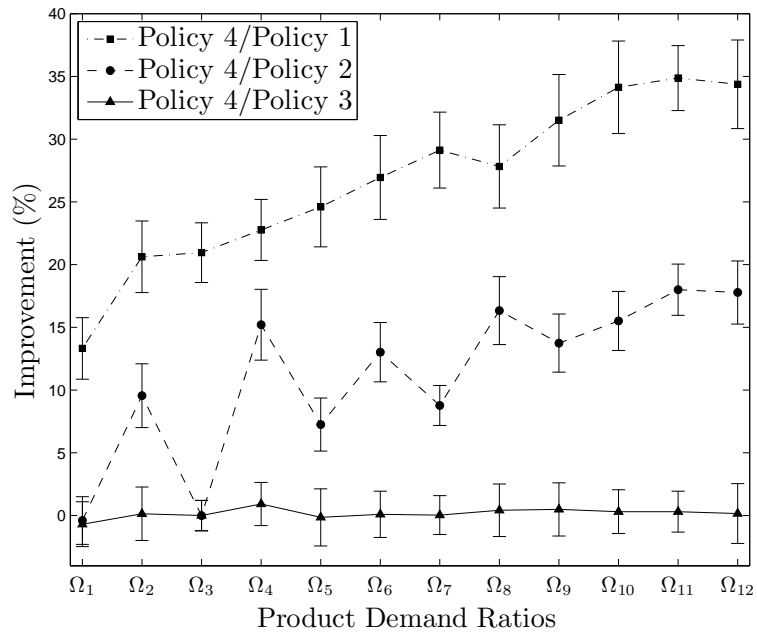


Figure 4.4: Improvement as a function of product demand ratio

Table 4.7: Simulation results with different product demand ratios

Scenario	Policy 1	Policy 2	Policy 3	Policy 4	P4/P1 (%)	P4/P2 (%)	P4/P3 (%)
Ω_1	59.5±1.2	67.7±1.1	67.9±0.9	67.5±1.2	13.3	-0.4	-0.7
Ω_2	80.4±1.8	88.5±2.0	96.9±1.8	97.0±1.8	20.6	9.6	0.1
Ω_3	135.9±3.0	164.4±1.8	164.4±1.8	164.4±1.7	21.0	0.0	0.0
Ω_4	140.3±2.5	149.5±3.8	170.6±2.1	172.2±3.0	22.8	15.2	0.9
Ω_5	75.5±1.9	87.7±1.2	94.2±1.9	94.1±1.9	24.6	7.3	-0.1
Ω_6	87.9±2.5	98.7±1.9	111.4±1.7	111.5±1.9	26.9	13	0.1
Ω_7	118.5±3.2	140.7±1.9	153.0±2.3	153.0±1.9	29.1	8.8	0.0
Ω_8	124.1±3.2	136.4±3.0	158.0±2.8	158.7±3.0	27.8	16.3	0.4
Ω_9	81.0±2.5	93.6±1.8	105.9±2.2	106.5±1.8	31.5	13.7	0.5
Ω_{10}	101.2±3.0	117.5±2.3	135.3±1.9	135.8±2.3	34.1	15.5	0.3
Ω_{11}	119.9±2.5	137.1±2.5	161.2±2.7	161.7±1.8	34.9	18	0.3
Ω_{12}	100.8±2.6	115.0±2.0	135.2±2.9	135.4±2.8	34.4	17.8	0.2

4.4.3.2 The effect of degradation

We investigate how much Policy 4 can improve the system performance when the degradation rates of a system are different. We assume that the yield model, reward, and production requirement follow Y_1 , R_{13} , and Ω_{12} , respectively, while we change degradation processes from P_1 (slower) to P_4 (faster) as reported in Table 4.8. Figure 4.5 illustrates that Policy 4 turns out to be more effective than Policies 1, 2, and 3 as systems produce under cruder manufacturing environment. Since cleaning and repairing occur more frequently with the increased deterioration rate, job sequencing and maintenance decisions under Policy 4 significantly contribute the enhancement of system performance.

Table 4.8: Transition probability matrix for degradation

Scenario	Transition Probability Matrix	Degradation Processes
D_1	P_1	degrades slower
D_2	P_2	degrades slow
D_3	P_3	degrades fast
D_4	P_4	degrades faster

