VARIATION MODELING, ANALYSIS AND CONTROL FOR MULTISTAGE WAFER MANUFACTURING PROCESSES

A Thesis
Presented to
The Academic Faculty

by

Ran Jin

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
August, 2011

VARIATION MODELING, ANALYSIS AND CONTROL FOR MULTISTAGE WAFER MANUFACTURING PROCESSES

Approved by:

Dr. Jianjun Shi, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. C. F. Jeff Wu
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Roshan Joseph Vengazhiyil
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Nagi Gebraeel
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Shreyes N. Melkote
School of Mechanical Engineering
Georgia Institute of Technology

Date Approved: April 21, 2011

To my beloved father Qun Jin and mother Ping Yuan

ACKNOWLEDGEMENTS

I would like to express my utmost gratitude to my advisor Professor Jianjun Shi, for his supervision and encouragement throughout my research. I will never forget his inspiration and guidance as I was facing the obstacles in the completion of my thesis. Without his help, this thesis would not have been possible.

I would like to express my sincere gratitude to my other thesis committee members, including Professor Nagi Gebraeel, Professor Shreyes N. Melkote, Professor Roshan Joseph Vengazhiyil and Professor C. F. Jeff Wu for their insightful advice and constructive comments to improve my research, and for their considerations and support during my Ph.D study.

I would like to thank my professors in the academic society, who provides me invaluable advice and continuous support for my academic career, including Professor Jionghua Jin, Professor Lawrence M. Seiford and Professor Ji Zhu from University of Michigan, Professor Yu Ding from Texas A&M University, and Professor Kwok-Leung Tsui from Georgia Institute of Technology.

I would like to thank Dr. Tzyy-Shuh Chang, Dr. Hongbin Jia, and all their colleagues in OG Technologies, Inc., and my other industrial collaborators for the technical support and help in the data collection.

I cherish my collaboration with Professor Kaibo Wang from Tsinghua University, Beijing. Professor Wang provides me many constructive comments for my research and help on data collection and research validation. I also want to express my sincere gratitude to him.

I would also like to thank all my colleagues, friends and visiting scholars in our research lab for great friendships and constructive comments in my research, including but not limited to my friends at University of Michigan: Dr. Jian Guo, Dr. Eduardo Izquierdo, Dr. Yong Lei, Dr. Jing Li, Mr. Qiang Li, Dr. Jian Liu, Dr. Jing Zhong, Dr. Zhisheng Zhang, and my friends at Georgia Institute of Technology: Mr. Shan Ba, Ms. Chia-Jung Chang, Dr. Xinwei Deng, Ms. Li Hao, Dr. Lulu Kang, Mr. Kaibo Liu, Mr. Matthew Plumlee, Dr. Fei Sun, Dr. Su Wu, Dr. Liang Ye, Dr. Weidong Zhang and Dr. Hongxu Zhao. A special gratitude goes to Dr. Xinwei Deng and Dr. Jian Liu, for giving me tremendous encouragement and help throughout my Ph.D study, especially when I was at the most difficult time in my research.

Last, but not the least, I want to thank my father Qun Jin, mother Ping Yuan, and my other family members. Their warm encouragement and understanding give me the strength to overcome the difficulties for many years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation

Geometric quality characteristics of wafers, such as BOW and WARP, are critical in their applications. A large variation of these quality variables reduces the number of conforming products in the downstream production. Therefore, it is important to reduce the variation by modeling the variation propagation, and further by developing the variation reduction methodology. However, a wafer manufacturing process is a very complex process involving mechanical and chemical operations on wafers. Typical operations include slicing, lapping, chemical etching, chemical vapor deposition, and polishing. There are no engineering models available to model the multistage variation propagation. On the other hand, with the rapid development of sensing technology, massive observational data may be obtained from the wafer manufacturing processes. These observational data characterize the wafer manufacturing processes by providing quality, process and material property measurements. This data-rich environment provides opportunities to advance the research in quality control methodology, while it poses the challenges including the high dimensionality and heterogeneity of the data, and effectiveness in complex manufacturing process modeling. To address these challenges, a unified variation modeling, analysis and control methodology is developed for multistage wafer manufacturing processes (MWMPs).

1.2 State-of-the-art

In a multistage manufacturing process, there are different ways to model the variation propagation and improve the quality. One methodology is called Stream of Variation (SoV), which is developed based on state space models (Jin and Shi, 1999; Shi, 2006; Shi and Zhou, 2009). The SoV approaches are capable of reducing the variation through control (Djurdjanovic and Zhu, 2005; Izquierdo *et al*., 2007; Jiao and Djurdjanovic, 2010). However, the variation reduction performance of this type of approaches depends on the validity and accuracy of state space models. Other methodologies are developed based on regression models, such as Robust Parameter Design (RPD) based feedforward control (Joseph, 2003) and DOE-based automatic process control (APC) (Jin and Ding, 2004; Zhong *et al.*, 2010). These regression models are estimated from the experimental data, which may be too expensive to obtain in a production system with many potential factors. Moreover, the single regression model strategy can not address complex situations in a multistage manufacturing process when the data structure is nonlinear. Therefore, there is a pressing need to develop advanced models from the data with high dimensionality and heterogeneity.

To develop advanced models based on the observational data, it is important to obtain important quality features for a wafer more quickly. Fast and accurate measurements of those features are crucial for variation reduction and feedforward control. Due to the advancement of sensing technology, these quality features may be measured as highly spatial correlated profile data, such as geometric profiles in wafer manufacturing processes. However, current wafer profile measurement scheme is time consuming, which is essentially an off-line technology and hence unable to provide quick

assessment of wafer quality in a timely manner. It is desirable to develop a measurement strategy to select the representative samples and develop models for the profile data. There are different ways to select the representative samples, such as grid spacing approaches in spatial statistics (Curran and Williamson, 1986; Curran 1988; McBratney and Webster, 1983a; McBratney and Webster, 1983b; Atkinson *et al.*, 1992; Atkinson *et al.*, 1994; Wang *et al.*, 2005; Xiao *et al.*, 2005; Anderson *et al.*, 2006), Sequential Monte Carlo methods (Liu and Chen, 1998; Doucet *et al.*, 2000; Doucet *et al.*, 2001; Guo and Wang, 2004) and design of computer experiments (Schonlau *et al.*, 1998; Williams *et al.*, 2000; Park *et al.*, 2002; Kleijnen and Beers, 2004; Huang *et al.*, 2006). However, these approaches have limitations in linking the local variability directly with the sample locations, or the computation is too intensive to be used for online measurement. It is highly desirable to develop methodology to reduce the measured sample size, while achieving required accuracy for online applications.

Based on the observational data and developed models to link the quality variables with process and material property measurements, a quick detection of changes in a multistage manufacturing process is also important for quality assurance and improvement. The conventional statistical process control (SPC) (Lowry and Montgomery, 1995; Woodall and Montgomery, 1999) monitors the final product quality without consideration of the inter-stage relationships. Thus, it is difficult to identify the stages with assignable causes. Regression model based risk-adjusted approaches (Hawkins, 1991, 1993; Shu *et al*., 2004a; Zhang, 1985, 1992; Shu *et al*., 2004b) and engineering model based risk-adjusted approaches (Xiang and Tsung, 2008) monitors the residuals and the covariates, thus distinguishing the process change at the current stage or

that from the upstream stages. However, these approaches assume only one baseline model under normal conditions, which may not be true in a complex manufacturing process, such as a MWMP. It is important to develop monitoring methodology to monitoring such a manufacturing process with multiple baseline models linked in multistage, which we call a multistage multimode process (MMOP).

1.3 Research Objectives

The objectives of this research are:

- identifying the unique characteristics of the observational data and extracting pertinent knowledge about wafer manufacturing systems for quality control by the integration of statistics, domain knowledge, and control;
- developing control strategy with the consideration of intermediate quality measurements and sensing noise for variation reduction of wafer geometric variables;
- developing efficient measurement strategy for the modeling of wafer geometric profile data;
- studying the monitoring of a multistage multimode wafer manufacturing process considering the modeling uncertainty.

1.4 Organization of the Thesis

This thesis presents variation modeling, analysis and control for multistage wafer manufacturing processes in a multiple manuscript format. Chapters 2, 3, 4 and 5 are written as research papers. The relationship among these chapters is shown in Figure 1.1.