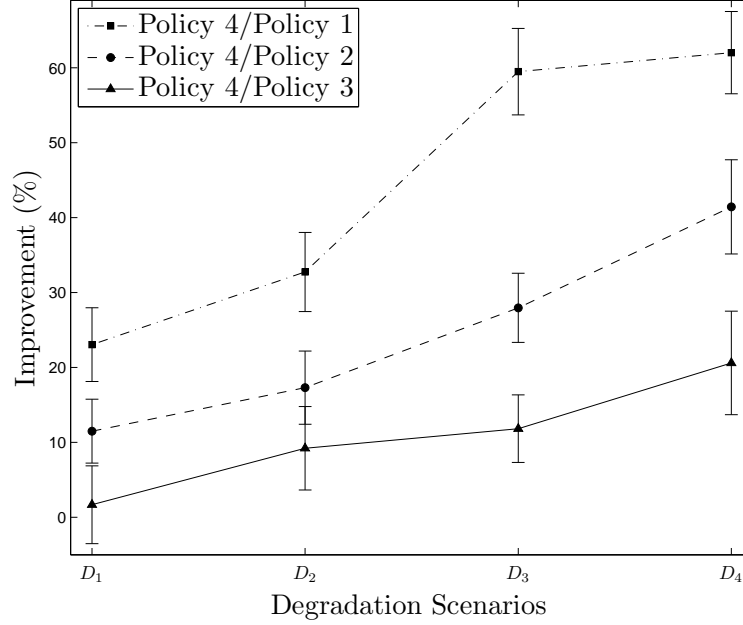


Figure 4.5: Improvement as a function of degradation

Table 4.9: Simulation results with different degradation processes

Scenario	Policy 1	Policy 2	Policy 3	Policy 4	P4/P1 (%)	P4/P2 (%)	P4/P3 (%)
D_1	89.5±2.0	98.8±1.1	108.3±6.2	110.1±6.9	23.1	11.5	1.7
D_2	62.8±2.6	71.1±3.3	76.4±5.2	83.4±4.3	32.7	17.3	9.2
D_3	49.7±2.2	61.9±2.7	70.8±3.7	79.2±3.3	59.5	28.0	11.8
D_4	40.5±1.3	46.4±2.6	54.4±4.5	65.6±3.1	62.0	41.4	20.6

4.4.3.3 The effect of product prices

To explore the effect of different product prices, we have five different product prices from R_1 to R_5 . Note that these five different price combinations seen from Table 4.3 are chosen in such a way that the mean values of product prices are the same so that we can avoid impacts of maintenance cost. However, no clear trend can be found throughout the different scenarios, as shown in Figure 4.6 and Table 4.10. Instead, improvement seems constant about 42 % (Policy 4 over Policy 1), 19 % (Policy 4 over Policy 2), and 0 % (Policy 4 over Policy 3) with the system parameters of Y_1 , P_3 , and Ω_{12} . The results highlight that variations of rewards have no influence on system performances.

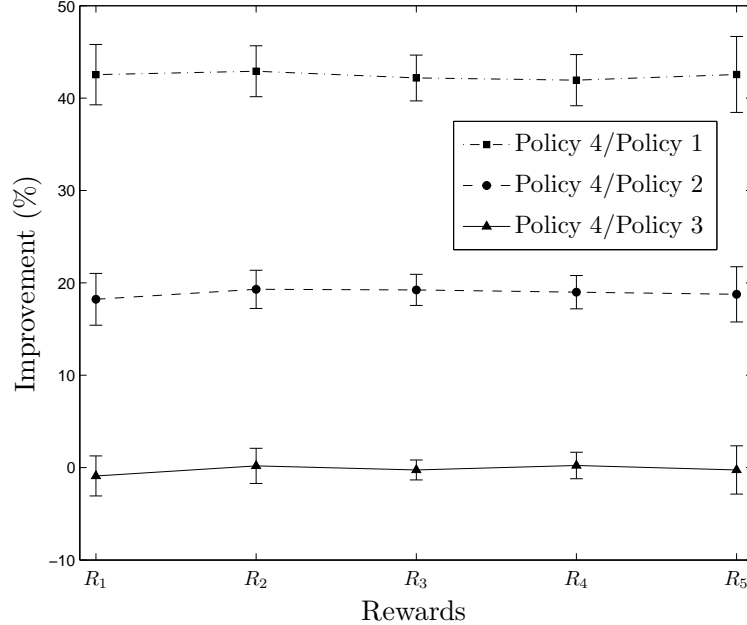


Figure 4.6: Improvement as a function of rewards

Table 4.10: Simulation results with different rewards

Scenario	Policy 1	Policy 2	Policy 3	Policy 4	P4/P1 (%)	P4/P2 (%)	P4/P3 (%)
R_1	98.4±1.2	118.6±1.7	141.5±1.5	140.2±2.7	42.5	18.2	-0.9
R_2	99.0±1.3	118.6±1.2	141.3±1.8	141.5±2.0	42.9	19.3	0.2
R_3	99.6±1.6	118.8±1.4	142.0±1.2	141.7±1.0	42.2	19.2	-0.3
R_4	100.0±1.6	119.3±1.3	141.7±1.3	142.0±1.5	41.9	19.0	0.2
R_5	99.5±2.1	119.5±1.9	142.2±2.5	141.9±2.8	42.6	18.8	-0.2

4.4.3.4 The effect of maintenance cost

To see the effect of Policy 4 on maintenance cost, we set the same production reward while changing maintenance costs (including cleaning and repair costs), as shown in Table 4.4. Y_1 and Ω_{12} are arbitrarily chosen for a yield model and production demand requirement, respectively. As maintenance costs become higher, Policy 4 provides increased improvement over Policies 1, 2, and 3. For example, the improvement of Policy 4 over Policy 1 ranges from 23.0 % to 68.4 % as the maintenance costs increase. The trend from Figure 4.7 demonstrates that the proposed Policy 4 is more effective when the maintenance costs are critical for decision-making process.

Table 4.11: Simulation results with different maintenance costs

Scenario	Policy 1	Policy 2	Policy 3	Policy 4	P4/P1 (%)	P4/P2 (%)	P4/P3 (%)
R_6	110.5±1.9	120.0±1.4	131.7±3.5	135.9±3.7	23.0	13.3	3.2
R_7	102.7±2.1	112.3±2.7	127.0±3.2	134.0±2.7	30.5	19.4	5.6
R_8	95.0±3.9	105.8±2.0	123.1±3.1	127.6±4.1	34.4	20.7	3.7
R_9	89.7±2.5	105.0±3.4	119.0±3.8	127.7±2.1	42.3	21.6	7.3
R_{10}	82.8±3.9	97.0±2.7	111.0±3.5	119.9±4.6	44.8	23.6	8.1
R_{11}	79.1±2.6	91.2±0.8	106.9±3.5	115.2±5.6	45.6	26.3	7.8
R_{12}	72.5±4.3	85.0±4.5	99.5±3.2	110.9±1.4	53.0	30.4	11.4
R_{13}	65.2±4.1	84.0±2.4	94.3±3.8	109.8±4.9	68.4	30.8	16.4

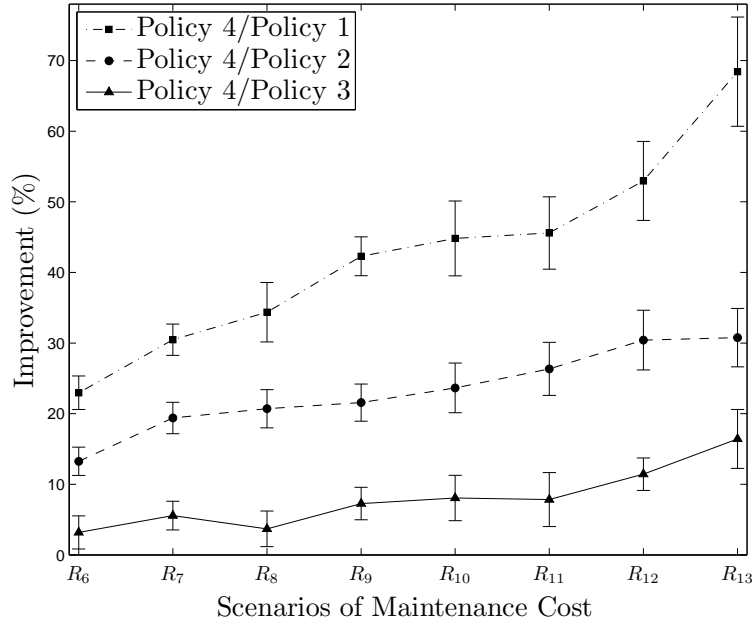


Figure 4.7: Improvement as a function of maintenance costs

4.4.3.5 The effect of buffer capacity

We examine the appropriate buffer level with a number of different buffer capacities, ranging from 2 to 30. As we expected, Figure 4.8 highlights that stations are often idle if the capacity of intermediate buffers is small. Empty buffers cannot provide material for downstream stations, while full buffers in downstream block a material flow. On the contrary, average WIP holding costs increase as the size of buffers becomes larger. Therefore, in order to keep the balance between the utilization of machines and inventory costs, it is important to consider the tradeoff as shown in Figure 4.8.

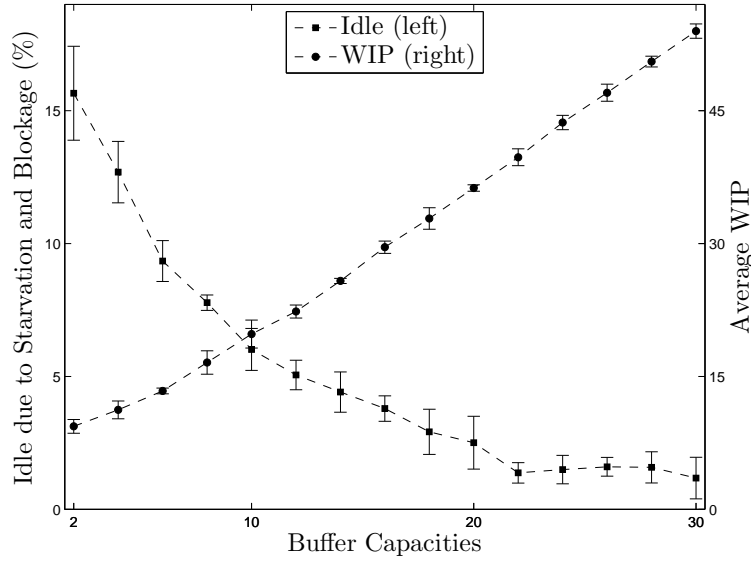


Figure 4.8: Impact of initial buffer contents

4.4.3.6 The effect of maintenance staffing

We change the number of maintenance personnel N_r in Equation (4.19) to find an appropriate maintenance staffing level. We measure percentage of idle times, caused by machine breakages, to evaluate the maintenance performance. Two degradation processes, such as P_1 and P_4 from Table 4.2, are tested with the system of 30 stations. Figure 4.9 reveals that an appropriate maintenance staffing level depends on degradation processes of the given system. For a slowly degrading system (P_1), we do not need more than 1 maintenance person to provide significant improvement because the chance that more than 1 machine fail simultaneously at the given system is unlikely. However, at least 2 maintenance personnel are necessary for a quickly degrading system (P_4). As we expected, the manufacturing system requires more maintenance resources for the production smoothness if machine in its system are not reliable.