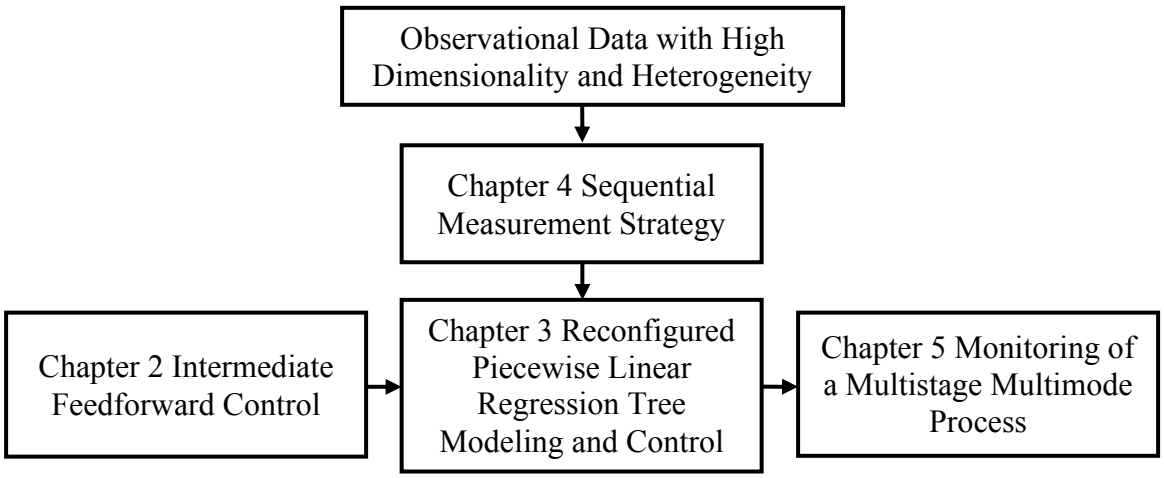


Figure 1.1: Thesis chapters

In Chapter 2 (Jin and Shi, manuscript), a group of regression models is used to capture the stage-to-stage variation. An intermediate feedforward control strategy is developed to adjust and update the control actions based on the online measurements of intermediate wafer quality measurements. The control performance is evaluated in a MWMP to transform ingots into polished wafers.

In Chapter 3 (Jin and Shi, 2011), piecewise linear regression tree (PLRT) models are used to address nonlinear relationships in MWMP to improve the model prediction performance. The obtained PLRT model is further reconfigured to comply with the physical layout of the MWMP for feedforward control purposes. The model complexity is further reduced by merging the leaf nodes with satisfied control accuracy. The procedure and effectiveness of the proposed method is shown in a case study of a MWMP.

In Chapter 4 (Jin *et al*., 2011), a sequential measurement strategy is proposed to reduce the number of samples measured in a wafer, yet to provide adequate accuracy for the quality feature estimation. A Gaussian process model is used to estimate the true

profile of a wafer. The predicted profile and its variance serve as guidelines to determine the measurement locations, thus to improve the sensing efficiency.

In Chapter 5 (Jin *et al*., manuscript), we study the monitoring problem of a MMOP. We propose to use PLRTs to inter-relate the variables in a MMOP. A unified charting system is developed based on the PLRTs for process monitoring. Because of the challenges to capture the baseline models to represent multimode processes, we further study the run length distribution, and optimize the control chart system by considering the modeling uncertainties. Finally, we compare the proposed method with the risk adjustment type of control chart systems based on global regression models, for both simulation study and a MWMP.

Finally, Chapter 6 concludes the thesis, summarizes the original contributions and discusses the future research directions.

CHAPTER 2

INTERMEDIATE FEEDFORWARD CONTROL IN MULTISTAGE WAFER MANUFACTURING PROCESSES

2.1 Introduction

Wafer manufacturing is a very complicated process involving mechanical and chemical operations on wafers. A typical process involves multiple operations to transform a silicon ingot into polished wafers with thin films deposited on one side of the wafers. Typical operations include slicing (wire sawing), lapping, chemical etching, chemical vapor deposition (CVD), and polishing. Therefore, the process is called a "multistage wafer manufacturing process (MWMP)". In a MWMP, the first manufacturing stage is to slice an ingot into wafers with rough surfaces through a wire sawing operation. Then lapping and chemical etching processes are used to improve the surfaces of wafers through by removing the mechanical cracks and reduce the roughness of the wafers' surfaces. After the cleaning process, thin films such as polysilicon or silicon dioxide may be deposited on the surfaces of wafers, typically completed in low pressure chemical vapor deposition (LPCVD) chambers or belt type conveyers. After the thin film deposition, the wafers are polished to achieve mirror-like surfaces.

In a MWMP, the geometric quality variables, such as BOW and WARP, are very important quality indexes to measure the surface roughness and flatness for downstream productions of wafers. The smaller the geometric quality variables are, the better the quality of the wafers are. A large variation of the quality variables will increase the

nonconforming products in the downstream stage. Therefore, the quality of wafers needs to be improved during the production to reduce energy and material wastes. In a MWMP, the quality of wafers is changed by its potential factors in a complex mechanism, not only in material removal processes, but also by the stress of thin films. Important factors that introduce variation in geometric variables are the process variables and the material property of wafers. To improve the geometric quality at the final stage, these important factors should be set or adjusted at different stages.

There are three typical methods to set or adjust the important factors: Robust Parameter Design (RPD), Stream of Variation (SoV) and Design of Experiment based Automatic Process Control (DOE-based APC). The RPD (Taguchi *et al.*, 1989) builds linear regression models based on the experimental data. Then it determines the optimal settings of controllable variables off-line by solving the nominal-the-best or the smaller-the-better problems. These settings are used to reduce the sensitivity of controllable variables to the noise factors. This approach provides a robust performance in a manufacturing process with fixed settings of controllable variables during operation.

The second typical method, i.e., the SoV method, uses a state space model to characterize the variation and its propagations in an MMP. In this method, the controllable variables minimize the deviation or variation of the final quality variables (Jin and Shi, 1999; Djurdjanovic and Zhu, 2005; Izquierdo *et al.*, 2007). These approaches successfully identify the control actions considering the physical specification and intermediate quality measurements. However, the SoV usually assumes a Markov property of stages. It may also require engineering knowledge in model construction.

When the Markov property is insufficient or the engineering knowledge is inadequate, the third typical method, i.e., the DOE-based APC, builds regression models from the experimental data. It adjusts the controllable process variables automatically during the production. Different types of variables and variation sources are considered, such as controllable variables, and measurable and immeasurable noise variables. For example, the reaction time in CVD process is a controllable variable; the sensor noise of WARP is a measurable noise variable; and the slurry distribution during the lapping process is an immeasurable noise variable. A cautious control strategy addresses the sensing and modeling errors in nominal-the-best problems (Shi, 2006; Jin and Ding, 2004; Shi *et al*., 2005; Zhong *et al.*, 2009). A DOE-based APC has much better performance in variation reduction than the traditional offline RPD. However, it is difficult to directly implement the DOE-based APC in an MMP because it models an MMP with a single regression model. Furthermore, the DOE-based APC is not applicable for the cases where online control actions are needed during intermediate stages of an MMP.

In addition to the three typical methods used for variation reduction, one uses regression models to identify the variation sources distributed at different stages, and model the variation propagation in an MMP (Lawless *et al.*, 1999; Agrawal *et al.*, 1999). In this method, a quality variable at the k^{th} stage is predicted by the quality variable at the $(k-1)^{th}$ stage and covariates at the k^{th} stage, shown as

$$Y_k = \alpha_k + \beta_k Y_{k-1} + \gamma_k z_k + \varepsilon_k \tag{2.1}$$

where Y_{k-1} and Y_k are the quality variables at the $(k-1)^{th}$ stage and the k^{th} stage, respectively; z_k are the covariates at the k^{th} stage; α_k, β_k and γ_k are the corresponding parameters; ε_k is the residual. This model is successful in variation analysis. However,

there are two limitations in this method: (1) One assumes the quality variables have Markov property, i.e., Y_k is independent of $Y_1, Y_2, ..., Y_{k-2}$ conditioning on Y_{k-1}. Some manufacturing processes do not hold the Markov properties. We illustrate one example in the case study in Section 2.3. (2) Another limitation is that the covariates z_k may not include the controllable variables, thus, we may not use these models in a control application.

This chapter proposes an integrated modeling and control strategy for variation reduction in an MMP, which is an extension based on Equation (2.1) by further considering process variables and the controllability of process variables. Here, we assume that the same quality variables can be measured repeatedly after each manufacturing stage. A *group* of regression models is constructed from the observational production data. These models predict the downstream quality variables stage-by-stage with the data obtained in their upstream process. With the help of the model group, we determine the control actions by solving constrained optimization problems. The proposed approach is based on the following three assumptions:

- A group of regression models describes the process and predicts the intermediate and final quality of future production with acceptable prediction error.
- The intermediate quality specifications and controllability of process variables can be presented as inequalities in the optimization problems.
- The control optimization problem at each controllable stage is solvable to minimize the final quality variation.

Based on the assumptions, the rest of the chapter is organized as follows: we propose the methodology for the variation propagation models and intermediate feedforward control in Section 2.2. We further use a five-stage MWMP as a case study to illustrate the modeling and control procedure in Section 2.3. Finally, we draw the conclusions in Section 2.4.

2.2 Intermediate Feedforward Control Strategy

We call our proposed method "intermediate feedforward control strategy", since we adjust the control actions at intermediate stage based on a group of regression models. In this section, we first provide an overview of the methods. Then we introduce in detail the regression modeling, intermediate feedforward control strategy formulation, and control action determination. Finally, we discuss the impact of sensing noise on the control objective function, since the quality measurements are important to adjust the control actions.