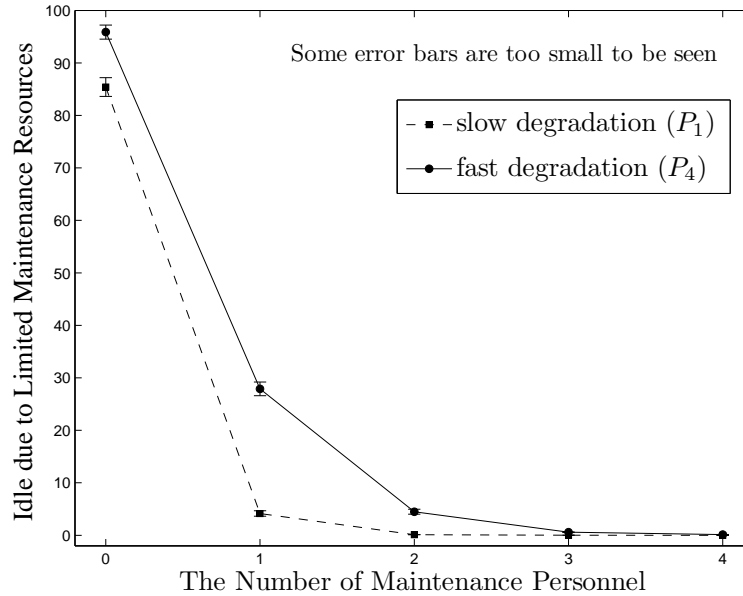


Figure 4.9: Impact of limited maintenance resource

4.5 Case Study with Semiconductor Manufacturing Process Data

In this section, we illustrate and validate our proposed Policy 4 with a set of industrial data from the semiconductor manufacturing processing. Data have been collected from chamber tools with two different recipes (i.e., two different types of products). Nine process parameters and the succeeding metrology measurement are periodically monitored in conjunction with associated process events.

The degradation processes have been modeled using a 5-state discrete time Markov chain. The corresponding probability transition matrix P , displayed in Table 4.12 is obtained from a set of manufacturing process data using an HMM. Since the underlying chamber degradation condition is not directly monitored or measured, we have to estimate them by applying the HMM, addressed in [76]. An HMM enables us to estimate machine condition from a sequence of measurements (on-wafer particle counts, temperature, pressure, etc.). The procedure of finding probability transition matrix

from a set of industrial data via an HMM is provided in Liu [76] in detail. Then, we derive the reliabilities of this chamber tool from the Markov chain by calculating the steady state probabilities. The resulting reliability distributions show that recipe 1 requires less harsh condition than that of recipe 2 although the current maintenance policy does not consider this difference for the cleaning decision.

Table 4.12: 5-state Markov chain with the corresponding transition probability matrix P

P_1					P_2				
0.972	0.020	0.008	0	0	0.315	0.685	0	0	0
0	0.926	0.054	0.020	0	0	0.968	0.025	0.005	0.002
0	0	0.987	0.012	0.001	0	0	0.703	0.280	0.017
0	0	0	0.974	0.026	0	0	0	0.228	0.772
0	0	0	0	1	0	0	0	0	1

Table 4.13: Benchmark results

	Policy 1	Policy 4	Policy4/Policy1 (%)
Reward values	111.13±1.08	146.39±0.96	31.73

As we can see in Table 4.13, the proposed Policy 4 improves the expected average reward by 31.73 %. As we have seen from the previous sections, this significant improvement in system performance can be attributed to the two advantages of the Policy 4:

- 1) simultaneously considering job sequencing and cleaning,
- 2) carrying out system level repairing decision based on real time information

Manufactures are generally conservative for production and maintenance planning. We learned that Policy 1 has been currently employed as a maintenance and sequencing rule suggested by chamber tool manufacturers. In other words, they clean the chamber after producing two wafers regardless of the wafer type and dispatch

wafers on a FCFS basis without considering the relationship between yield loss and a degraded chamber. In contrast, Policy 4 that we are proposing recommends to dispatch a wafer p_2 to the less degraded state while dispatching a wafer p_1 to the more degraded state. When it comes to cleaning, condition-based cleaning is suggested instead of a wafer-based cleaning.

4.6 Conclusion

The purpose of this chapter was to develop an advanced job sequencing/maintenance policy based on both online condition monitoring information and the dynamic relationship between machine degradation and product quality. The problem was motivated by an application in semiconductor fabrications where machine deterioration has different influences on the yield of different types of products. We proposed an integrated decision-making on maintenance scheduling and production planning. In this proposed model, the long-term decision focuses on the stochastic degradation process of each station to make long-term planning on repair scheduling and job sequencing. On the other hand, the short-term decision focuses on the dynamic interdependencies between stations to make short-term system level maintenance schedules. A simulation model is built to demonstrate the advantages of the integrated policy over conventional policies. It is shown that using this proposed model, we are able to derive a policy with simple structure, which can be easily implemented in practice. In addition, this model is shown to significantly improve the system performance in terms of expected average profit and total repair costs through several numerical experiments and a case study with a real manufacturing process data.

In terms of the future work, the proposed integrated model could be expanded to a more complex application involving re-entrant loops, multi-layers, and larger sys-

tems having more machines with different configurations and producing more product types. Furthermore, inventory planning was out of the scope of our study because yield is more emphasized than cycle time for the make-to-stock system. Including this issue into joint decision making on maintenance and job sequencing would be also an interesting topic to be explored in the future.

CHAPTER V

Conclusions and Contributions

5.1 Conclusions

This research has investigated a variety of stochastic modeling techniques, focusing on a diagnosis and a maintenance decision-making in manufacturing systems. Since maintenance costs become a large portion of a company expense, more attention on maintenance has to be paid in order to reduce overall maintenance cost.

To detect and isolate unknown faults as early as possible, we develop the modified hidden Markov model algorithm. A conventional hidden Markov model has been combined with the Hotelling multivariate control chart technique in statistical process control for unknown state detection. This reinforcement learning algorithm enables us to detect anomalous behaviors at the early stages so that appropriate condition-based maintenance can be planned accordingly. The mHMM which is applied to a turning process illustrates that it is able to not only estimate a tool wear, but also detect shortage of coolant. It has also been shown that a mHMM provides higher estimation accuracy than other classification algorithms such as neural networks, Gaussian mixture model, and K-means.

Stochastic models of joint equipment degradation processes and maintenance actions have been proposed for a preventive maintenance decision-making. These mod-

els are used to find the optimal PM intervals in terms of the system availability, productivity, and profit. We account for non-negligible repair times and periodic inspections in Markov models so that we are able to represent more realistic maintenance characteristics. Furthermore, these modeling techniques are extended to maintenance decision-making problems with two machine systems. With the case study from a semiconductor manufacturing, we show that the current production system has been over-maintained and recommend the new optimal PM interval with a sensitivity analysis. We have shown from this case study that maintenance decision-making depends on degradation process, maintenance (PM and RM) costs, and machine configurations.

For the system which can manufacture multiple products, the relationship between machine condition and product quality has been considered to develop advanced joint maintenance and product sequencing policy. This joint policy can provide more opportunities to select an appropriate product among multiple products. The short-term decisions have been added in the joint policy in order to intelligently respond the system dynamics. This integrated policy is proven to significantly improve the system performance in terms of average profit and total repair costs through numerical experiments and case study with a real semiconductor manufacturing process data.

5.2 Contributions

Comprehensive research of maintenance strategies for complex manufacturing systems aims to provide more cost-effective maintenance programs using stochastic modeling techniques.

The scientific contributions of this research are summarized as follows:

- A modified hidden Markov model combined with reinforcement learning algo-

rithms has been proposed for online condition monitoring and unknown fault isolation.

- We propose to use an approximation of a non-exponential distribution in a Markov chain using phase-type distributions. The effects of different machine configurations on preventive maintenance policies have been analytically found.
- The links between machine condition and corresponding product quality have been used for integrated production and maintenance rules to increase the overall system profit.

5.3 Future Work

Future work will involve further experimental validations of the hidden Markov Model algorithm with many other applications. A current modified hidden Markov model is limited to the Gaussian density distribution for the monitoring signals. We have to release this assumption to handle general random distributions.

Additionally, it might be difficult to analytically find the optimal PM intervals for a system which has more than two machines as the complexity of the proposed method grows exponentially according to the number of machines. Therefore, an efficient algorithm to handle the computational problem has to be explored.

For the multiple product system, the proposed policy model can be expanded to a more complex application involving different system configurations. Furthermore, considering setup or changeover times between different types of products can make our model more realistic in most manufacturing processes.

BIBLIOGRAPHY

BIBLIOGRAPHY

- [1] Brick, J. M., Michael, J. R., and Morganstein, D., 1989, “Using statistical thinking to solve maintenance problems,” *Quality Progress*, **22**(5), pp. 55–60.
- [2] Valdez-Flores, C. and Feldman, R. M., 1989, “A survey of preventive maintenance models for stochastically deteriorating single-unit systems,” *Naval Research Logistics*, **36**(4), pp. 419–446.
- [3] Wang, H., 2002, “A survey of maintenance policies of deteriorating systems,” *European Journal of Operational Research*, **139**(3), pp. 469–489.
- [4] Iakovou, E., Ip, C. M., and Koulamas, C., 1999, “Throughput dependent periodic maintenance policies for general production units,” *Annals of Operations Research*, **91**(0), pp. 41–47.
- [5] Barlow, R. and Hunter, L., 1960, “Optimum preventive maintenance policies,” *Operations Research*, **8**(1), pp. 90–100.
- [6] Barlow, R. E. and Proschan, F., 1996, *Mathematical theory of reliability*, Society for Industrial Mathematics.
- [7] Handlarski, J., 1980, “Mathematical analysis of preventive maintenance schemes,” *The Journal of the Operational Research Society*, **31**(3), pp. 227–237.
- [8] McCall, J. J., 1965, “Maintenance policies for stochastically failing equipment: a survey,” *Management Science*, **11**(5), pp. 493–524.
- [9] Pierskalla, W. P. and Voelker, J. A., 1976, “A survey of maintenance models: the control and surveillance of deteriorating systems,” *Naval Research Logistics Quarterly*, **23**(3), pp. 353–388.
- [10] Scarf, P. A., 1997, “On the application of mathematical models in maintenance,” *European Journal of Operational Research*, **99**(3), pp. 493–506.
- [11] Zeng, S. W., 1997, “Discussion on maintenance strategy, policy and corresponding maintenance systems in manufacturing,” *Reliability engineering & system safety*, **55**(2), p. 151.
- [12] 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), pp. 1483–1510.