2.2.1 Overview of the Proposed Methods

To show the intermediate feedforward control strategy, we illustrate the procedure in Figure 2.1. First, we observe the initial quality variables and material property variables at the beginning of the production. Once the production starts, we identify if the next stage is controllable, i.e., the stage has controllable variables. If the stage is not controllable, we go to the next stage. Otherwise, we optimize the predicted final quality by determining a set of optimal control actions at all downstream stages. From the set of optimal control actions, we only take the control actions at the current stage. After taking the control actions, the intermediate quality measurements of the current stage become available. If we need additional adjustment of the control variables at the downstream

stages, we further use the intermediate quality measurements to update the downstream control actions in iterations. Otherwise, the control ends when there are no additional control actions to be determined.

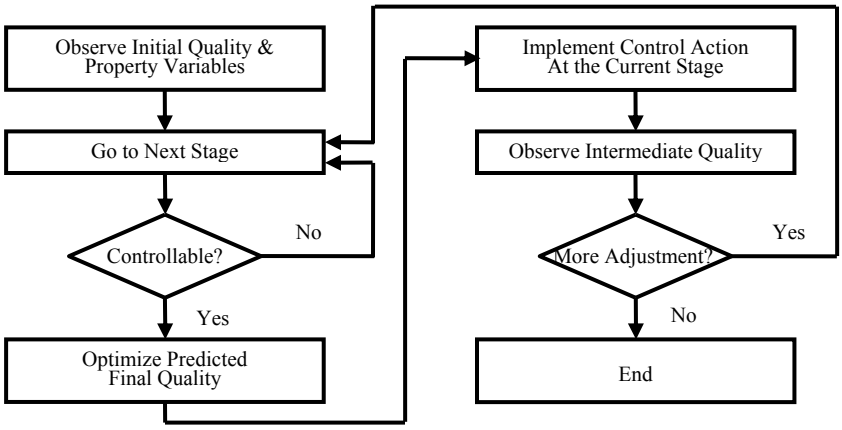


Figure 2.1: Intermediate feedforward control procedure

2.2.2 Regression Model Group

In the intermediate feedforward control, a regression model group predicts the final quality in the optimization problem. This model group also describes the variation propagation in a MMP. Figure 2.2 shows a layout of an MMP with N stages. Four types of variables describe the process, including the quality measurements at the kth stage ($\mathbf{Y}(\mathrm{k}) \in \Re^{m_\mathrm{k} \times 1}$), the continuous online controllable variables at the kth stage ($\mathbf{U}_\mathrm{k} \in \Re^{r_\mathrm{k} \times 1}$), the offline setting variables at the kth stage ($\mathbf{X}_\mathrm{k} \in \Re^{n_\mathrm{k} \times 1}$), and the material property variables independent of stages ($\mathbf{M} \in \Re^{t \times 1}$). Furthermore, u_{ik} is the ith continuous online controllable variable at the kth stage ($\mathrm{i} = 1, \cdots, r_\mathrm{k}$); and x_{ik} is the ith offline setting variable at the kth stage ($\mathrm{i} = 1, \cdots, n_\mathrm{k}$).

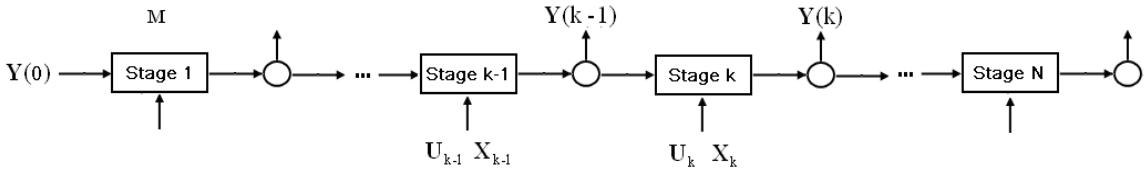


Figure 2.2: A layout of an MMP

To model the variation propagation, we predict $\mathbf{Y}(\mathrm{k})$ ($\mathrm{k}=1,2,\cdots,\mathrm{N}$) by the quality and the process variables measured in all upstream stages, and the material property variables. In this way, even if the MMP does not have Markov property, the MMP can still be modeled by a group of regression models. Denote $\mathbf{Y}^{l}=\left[\mathbf{Y}(0)^{\mathrm{T}} \quad \mathbf{Y}(1)^{\mathrm{T}} \quad \cdots \quad \mathbf{Y}(l)^{\mathrm{T}}\right]^{\mathrm{T}}$, $\mathbf{U}^{l}=\left[\mathbf{U}_1^{\mathrm{T}} \quad \mathbf{U}_2^{\mathrm{T}} \quad \cdots \quad \mathbf{U}_l^{\mathrm{T}}\right]^{\mathrm{T}}$, and $\mathbf{X}^{l}=\left[\mathbf{X}_1^{\mathrm{T}} \quad \mathbf{X}_2^{\mathrm{T}} \quad \cdots \quad \mathbf{X}_l^{\mathrm{T}}\right]^{\mathrm{T}}$ ($l=1,\cdots,\mathrm{N}$), the prediction model is:

$$\mathbf{Y}(\mathrm{k})_{\mathrm{j}}=\mathbf{\Phi}_{\mathrm{k}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{k,k}}^{\mathrm{j,\Phi}}\mathbf{\Phi}_{\mathrm{k}}+\boldsymbol{\psi}_{\mathrm{k}}^{\mathrm{T}}(\boldsymbol{\beta}_{\mathrm{k,k}}^{\mathrm{j,\psi,U}})\mathbf{U}_{\mathrm{k}}+\boldsymbol{\psi}_{\mathrm{k}}^{\mathrm{T}}(\boldsymbol{\beta}_{\mathrm{k,k}}^{\mathrm{j,\psi,X}})\mathbf{X}_{\mathrm{k}}+\varepsilon_{\mathrm{kj}} \tag{2.2}$$

where $\mathbf{Y}(\mathrm{k})_{\mathrm{j}}$ is the j^{th} quality variable at the k^{th} stage ($\mathrm{j}=1,2,...,m_{\mathrm{k}}$); $\mathbf{\Phi}_{\mathrm{k}}=[1 \quad \mathbf{Y}^{\mathrm{k\text{-}1}^{\mathrm{T}}} \quad \mathbf{U}^{\mathrm{k\text{-}1}^{\mathrm{T}}} \quad \mathbf{X}^{\mathrm{k\text{-}1}^{\mathrm{T}}} \quad \mathbf{M}^{\mathrm{T}}]^{\mathrm{T}}$; $\boldsymbol{\psi}_{\mathrm{k}}=[1 \quad \mathbf{Y}^{\mathrm{k\text{-}1}^{\mathrm{T}}} \quad \mathbf{M}^{\mathrm{T}}]^{\mathrm{T}}$; $\boldsymbol{\beta}_{\mathrm{k,k}}^{\mathrm{j,\Phi}}$, $\boldsymbol{\beta}_{\mathrm{k,k}}^{\mathrm{j,\psi,U}}$ and $\boldsymbol{\beta}_{\mathrm{k,k}}^{\mathrm{j,\psi,X}}$ are the corresponding coefficient matrices in the regression model with proper dimensions; and $\varepsilon_{\mathrm{kj}}$ is the residual. The term $\mathbf{\Phi}_{\mathrm{k}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{k,k}}^{\mathrm{j,\Phi}}\mathbf{\Phi}_{\mathrm{k}}$ is the contribution of the observed information up to the k^{th} stage. The terms $\boldsymbol{\psi}_{\mathrm{k}}^{\mathrm{T}}(\boldsymbol{\beta}_{\mathrm{k,k}}^{\mathrm{j,\psi,U}})\mathbf{U}_{\mathrm{k}}$ and $\boldsymbol{\psi}_{\mathrm{k}}^{\mathrm{T}}(\boldsymbol{\beta}_{\mathrm{k,k}}^{\mathrm{j,\psi,X}})\mathbf{X}_{\mathrm{k}}$ are the contribution of the control actions at the current stage k. In practice, we select the predictors by using both the engineering knowledge and statistical method to further reduce model complexity. The final model structure is determined by 10-fold cross validation using mean sum of square error.

In this way, we use Equation (2.2) to predict each quality variable before the operation takes place. In an MMP involving multiple quality variables, multiple

regression models forms a *regression model group* to predict these variables. The quality variables are predicted sequentially from the first stage to the last stage. When the intermediate quality variables are not available, we substitute the predicted values of the intermediate quality variables to predict the final quality variables.

2.2.3 Feedforward Control Strategy

Once we can predict the final quality variables, we determine the controllable variables to reduce the predicted final quality variation before the operations take place, illustrated in Figure 2.3. Figure 2.3 shows the procedure of intermediate adjustment between the (k-1)th stage and the kth stage. When the operations at the (k-1)th stage finish, the intermediate quality measurements $\hat{\mathbf{Y}}(k-1)$ become available. Due to the prediction error or uncertainties in an MMP, deviations exist between the predicted quality $\hat{\hat{\mathbf{Y}}}(k-1)$ at the last iteration and the actual measurements $\hat{\mathbf{Y}}(k-1)$. Therefore, the controllable variables of the downstream stages (from the kth stage to the Nth stage) need adjustments in the following steps: (1) we first collect $\mathbf{\Phi}_k$, the quality, process, and material information of all upstream stages; (2) then we predict the downstream quality variables sequentially using Equation (2.2); (3) by solving a constrained optimization problem shown in Equation (2.3), we update the control actions of the downstream stages; (4) we implement the optimized $\mathbf{U}_k$ and $\mathbf{X}_k$, the control actions at the kth stage; and (5) we move to the next stage for control action updating, until the last stage N is achieved.

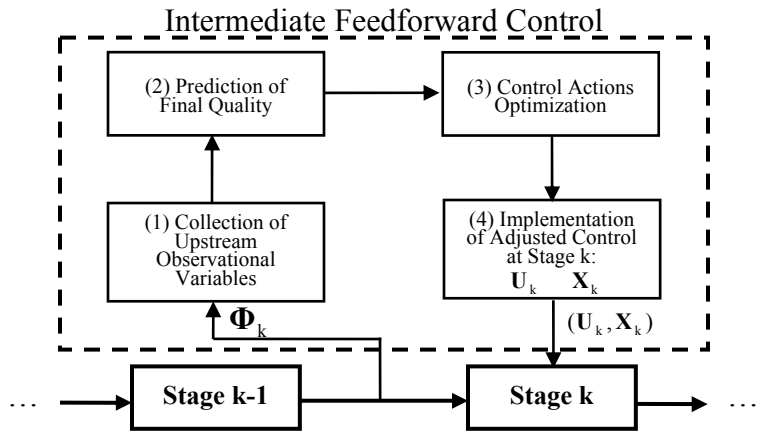


Figure 2.3: The procedure of adjustment between two stages

In Step (2), the constrained optimization problem uses the-smaller-the-better control objective in MWMPs. When the operations at the (k-1)th stage finish, the optimization problem between the (k-1)th stage and the kth stage is:

$$\begin{aligned} \min_{\mathbf{U}_l, \mathbf{X}_l, l=\mathrm{k},\mathrm{k}+1,\cdots,\mathrm{N}} \quad & J(\mathbf{U}_l, \mathbf{X}_l) = \sum_{\mathrm{j}=1}^{\mathrm{m}} \mathrm{c}_\mathrm{j} \mathrm{E}(\mathbf{Y}(\mathrm{N}|\mathrm{k})_\mathrm{j}^2) \\ s.t. \quad & \mathbf{Y}(l)_\mathrm{j} = \mathbf{\Phi}_l^{\mathrm{T}} \boldsymbol{\beta}_{l,l}^{\mathrm{j},\mathbf{\Phi}} \mathbf{\Phi}_l + \boldsymbol{\psi}_l^{\mathrm{T}} (\boldsymbol{\beta}_{l,l}^{\mathrm{j},\boldsymbol{\psi},\mathbf{U}}) \mathbf{U}_l + \boldsymbol{\psi}_l^{\mathrm{T}} (\boldsymbol{\beta}_{l,l}^{\mathrm{j},\boldsymbol{\psi},\mathbf{X}}) \mathbf{X}_l + \varepsilon_{l\mathrm{j}} \\ & g(\mathbf{Y}(s)_\mathrm{j}) < L_{\mathrm{j}s}, \\ & u_{il}^{\mathrm{L}} < u_{il} < u_{il}^{\mathrm{U}}, \forall \mathrm{i} \\ & x_{il} \in \{x_{il}\}, \forall \mathrm{i} \\ & l = \mathrm{k}, \mathrm{k}+1, \cdots, \mathrm{N} \\ & s = 1,2, \cdots, \mathrm{N} \end{aligned} \tag{2.3}$$

where the objective function $J(\mathbf{U}_l, \mathbf{X}_l)$ is a weighted summation of the second moment of m quality variables at the final stage N; c_j is the weight of the jth variable, determined by domain knowledge or requirements, such as the cost due to the inferior quality performance; and the decision variables are all downstream control variables $\mathbf{U}_l$ and $\mathbf{X}_l$ ($l = \mathrm{k}, \mathrm{k}+1, \cdots \mathrm{N}$). In the optimization problem, we also formulate three types of constraints. The first type of constraints is the group of regression models to predict the quality variables at downstream stages. The second type of constraints represent

intermediate quality specifications, modeled as $g(\mathbf{Y}(s)_j) < L_{js}$ for the j[th] quality variable at the s^{th} stage (s=1,2,…,N). In these equalities, $g(\cdot)$ is a quality statistic and L_{js} is the specification limit. As an example in a MWMP, the thickness of wafers should be within certain specifications in the lapping process. The third type of constraints describes the controllability of the process variables, i.e., the feasible range of control variables. These constraints form a group of inequalities $u_{il}^{L} < u_{il} < u_{il}^{U}$ for continuous variables, or a subset of possible values $x_{il} \in \{x_{il}\}$ for the discrete variables. As an example in a MWMP, the locations of wafers in a LPCVD chamber should be within a limited region.

2.2.4 Determination of the Control Actions

At each stage, we determine the control actions by minimizing the objective function of the predicted final quality variables. The decision variables can be continuous, or discrete, or a mixture of them. The multiple types of decision variables make the optimization problem difficult to solve. Here solutions are provided for three scenarios.

If the decision variables are continuous variables, we determine the optimal solutions by setting the partial derivative $\frac{\partial J(\mathbf{U})}{\partial \mathbf{U}} = 0$ for the quadratic objective function. If the solutions do not violate the constraints, then we find the optimal solution; otherwise the algorithm searches on the boundaries of constraints (Pierre, 1986).

If the decision variables are discrete variables, we treat the optimization problem as a combinatorial optimization problem. We use the Iterated Local Search (ILS) algorithm (Stutzle, 1998) to solve this optimization problem.

Table 2.1: Iterated local search procedure with both $\mathbf{U}_k$ and $\mathbf{X}_k$

Step1. Generate an initial feasible solution $\mathbf{u}_l^0$, $\mathbf{x}_l^0$ $(l = \mathrm{k}, \mathrm{k}+1, \cdots, \mathrm{N})$
Step2. Local search x_{ik} for every i to optimize the objective function until no more improvement is obtained, denote the local optimal solution as $\mathbf{x}^{loc,opt}$
Step3. Substitute $\mathbf{x}^{loc,opt}$, set the partial derivative $\frac{\partial J(\mathbf{U})}{\partial \mathbf{U}} = 0$ and search the constraints for the constrained local optimal solution $\mathbf{u}^{loc,opt}$
Step4. Modify the local optimal solution $\mathbf{x}^{loc,opt} \rightarrow \mathbf{x}^a$ by interchanging certain percentage of the solutions randomly
Step5. Local search x_{ik} for every i to optimize the objective function until no more improvement is obtained, denote the local optimal solution as $\mathbf{x}^{a\,loc,opt}$.
Step6. Substitute $\mathbf{x}^{a\,loc,opt}$, set the partial derivative $\frac{\partial J(\mathbf{U})}{\partial \mathbf{U}} = 0$ and search the constraints for the constrained local optimal solution $\mathbf{u}^{a\,loc,opt}$
Step7. If $\mathbf{u}^{a\,loc,opt}$ and $\mathbf{x}^{a\,loc,opt}$ has better objective value, then accept $\mathbf{u}^{a\,loc,opt}$ and $\mathbf{x}^{a\,loc,opt}$, i.e., $\mathbf{x}^{loc,opt} = \mathbf{x}^{a\,loc,opt}$, $\mathbf{u}^{loc,opt} = \mathbf{u}^{a\,loc,opt}$
Step8. If the termination conditions are met, terminate the optimization procedure; otherwise, go to Step 4

If the decision variables contain both continuous and discrete variables, we solve the optimization problem in an ILS framework, shown in Table 2.1. In this framework, Step 2 and Step 5 are "local search", which find a local optimal solution from an initial solution. Step 4 is "perturbation", which generates a new initial solution in iterations and prevents a solution trapped in a local optimal solution. A usual way of perturbation interchanges certain values of decision variables. In general, a higher percentage of interchange results in easier escape from the current local optimal solution, but may take longer time in finding a new local optimal solution, vice versa. The optimization process terminates when the optimal value is not improved for certain number of iterations. In the literature (Stutzle, 1998; Intellektik *et al.*, 1999; Lourenco *et al.*, 2002), one discusses

the choice of local search algorithm, interchange approach, and termination conditions, which will not be repeated in this chapter.

2.2.5 The Control Performance with Quality Sensing Noise

In the intermediate feedforward control strategy, the online quality measurements are the key factors in updating the control actions for downstream stages. However, the sensing noises may contaminate the online quality measurements. The "optimized" the control actions with quality sensing noise may not be the true optimal ones in variation reduction. Therefore, it is important to understand the impact of the sensing noise of quality variables on the control performance, thus to infer if the sensing noise is negligible or not.