- [13] Heng, A., Zhang, S., Tan, A. C. C., and Mathew, J., 2009, “Rotating machinery prognostics: state of the art, challenges and opportunities,” *Mechanical Systems and Signal Processing*, **23**(3), pp. 724–739.
- [14] Austerlitz, H., 2003, *Data acquisition techniques using PCs*, Academic Press, San Diego.
- [15] Yurish, S., Shpak, N., Deynega, V., and Kirianaki, N. V., 2002, *Data Acquisition and Signal Processing for Smart Sensors*, Halsted Press.
- [16] Baydar, N. and Ball, A., 2003, “Detection of gear failures via vibration and acoustic signals using wavelet transform,” *Mechanical Systems and Signal Processing*, **17**(4), pp. 787–804.
- [17] Gu, S., Ni, J., and Yuan, J., 2002, “Non-stationary signal analysis and transient machining process condition monitoring,” *International Journal of Machine Tools and Manufacture*, **42**(1), pp. 41–51.
- [18] Luo, G. Y., Osypiw, D., and Irle, M., 2003, “On-line vibration analysis with fast continuous wavelet algorithm for condition monitoring of bearing,” *Journal of Vibration and Control*, **9**(8), pp. 931–947.
- [19] Zhang, S., Mathew, J., Ma, L., and Sun, Y., 2005, “Best basis-based intelligent machine fault diagnosis,” *Mechanical Systems and Signal Processing*, **19**(2), pp. 357–370.
- [20] 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), pp. 597–612.
- [21] Williams, J. H., Davies, A., and Drake, P. R., 2002, *Condition-Based Maintenance And Machine Diagnostics*, Kluwer Academic Publishers.
- [22] Korbicz, J., Koscielny, J. M., Kowalczyk, Z., and Cholewa, W., 2004, *Fault diagnosis: models, artificial intelligence, applications*, Springer.
- [23] Sohn, H. and Farrar, C. R., 2001, “Damage diagnosis using time series analysis,” *Smart materials & structures*, **10**(3), p. 446.
- [24] Hong, D., Xiuwen, G., and Shuzi, Y., 1991, “An approach to state recognition and knowledge-based diagnosis for engines,” *Mechanical Systems and Signal Processing*, **5**(4), pp. 257–266.
- [25] Dornfeld, D. A. and DeVries, M. F., 1990, “Neural network sensor fusion for tool condition monitoring,” *CIRP Annals - Manufacturing Technology*, **39**(1), pp. 101–105.
- [26] Samanta, B., 2003, “Artificial neural network based fault diagnostics of rolling element bearings using time-domain features,” *Mechanical Systems and Signal Processing*, **17**(2), p. 317.

- [27] Samanta, B., 2004, “Gear fault detection using artificial neural networks and support vector machines with genetic algorithms,” *Mechanical Systems and Signal Processing*, **18**(3), p. 625.
- [28] Silva, R., 1998, “Tool wear monitoring of turning operations by neural network and expert system classification of a feature set generated from multiple sensors,” *Mechanical Systems and Signal Processing*, **12**(2), p. 319.
- [29] Wang, W. and Christer, A. H., 2000, “Towards a general condition based maintenance model for a stochastic dynamic system,” *Journal of the Operational Research Society*, **51**(2), pp. 145–155.
- [30] Wang, G., Luo, Z., Qin, X., Leng, Y., and Wang, T., 2008, “Fault identification and classification of rolling element bearing based on time-varying autoregressive spectrum,” *Mechanical Systems and Signal Processing*, **22**(4), pp. 934 – 947.
- [31] Randall, R. B., Antoni, J., and Chobsaard, S., 2001, “The relationship between spectral correlation and envelope analysis in the diagnostics of bearing faults and other cyclostationary machine signals,” *Mechanical Systems and Signal Processing*, **15**(5), pp. 945–962.
- [32] Bonato, P., 1997, “Bilinear time-frequency transformations in the analysis of damaged structures,” *Mechanical Systems and Signal Processing*, **11**(4), p. 509.
- [33] Paya, B., 1997, “Artificial neural network based fault diagnostics of rotating machinery using wavelet transforms as a preprocessor,” *Mechanical Systems and Signal Processing*, **11**(5), p. 751.
- [34] Jin, J. and Shi, J., 2000, “Diagnostic feature extraction from stamping tonnage signals based on design of experiments,” *Journal of Manufacturing Science and Engineering*, **122**(2), pp. 360–369.
- [35] Rabiner, L. R., 1989, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, **77**(2), pp. 257–286.
- [36] Ertunc, H. M., Loparo, K. A., and Ocak, H., 2001, “Tool wear condition monitoring in drilling operations using hidden Markov models,” *International Journal of Machine Tools and Manufacture*, **41**(9), pp. 1363–1384.
- [37] Wang, L., Mehrabi, M. G., and Kannatey-Asibu, J. E., 2002, “Hidden Markov model-based tool wear monitoring in turning,” *Journal of Manufacturing Science and Engineering*, **124**(3), pp. 651–658.
- [38] Li, Z., Wu, Z., He, Y., and Fulei, C., 2005, “Hidden Markov model-based fault diagnostics method in speed-up and speed-down process for rotating machinery,” *Mechanical Systems and Signal Processing*, **19**(2), pp. 329–339.
- [39] Smyth, P., 1994, “Markov monitoring with unknown states,” *IEEE Journal on Selected Areas in Communications*, **12**, pp. 1600–1612.