In the quality and process relationship, we predict the final quality variables $\mathbf{Y}(\mathrm{N})_{\mathrm{j}}$ using Equation (2.2) stage-by-stage. Without loss of generality, the model to predict the final quality variable $\mathbf{Y}(\mathrm{N})_{\mathrm{j}}$ can be re-written as :

$$\begin{aligned}\mathbf{Y}(\mathrm{N})_{\mathrm{j}} &= \boldsymbol{\Phi}_{\mathrm{N}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{N,N}}^{\mathrm{j,\Phi}}\boldsymbol{\Phi}_{\mathrm{N}} + \boldsymbol{\psi}_{\mathrm{N}}^{\mathrm{T}}(\boldsymbol{\beta}_{\mathrm{N,N}}^{\mathrm{j,\psi,U}})\mathbf{U}_{\mathrm{N}} + \boldsymbol{\psi}_{\mathrm{N}}^{\mathrm{T}}(\boldsymbol{\beta}_{\mathrm{N,N}}^{\mathrm{j,\psi,X}})\mathbf{X}_{\mathrm{N}} + \varepsilon_{\mathrm{Nj}} \\ &= \beta_{0,\mathrm{N}} + \boldsymbol{\beta}_{\mathbf{Y}}^{\mathrm{T}}\mathbf{Y}^{\mathrm{N-1}} + \boldsymbol{\beta}_{\mathbf{U}}^{\mathrm{T}}\mathbf{U}^{\mathrm{N}} + \boldsymbol{\beta}_{\mathbf{X}}^{\mathrm{T}}\mathbf{X}^{\mathrm{N}} + \boldsymbol{\beta}_{\mathbf{M}}^{\mathrm{T}}\mathbf{M} + (\mathbf{Y}^{\mathrm{N-1}})^{\mathrm{T}}\mathbf{B}_{1}\mathbf{U}^{\mathrm{N}} + \\ &\quad (\mathbf{Y}^{\mathrm{N-1}})^{\mathrm{T}}\mathbf{B}_{2}\mathbf{X}^{\mathrm{N}} + (\mathbf{Y}^{\mathrm{N-1}})^{\mathrm{T}}\mathbf{B}_{3}\mathbf{M} + \mathbf{M}^{\mathrm{T}}\mathbf{C}_{1}\mathbf{U}^{\mathrm{N}} + \mathbf{M}^{\mathrm{T}}\mathbf{C}_{2}\mathbf{X}^{\mathrm{N}} + \varepsilon_{\mathrm{Nj}} \qquad (2.4)\end{aligned}$$

where $\mathbf{Y}^{\mathrm{N-1}}$, $\mathbf{U}^{\mathrm{N}}$ and $\mathbf{X}^{\mathrm{N}}$ are denoted as the same as those in Equation (2.2); and $\beta_{0,\mathrm{N}}$, $\boldsymbol{\beta}_{\mathbf{Y}}$, $\boldsymbol{\beta}_{\mathbf{U}}$, $\boldsymbol{\beta}_{\mathbf{X}}$, $\boldsymbol{\beta}_{\mathbf{M}}$, $\mathbf{B}_{i}$ ($i = 1,2,3$) and $\mathbf{C}_{i}$ ($i = 1,2$) are corresponding parameters with proper dimensions. We further assume the interaction terms of $(\mathbf{Y}^{\mathrm{N-1}})^{\mathrm{T}}\mathbf{Y}^{\mathrm{N-1}}$, $\mathbf{U}^{\mathrm{N^T}}\mathbf{U}^{\mathrm{N}}$, $\mathbf{X}^{\mathrm{N^T}}\mathbf{X}^{\mathrm{N}}$, $\mathbf{U}^{\mathrm{N^T}}\mathbf{X}^{\mathrm{N}}$ and $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ are insignificant in Equation (2.4). This assumption is

based on the fact that the interaction of the quality variables and process or material variables are more important in variation reduction through feedforward control in the MMP, than other interaction terms. The assumption is also verified by data driven variable selection in the case study.

To analyze the impact of sensing noise to the control objective, we further make two assumptions:

- The online observers provide unbiased sensing noise in the intermediate quality measurements, denoted as $\hat{\mathbf{Y}}^{N-1} = \mathbf{Y}^{N-1} + \tilde{\mathbf{Y}}^{N-1}$, where $\mathbf{Y}^{N-1}$ is the true value, $\hat{\mathbf{Y}}^{N-1}$ is the observed value and $\tilde{\mathbf{Y}}^{N-1}$ is the sensing noise. Here $E(\tilde{\mathbf{Y}}^{N-1} \mid \hat{\mathbf{Y}}^{N-1}) = 0$ and $\text{cov}(\tilde{\mathbf{Y}}^{N-1} \mid \hat{\mathbf{Y}}^{N-1}) = \Sigma_{\tilde{\mathbf{Y}}^{N-1}}$. In addition, the sensing noises of the quality variables are independent, i.e.,

$$\Sigma_{\tilde{\mathbf{Y}}^{N-1}} = diag(\sigma^2_{\tilde{\mathbf{Y}}(0)_1}, \sigma^2_{\tilde{\mathbf{Y}}(0)_2}, \cdots, \sigma^2_{\tilde{\mathbf{Y}}(1)_1}, \sigma^2_{\tilde{\mathbf{Y}}(1)_2}, \cdots, \sigma^2_{\tilde{\mathbf{Y}}(N-1)_1}, \sigma^2_{\tilde{\mathbf{Y}}(N-1)_2}, \cdots, \sigma^2_{\tilde{\mathbf{Y}}(N-1)_{m_{N-1}}}).$$

- The sensing noise $\tilde{\mathbf{Y}}^{N-1}$ is independent of ε_{Nj}.

Thus, based on Equation (2.4), the control objective for the j^{th} quality variable is a summation of a bias term and a variance term considering the sensing noise as follows:

$$\begin{aligned} E(\mathbf{Y}(N)_j^2) &= (E_{\varepsilon,\mathbf{Y}}(\mathbf{Y}(N)_j \mid \hat{\mathbf{Y}})^2 + \text{Var}_{\varepsilon,\mathbf{Y}}(\mathbf{Y}(N)_j \mid \hat{\mathbf{Y}}) \\ &= [\beta_{0,N} + \boldsymbol{\beta}_{\mathbf{Y}}^T \hat{\mathbf{Y}}^{N-1} + \boldsymbol{\beta}_{\mathbf{U}}^T \mathbf{U}^N + \boldsymbol{\beta}_{\mathbf{X}}^T \mathbf{X}^N + \boldsymbol{\beta}_{\mathbf{M}}^T \mathbf{M} + (\hat{\mathbf{Y}}^{N-1})^T \mathbf{B}_1 \mathbf{U}^N + \\ &\quad (\hat{\mathbf{Y}}^{N-1})^T \mathbf{B}_2 \mathbf{X}^N + (\hat{\mathbf{Y}}^{N-1})^T \mathbf{B}_3 \mathbf{M} + \mathbf{M}^T \mathbf{C}_1 \mathbf{U}^N + \mathbf{M}^T \mathbf{C}_2 \mathbf{X}^N]^2 + \sigma^2_{\varepsilon_{Nj}} + \\ &\quad (\boldsymbol{\beta}_{\mathbf{Y}} + \mathbf{B}_1 \mathbf{U}^N + \mathbf{B}_2 \mathbf{X}^N + \mathbf{B}_3 \mathbf{M})^T \Sigma_{\tilde{\mathbf{Y}}^{N-1}} (\boldsymbol{\beta}_{\mathbf{Y}} + \mathbf{B}_1 \mathbf{U}^N + \mathbf{B}_2 \mathbf{X}^N + \mathbf{B}_3 \mathbf{M}) \end{aligned} \quad (2.5)$$

In the control action determination, Equation (2.5) is treated as the optimization objective function. When the quality measurement has sensing noise, i.e., $\Sigma_{\tilde{\mathbf{Y}}^{N-1}}$ is a non-zero matrix, the optimized control actions using the objective control function in Equation (2.5) may become different as those without sensing noise. The impact of the sensing noise of quality variables to the control objective in the optimization problem consists of two parts: (1) the direct impacts of the sensing noise on the current control actions, through term $(\boldsymbol{\beta}_{\mathbf{Y}} + \mathbf{B}_1\mathbf{U}^{N} + \mathbf{B}_2\mathbf{X}^{N} + \mathbf{B}_3\mathbf{M})^{T}\Sigma_{\tilde{\mathbf{Y}}^{N-1}}(\boldsymbol{\beta}_{\mathbf{Y}} + \mathbf{B}_1\mathbf{U}^{N} + \mathbf{B}_2\mathbf{X}^{N} + \mathbf{B}_3\mathbf{M})$ in Equation (2.5); and (2) the indirect impacts of the sensing noise on the implemented control actions in upstream stages.

In the optimization at the k^{th} stage, $\Sigma_{\tilde{\mathbf{Y}}^{N-1}}$ has a direct impact on the control objective. The quality variables $\hat{\mathbf{Y}}^{N-1}$ are decomposed as the measured ones $\hat{\mathbf{Y}}^{k-1}$ and the predicted ones $\hat{\hat{\mathbf{Y}}}^{k,N-1}$, where $\hat{\mathbf{Y}}^{k-1} = \left[\hat{\mathbf{Y}}(0)^{T} \quad \hat{\mathbf{Y}}(1)^{T} \quad \cdots \quad \hat{\mathbf{Y}}(k-1)^{T}\right]^{T}$ are measured quality variables up to the $(k\text{-}1)^{th}$ stage and $\hat{\hat{\mathbf{Y}}}^{k,N-1} = \left[\hat{\hat{\mathbf{Y}}}(k)^{T} \quad \hat{\hat{\mathbf{Y}}}(k+1)^{T} \quad \cdots \quad \hat{\hat{\mathbf{Y}}}(N-1)^{T}\right]^{T}$ are the predicted quality variables substituted in the prediction models to predict the final quality variables. Therefore, $\Sigma_{\tilde{\mathbf{Y}}^{N-1}}$ represents both the sensing noise of $\hat{\mathbf{Y}}^{k-1}$ and the uncertainty of $\hat{\hat{\mathbf{Y}}}^{k,N-1}$. Here the uncertainty of $\hat{\hat{\mathbf{Y}}}^{k,N-1}$ is contributed by the prediction errors and the sensing noise of $\hat{\mathbf{Y}}^{k-1}$.

Beside the direct impact, the sensing noise of the quality variables also has indirect impacts on the control objective through the implemented control actions. Here we decompose the control actions $\mathbf{U}^{N} = \left[(\mathbf{U}^{k\text{-}1})^{T} \quad (\mathbf{U}^{k,N})^{T}\right]^{T}$ and

$\mathbf{X}^N = \left[(\mathbf{X}^{k-1})^T \quad (\mathbf{X}^{k,N})^T\right]^T$, where $\mathbf{U}^{k-1}$ and X^{k-1} are implemented control actions before the k^{th} stage; and $\mathbf{U}^{k,N} = \left[\mathbf{U}_k^T \quad \mathbf{U}_{k+1}^T \quad \cdots \quad \mathbf{U}_N^T\right]^T$ and $\mathbf{X}^{k,N} = \left[\mathbf{X}_k^T \quad \mathbf{X}_{k+1}^T \quad \cdots \quad \mathbf{X}_N^T\right]^T$ are the unimplemented control actions, i.e., the decision variables in the control optimization of the k^{th} stage. Because of the sensing noise before the k^{th} stage, the implemented control actions $\mathbf{U}^{k-1}$ and $\mathbf{X}^{k-1}$ may not be optimal to minimize the control objectives. The deviation of the control actions will further impact on the control optimization at the k^{th} stage.

By combining the effects from both the direct and indirect impacts of the sensing noise, the final implemented control actions may be different from the optimal ones without sensing noise, thus to degrade the variation reduction performance. Depending on the stage and magnitude of sensing noise, we expect different kinds of impact on the control objective. We study the impact of sensing noise in the case study.

2.3 Case Study

To show the performance of the intermediate feedforward control strategy, we conduct a case study in a MWMP with five major stages shown in Figure 2.4.

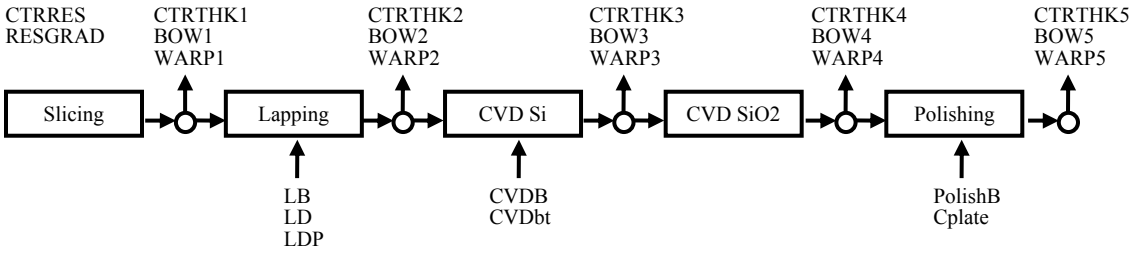


Figure 2.4: The layout of MWMP in the case study

In Figure 2.4, the slicing stage represents the slicing and cleaning process; the lapping stage represents lapping process, chemical etching and cleaning process; and

CVD polysilicon stage and CVD silicon dioxide stage consist of CVD process and their cleaning processes.

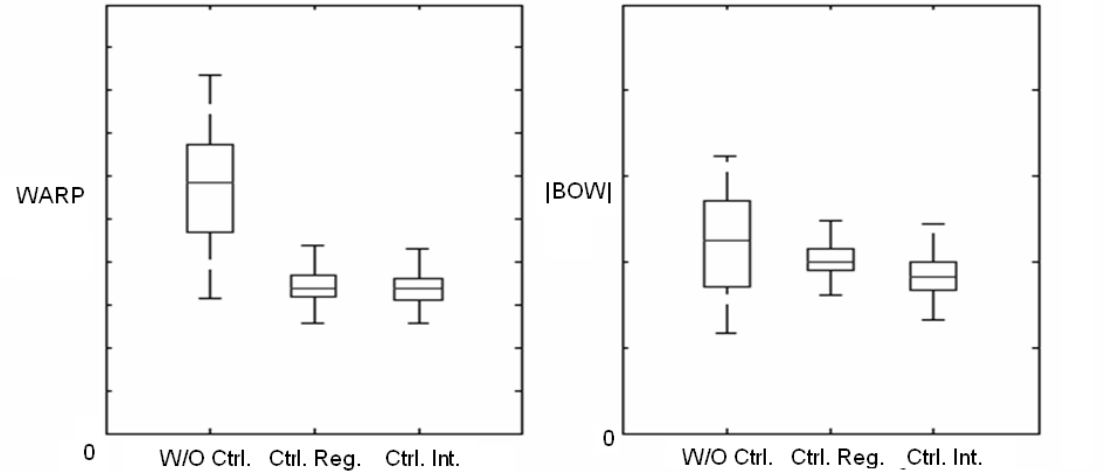
Table 2.2: Measured variables in a MWMP

Variable Type	Variable Name	Discrete / Continuous	Measured Stage	Physical Meanings
Process Variables	LB	Discrete	Lapping	Lapping batch, representing processing time with 15 levels
	LD	Discrete	Lapping	Lapping disk, representing pulley discs with 5 levels
	LDP	Discrete	Lapping	Positions in lapping with 6 levels
	CVDB	Discrete	CVD Si	CVD batch, representing different tubes with 5 levels
	CVDbt	Discrete	CVD Si	CVD boat, representing wafers' position in CVD tube with 4 levels
	PolishB	Discrete	Polishing	Polishing batch, representing age of slurry and pad with 12 levels
	Cplate	Discrete	Polishing	Ceramic plate, representing the alignment of ceramic plate holders with 4 levels
Material property Variables	CTRRES	Continuous	Na	Central resistivity of wafers
	RESGRAD	Continuous	Na	Resistivity gradient of wafers
Quality Variables	BOW	Continuous	All	Local warp at the center of a wafer
	WARP	Continuous	All	Maximum local warp
	CTRTHK	Continuous	All	Central thickness of wafer

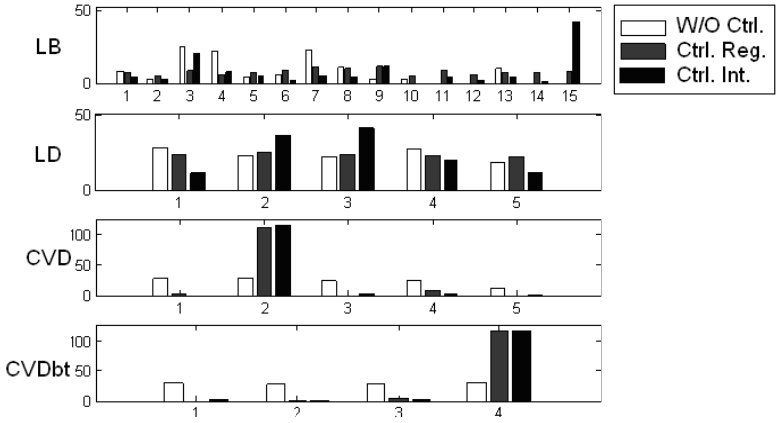
In this MWMP, we measured three types of variables to describe the manufacturing process, including quality variables, discrete offline setting variables, and material property variables. We summarize the detail definitions of these variables in Table 2.2. The number in the name of a quality variable represents the stage where it is measured, from Stage 1 to Stage 5. In the production, the objective is to minimize the magnitude of WARP5 and BOW5 after polishing stage. In the case study, we collect a total of 373 wafers of the observational data from a real production.

We split these wafers into training data set (250 wafers) and testing data set (123 wafers). Following the procedure in Section 2.2, we first construct the regression model

group based on the training data set, and then evaluate the control performance on the testing data set.