- [40] Smyth, P., 1994, "Hidden Markov-models for fault-detection in dynamic-systems," *Pattern recognition*, **27**(1), pp. 149–164.
- [41] Tang, K., Williams, W. W., Jwo, W., and Gong, L. G., 1999, "Performance comparison between on-line sensors and control charts in manufacturing process monitoring," *IIE transactions*, **31**(12), pp. 1181–1190.
- [42] Ross, S. M., 1996, *Stochastic Processes*, Probability and Statistics, John Wiley & Sons Inc.
- [43] Viterbi, A., 1967, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *Information Theory, IEEE Transactions on*, **13**(2), pp. 260–269.
- [44] Forney, J., G. D., 1973, "The Viterbi algorithm," *Proceedings of the IEEE*, **61**(3), pp. 268–278.
- [45] Baum, L. E. and Petrie, T., 1966, "Statistical inference for probabilistic functions of finite state Markov chains," *The Annals of Mathematical Statistics*, **37**(6), pp. 1554–1563.
- [46] Dempster, A. P., Laird, N. M., and Rubin, D. B., 1977, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), pp. 1–38.
- [47] Duda, R. O., Laird, N. M., and Stork, D. G., 2001, *Pattern Classification*, John Wiley & Sons Inc.
- [48] Montgomery, D. C., 2004, *Introduction to Statistical Quality Control*, John Wiley & Sons Inc.
- [49] Lowry, C. A. and Montgomery, D. C., 1995, "A review of multivariate control charts," *IIE transactions*, **27**(6), p. 800.
- [50] Jackson, J. E., 1985, "Multivariate quality-control," *Communications in statistics. Theory and methods*, **14**(11), pp. 2657–2688.
- [51] Bishop, C. M., 2006, *Pattern Recognition and Machine Learning*, Springer.
- [52] Ryan, T. P., 2000, *Statistical Methods for Quality Improvement*, John Wiley & Sons Inc.
- [53] Nelson, L., 1984, "The Shewhart control chart - tests for special causes," *Journal of Quality Technology*, **16**(4), pp. 237–239.
- [54] Jain, A. K., Mao, J., and Mohiuddin, K. M., 1996, "Artificial neural networks: a tutorial," *IEEE Computer*, **29**(3), pp. 31–44.
- [55] Jain, A., Murty, M., and Flynn, P., 1999, "Data clustering: a review," *ACM computing surveys*, **31**(3), pp. 264–323.

- [56] Dekker, R., 1996, “Applications of maintenance optimization models: a review and analysis,” *Reliability Engineering & System Safety*, **51**(3), pp. 229–240.
- [57] William P. Pierskalla, J. A. V., 1976, “A survey of maintenance models: the control and surveillance of deteriorating systems,” *Naval Research Logistics Quarterly*, **23**(3), pp. 353–388.
- [58] Ciriaco Valdez-Flores, R. M. F., 1989, “A survey of preventive maintenance models for stochastically deteriorating single-unit systems,” *Naval Research Logistics*, **36**(4), pp. 419–446.
- [59] Chan, G. K. and Asgarpour, S., 2006, “Optimum maintenance policy with Markov processes,” *Electric Power Systems Research*, **76**(6-7), pp. 452–456.
- [60] Chen, D. and Trivedi, K. S., 2005, “Optimization for condition-based maintenance with semi-Markov decision process,” *Reliability Engineering & System Safety*, **90**(1), pp. 25–29.
- [61] Maillart, L. M., 2006, “Maintenance policies for systems with condition monitoring and obvious failures,” *IIE Transactions*, **38**(6), p. 463(13).
- [62] Amari, S. V., McLaughlin, L., and Pham, H., 2006, “Cost-effective condition-based maintenance using Markov decision processes,” *Proceedings of the RAMS '06. Annual Reliability and Maintainability Symposium, 2006.*, pp. 464–469.
- [63] Sim, S. H. and Endrenyi, J., 1988, “Optimal preventive maintenance with repair,” *Reliability, IEEE Transactions on*, **37**(1), pp. 92–96.
- [64] Sim, S. H. and Endrenyi, J., 1993, “A failure-repair model with minimal and major maintenance,” *Reliability, IEEE Transactions on*, **42**(1), pp. 134–140.
- [65] Das, T. K., Gosavi, A., Mahadevan, S., and Marchallick, N., 1999, “Solving semi-Markov decision problems using average reward reinforcement learning,” *Management Science*, **45**(4), pp. 560–574.
- [66] Berenguer, C., Chu, C. B., and Grall, A., 1997, “Inspection and maintenance planning: an application of semi-Markov decision processes,” *Journal of Intelligent Manufacturing*, **8**(5), pp. 467–476.
- [67] Cho, D. I. and Parlar, M., 1991, “A survey of maintenance models for multunit systems,” *European Journal of Operational Research*, **51**(1), pp. 1–23.
- [68] Wang, W. and Christer, A. H., 2003, “Solution algorithms for a nonhomogeneous multi-component inspection model,” *Computers & Operations Research*, **30**(1), pp. 19–34.
- [69] Sherif, Y. S. and Smith, M. L., 1981, “Optimal maintenance models for system subject to failure - a review,” *Naval Research Logistics*, **28**(1), pp. 47–74.

- [70] Assaf, D. and Shanthikumar, J. G., 1987, “Optimal group maintenance policies with continuous and periodic inspections,” *Management Science*, **33**(11), pp. 1440–1452.
- [71] Das, K., Lashkari, R. S., and Sengupta, S., 2007, “Machine reliability and preventive maintenance planning for cellular manufacturing systems,” *European Journal of Operational Research*, **183**(1), pp. 162–180.
- [72] Osogami, T. and Harchol-Balter, M., 2006, “Closed form solutions for mapping general distributions to quasi-minimal PH distributions,” *Performance Evaluation*, **63**(6), pp. 524–552.
- [73] David, A. and Larry, S., 1987, “The least variable phase type distribution is Erlang,” *Stochastic Models*, **3**(3), pp. 467 – 473.
- [74] Ruiz-Castro, J. E., Fernandez-Villodre, G., and Perez-Ocon, R., 2009, “A level-dependent general discrete system involving phase-type distributions,” *IIE Transactions*, **41**(1), pp. 45 – 56.
- [75] Meerkov, S. M. and Zhang, L., 2008, “Transient behavior of serial production lines with bernoulli machines,” *IIE Transactions*, **40**(3), pp. 297–312.
- [76] Liu, Y., 2007, “Predictive modeling of multivariate stochastic dependencies in semiconductor manufacturing,” SRC Deliverable Report Task ID 1222.001.
- [77] Sloan, T. W. and Shanthikumar, J. G., 2002, “Using in-line equipment condition and yield information for maintenance scheduling and dispatching in semiconductor wafer fabs,” *IIE transactions*, **34**(2), p. 191.
- [78] Sloan, T. W. and Shanthikumar, J. G., 2000, “Combined production and maintenance scheduling for a multiple-product, single-machine production system,” *Production and Operations Management*, **9**(4), pp. 379–399.
- [79] Takahashi, K. M. and Daugherty, J. E., 1996, “Current capabilities and limitations of *in situ* particle monitors in silicon processing equipment,” *Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films*, **14**(6), pp. 2983–2993.
- [80] Miyashita, H., Kikuchi, T., Kawasaki, Y., Katakura, Y., and Ohsako, N., 1999, “Particle measurements in vacuum tools by *in situ* particle monitor,” *Journal of Vacuum Science & Technology A: Vacuum Surfaces and Films*, **17**(3), pp. 1066–1070.
- [81] Borden, P. G. and Larson, L. A., 1989, “Benefits of real-time, *in situ* particle monitoring in production medium current implantation,” *Semiconductor Manufacturing*, *IEEE Transactions on*, **2**(4), pp. 141–145.
- [82] Djurdjanovic, D. and Liu, Y., 2006, “Survey of predictive maintenance research and industry best practice,” University of Michigan, Ann Arbor, MI.

- [83] Fernandez, E., Fu, M., and Marcus, S., 2002, "Survey of current best practices in preventive maintenance scheduling in semiconductor manufacturing," SRC Deliverable Report Task ID 877-001.
- [84] Yano, C. A. and Lee, H. L., 1995, "Lot-sizing with random yields: a review," *Operations Research*, **43**(2), p. 311.
- [85] Kim, C., Hong, Y., and Chang, S., 2001, "Optimal production run length and inspection schedules in a deteriorating production process," *IIE Transactions*, **33**(5), pp. 421–426.
- [86] Lee, H. L. and Rosenblatt, M. J., 1989, "A production and maintenance planning model with restoration cost dependent on detection delay," *IIE Transactions*, **21**(4), pp. 368 – 375.
- [87] Makis, V. and Fung, J., 1998, "An EMQ model with inspections and random machine failures," *The Journal of the Operational Research Society*, **49**(1), pp. 66–76.
- [88] Freimer, M., Thomas, D., and Tyworth, J., 2006, "The value of setup cost reduction and process improvement for the economic production quantity model with defects," *European Journal of Operational Research*, **173**(1), pp. 241–251.
- [89] Sloan, T., 2008, "Simultaneous determination of production and maintenance schedules using in-line equipment condition and yield information," *Naval Research Logistics*, **55**(2), pp. 116–129.
- [90] Zhou, J., Djurdjanovic, D., Ivy, J., and Ni, J., 2007, "Integrated reconfiguration and age-based preventive maintenance decision making," *IIE Transactions*, **39**(12), pp. 1085 – 1102.
- [91] Monden, Y., 1983, *Toyota Production System: Practical Approach to Production Management*, Industrial Engineering and Management Press.
- [92] Shingo, S., 1985, *A Revolution in Manufacturing: The SMED System*, Productivity Press.
- [93] Lee, S., Li, L., and Ni, J., 2010, "Online degradation assessment and adaptive fault detection using modified hidden Markov model," *Journal of Manufacturing Science and Engineering*, **132**(2), pp. 021010–11.
- [94] Bellman, R., 1957, "A Markovian decision process," *Journal of Mathematics and Mechanics*, **6**(5), pp. 679–684.
- [95] Manne, A. S., 1960, "Linear programming and sequential decisions," *Management Science*, **6**(3), pp. 259–267.
- [96] Ambani, S., Li, L., and Ni, J., 2009, "Condition-based maintenance decision-making for multiple machine systems," *Journal of Manufacturing Science and Engineering*, **131**(3), pp. 031009–9.

- [97] Yang, Z., Chang, Q., Djurdjanovic, D., Ni, J., and Lee, J., 2007, “Maintenance priority assignment utilizing on-line production information,” *Journal of Manufacturing Science and Engineering*, **129**(2), pp. 435–446.
- [98] Chang, Q., Ni, J., Bandyopadhyay, P., Biller, S., and Xiao, G., 2007, “Maintenance staffing management,” *Journal of Intelligent Manufacturing*, **18**(3), pp. 351–360.