(a) WARP5 (b) |BOW5|



(c) Uncontrolled vs. controlled process variables

Figure 2.5: Uncontrolled vs. controlled final quality and process variables

In this process, the quality variables do not have Markov property. For example, an important factor to change BOW5 is the stress of polysillicon thin film on one face of the wafers, which is changed by both the CVD polysillicon and the polishing process, i.e., the BOW5 is correlated with the BOW3 given the BOW4. Many MMPs without Markov property may use the proposed modeling method to obtain an adequate

prediction model. By using the regression models in Equation (2.2), we identify four significant controllable variables: LB, LD, CVDB and CVDbt to predict both WARP5 and BOW5.

By using these controllable variables, we further determine the optimal settings of these variables in the control optimization. Without loss of generality, we set $C_j = 1$ in the control objective function in Equation (2.3). The intermediate quality specifications are the ranges of the wafer thickness. The controllability of the four significant controllable variables is the possible settings of these variables.

After implementing the intermediate feedforward control strategy, we compare the WARP5 and BOW5 in three scenarios as shown in Figure 2.5: (1) the quality performance without control ("W/O Ctrl."); (2) the quality performance using feedforward control based on a single regression model ("Ctrl. Reg."); and (3) the quality performance using the intermediate feedforward control based on the regression model group ("Ctrl. Int."). In Scenario (2), we solve a similar control optimization problem as Scenario (3) based on a single regression model.

Figure 2.5 (a) and (b) show the box plots of WARP5 and BOW5 for the testing wafers. In this figure, the controlled quality variables have smaller mean and variance than the uncontrolled quality variables. The proposed method ("Ctrl. Int.") has similar control performance in WARP5, but provides better control performance with smaller mean of BOW5, comparing to the feedforward control based on a single regression model ("Ctrl. Reg."). Figure 2.5 (c) shows the histograms of the final implemented process variables. The horizontal axis represents different settings; the vertical axis represents the counts of wafers set to certain value; and the left bar, the middle bar and

the right bar represent the counts for "W/O Ctrl.", "Ctrl. Reg." and "Ctrl. Int.", respectively. It is clear that some of the controlled process variables concentrate to a subset of settings to achieve better final quality, such as LB=15, CVDB=2 and CVDbt=4. The result indicates that these settings are better than others to reduce final variation of WARP5 and BOW5.

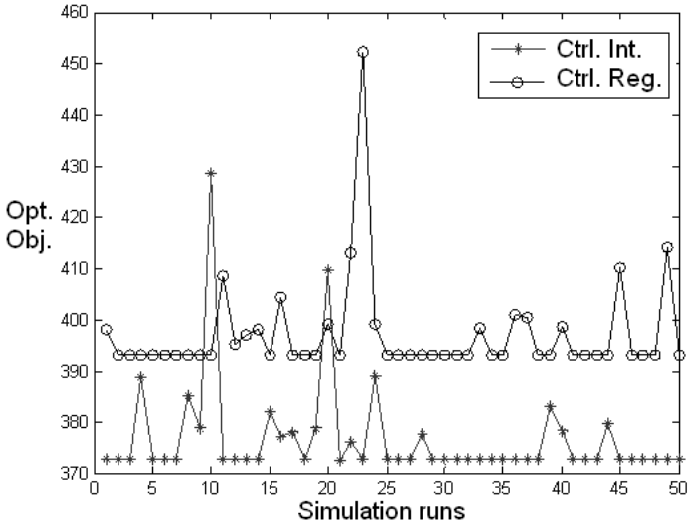


Figure 2.6: Control objective values of 50 simulation runs

To further evaluate the two types of controlled performance, we conduct 50 runs of control simulations, shown in Figure 2.6. In Figure 2.6, both control strategies yield better quality performance with smaller control objectives than the uncontrolled objective value 1027.54. To evaluate the control improvement, the Quality Index (*QI*) is defined as

$$QI = \frac{J^{w/octrl} - J^{w/ctrl}}{J^{w/octrl}} \times 100\% \tag{2.6}$$

where $J^{w/octrl}$ is the uncontrolled objective value; and $J^{w/ctrl}$ is the mean of the controlled optimal objective values. In this case study, *QI* is 63.34% for the intermediate feedforward control, which is a significant reduction of the objective value. Moreover,

the proposed intermediate feedforward control (stars) has better control performance than the control based on a regression model (circles) in most of the simulation runs. A few spikes show in the optimal objectives, since the ILS algorithm may not obtain global optimal in limited iterations.

Table 2.3: Optimal control objectives under different sensing noise scenarios

Scenarios	$\mathbf{Y}(1)$	$\mathbf{Y}(2)$	$\mathbf{Y}(3)$	$\mathbf{Y}(4)$	$\mathbf{Y}(5)$	Min. Opt. Ctrl. Obj.	Ctrl. Obj
1. No Noise	0	0	0	0	0	372.72	381.08 ± 11.23
2. Slicing	$\sigma^2_{\tilde{\mathbf{Y}}}$	0	0	0	0	379.00	394.52 ± 29.00
3. Lapping	0	$\sigma^2_{\tilde{\mathbf{Y}}}$	0	0	0	372.44	386.50 ± 28.14
4. Slicing & Lapping	$\sigma^2_{\tilde{\mathbf{Y}}}$	$\sigma^2_{\tilde{\mathbf{Y}}}$	0	0	0	379.00	394.52 ± 29.00
5. All Stages	$\sigma^2_{\tilde{\mathbf{Y}}}$	$\sigma^2_{\tilde{\mathbf{Y}}}$	$\sigma^2_{\tilde{\mathbf{Y}}}$	$\sigma^2_{\tilde{\mathbf{Y}}}$	$\sigma^2_{\tilde{\mathbf{Y}}}$	378.42	392.36 ± 32.30

In addition to the control performance comparison, we further analyze the impact of the sensing noise to the control optimization objectives. We show that (1) the sensing noise with same variance but from different stages may impact the final quality performance differently, and (2) quality performance becomes worse as the variance of the sensing noise increases. We use the result to conclude that the sensing noise in the case study is negligible to control objective.

In the simulation, we assume the distribution of sensing noise for each quality variables at the same stage follows the same distribution, and the maximum variance of the sensing noise is the largest modeling error $\sigma^2_{\max}$ from the regression model group used for prediction, which is $\sigma^2_{\tilde{\mathbf{Y}}} = \sigma^2_{\max} = 3.13\,\mu\text{m}^2$.

To evaluate the sensing noise from different stages, we assume the sensors at certain stages have the same noise distribution following $N(0, \sigma^2_{\tilde{\mathbf{Y}}})$, classified into five scenarios: without sensing noise, sensing noise at slicing stage, sensing noise at lapping

stage, sensing noise at both slicing and lapping stages, and sensing noise at all stages. In the case study, there are no further control actions to adjust after the lapping stage, therefore, it is not necessary to simulate the impact of sensing noise at downstream stages of lapping stage. For each scenario, we conduct 50 simulation runs and summarize the result in Table 2.3.

When there is no sensing noise (Scenario 1), the minimal control objective in these 50 simulation runs is 372.72, and the mean and standard deviation is 381.08 ± 11.23. By comparing the control objectives in different scenarios, we find that (1) the sensing noise at the slicing stage (Scenario 2) changes the control objective, and the sensing noise at the lapping stage (Scenario 3) does not influence the control objective significantly. Therefore, the sensing noise at both slicing and lapping stages (Scenario 4) has similar contribution to the control objective as that in Scenario 2. (2) With sensing noise (Scenario 2~5), the standard deviations of control objectives increase, which indicates the ILS algorithm need more iterations to obtain the global optimal solutions. And (3) the sensing noise after the last controllable stage (lapping stage) will have no impact on the control performance.

Based on the simulation result, the sensing noise of the intermediate quality measurements from lapping to polishing has less impact on the control objectives. However, it does not indicate that these online quality measurements are insignificant. This is because the simulation is conducted by assuming the maximal sensing noise is bounded by the largest prediction error of the regression models. When the sensing noise is very large or the prediction performance is inferior, the control objective will be significantly changed.

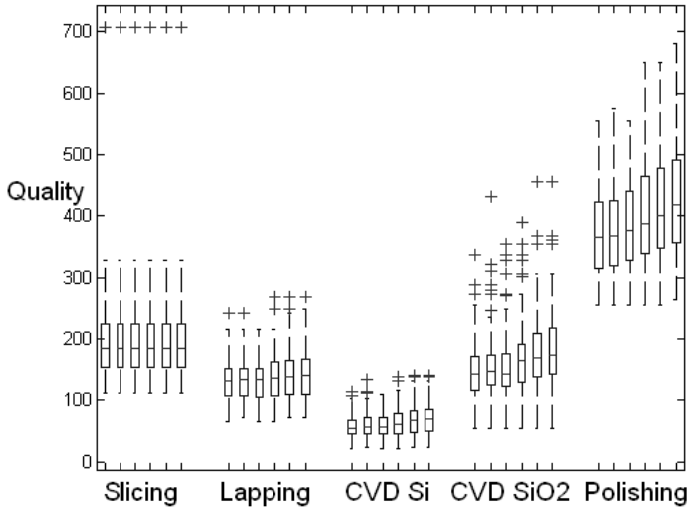


Figure 2.7: Optimized quality performance with sensing noise by stages

To evaluate the impact of sensing noise variances, we assume the sensing noises have the same variance for all intermediate and final quality variables, with result shown in Figure 2.7 and Figure 2.8. In Figure 2.7, we compare the quality performance with different variance of sensing noise stage-by-stage in one simulation run. The quality performance (vertical axis) is the summation of squared WARP and squared BOW of each wafer. We group the box plots of the wafer quality by stages. Within each stage, there are six box plots representing the wafer quality without sensing noise, and with sensing noise standard deviation as $0.2\sigma_{\max}$, $0.4\sigma_{\max}$, $0.6\sigma_{\max}$, $0.8\sigma_{\max}$ and $\sigma_{\max}$, from the left to the right. In Figure 2.7, because the sensing noise may exist in quality measurements, the optimized control actions may be different when the sensing noise variance varies. Thus, the quality with sensing noise at each stage becomes worse comparing to the one without sensing noise (the leftmost one within each stage). Moreover, the variance of quality performance is increasing as the variance of sensing noise becomes larger.

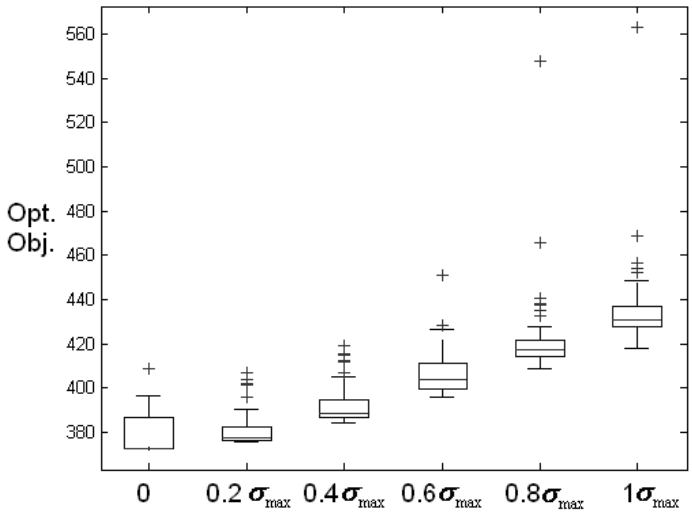


Figure 2.8: Optimal control objectives with different variance of sensing noise

In addition to impact of sensing noise at each stage, it is important to understand the relationship of sensing noise and final quality performance to evaluate if sensing noise is negligible. Figure 2.8 shows such a relationship. In Figure 2.8, the horizontal axis represents the sensing noise, from standard deviation 0 to $\sigma_{\max}$. The vertical axis represents the optimal objective values in the box plots. Each box plot accounts for the optimized control objectives of 50 simulation runs. It is clear that when the sensing noise is small, such as that the standard deviation is smaller than $0.2\sigma_{\max} = 0.354\mu$m in this case study, the optimal control objective is comparable to the one without sensing noise. When the sensing noise becomes larger, the final quality performance becomes worse. In the case study, the sensing noise to measure the quality variables is 0.1 μm based on the gauge precision. Therefore, the sensing noise is negligible and the final optimized control actions will not degrade the control performance in this case study.

2.4 Conclusion

It is very important to reduce variation of wafer quality in a MWMP, because the wafer quality variables are important quality specifications for downstream process, such as solar cell manufacturing, or integrated circuit fabrication. Large variation would result in huge lost in both energy and materials. It is also a challenge task for variation reduction. The challenges not only lie in the complexity of a typical MWMP, but also the variation propagation among different stages.

In this chapter, we first propose a group of regression models to model the variation propagation of quality variables based on observational production data. Then, we develop an intermediate feedforward control strategy by solving a sequence of constrained optimization problems. In the control strategy, we use the intermediate quality measurements to update the control actions. The modeling and control procedure is demonstrated in a typical MWMP to improve BOW and WARP. By implementing the proposed method, the quality of wafers is significantly improved by 63.34%. Moreover, we use simulation to study the impact of the sensing noise of quality variables to the control objective, from different stages and of different variances. The sensing noise is negligible to the control objective in the case study.

In the future research, we will improve the control performance by developing models with higher prediction accuracy, such as advanced statistical model from data mining. The engineering knowledge within each stage will be used to construct better model for control.

CHAPTER 3

RECONFIGURED PIECEWISE LINEAR REGRESSION TREE FOR MULTISTAGE MANUFACTURING PROCESS CONTROL

3.1 Introduction

A multistage manufacturing process (MMP) refers to a manufacturing system consisting of multiple units, stations, or operations to finish a final product. In most cases, the final product quality of a MMP is determined by complex interactions among multiple stages. The quality characteristics of one stage are not only influenced by the local variations at that stage but also by the propagated variations from upstream stages. A MMP presents significant challenges, as well as opportunities, for quality engineering research. Two of the common challenges are how to model the variation and its propagations along the production stages, and how to further use the model to reduce the final product variation.

Various methodologies have been developed for modeling and control of system variability in MMPs. The feedforward control is one of the commonly adopted methodologies for such purposes. There are three typical feedforward control strategies reported in the literature based on the models used to represent a MMP.

One methodology is called Stream of Variation (SoV) based on a state space model (Jin and Shi, 1999; Shi, 2006). A SoV model is typically obtained from engineering knowledge, such as design information and physical laws of the process. Studies of feedforward control under the SoV framework includes the adjustment of the fixture position and the tool path in a machining process (Djurdjanovic and Zhu, 2005), and variation reduction in an assembly process when taking the controllability and

measurement noises into account (Izquierdo *et al*., 2007). In recent years, a new control strategy is developed based on a one-step ahead optimal criterion. The control actions are updated iteratively as the operations move on (Jiao and Djurdjanovic, 2010). The control performance of this type of approaches depends on the validity and accuracy of the state space model. The SoV based feedforward control may not be applicable (1) if the SoV model cannot be obtained based on the physics and engineering knowledge due to the system complexity; and (2) there are strong nonlinear relationships among process variables and quality variables in a complex MMP. In this situation, an effective data-driven modeling method is desirable to address nonlinear properties of the observational data.

Other methodologies are developed based on regression models, such as Robust Parameter Design (RPD) based feedforward control (Joseph, 2003) and DOE-based automatic process control (APC) (Jin and Ding, 2004). DOE-based APC determines the control actions by minimizing the predicted control objective function from a global regression model. The certainty equivalence control or cautious control strategies are employed in the APC context (Jin and Ding, 2004). Recently, Zhong *et al.,* (2010) has also investigated the impacts of model uncertainties and sensing errors on the control performances. The DOE-based APC approach yields better performance for variability reduction than the traditional RPD does. However, the DOE-based APC approach has two limitations: (1) the global regression model predicts the final quality variables when information at all stages are known. Thus, it cannot be used to control at an intermediate stage when only its upstream stage information is available; (2) The single regression

model strategy can not address complex situations in a MMP when the data structure is nonlinear.

With abundant observational data available in a modern MMP, there are timely information provided about the process variables, material properties, and intermediate quality measures. With the help of these data, data mining techniques can be used to model the interrelationships among those variables. The regression tree models are one of effective approaches to model nonlinear data structure with high prediction accuracy and explicit interpretation of predictors. Therefore, the regression tree models are adopted in this chapter to model the variation and its propagations in MMPs.

There are three typical methods to model a regression tree, which are greedy search, Bayesian tree, and statistical test. In general, the greedy search approaches are biased in splitting variable selection and computational intensive, such as AID algorithm (Morgan and Sonquist, 1963) and Classification and Regression Tree (CART) (Breiman *et al*., 1984). To improve the computation efficiency, Bayesian tree is developed by proposing the priors distributions for both tree structure and parameters (Chipman *et al.*, 1998, 2002; Dennison *et al.*, 2002). The MCMC method is used to determine the posterior distributions. Another type of approaches uses statistical tests to determine splitting variables, such as Smoothed and Unsmoothed Piecewise-polynomial Regression Trees (SUPPORT) (Chaudhuri *et al*., 1994) and Generalized, Unbiased Interaction Detection and Estimation (GUIDE) (Loh, 2002; Kim *et al*., 2007). In these approaches, the residuals of piecewise models are tested with better computational efficiency.

In this chapter, piecewise linear regression trees (PLRTs) estimated by GUIDE are adopted to model MMPs for process control. The reasons for selection of the PLRTs

from GUIDE are: (1) A PLRT from GUIDE has a better prediction accuracy for nonlinear data structure than a global regression model (Loh, 2002; Kim *et al.*, 2007; Loh *et al*., 2007). (2) The interpretation of the PLRT is explicit. The predictors in the tree structure are explained as important factors under different scenarios or splitting conditions. (3) GUIDE has several superior properties over other estimation methods. For example, both categorical and continuous predictors can be assigned to different roles, such as splitting only, regression only, or both. It also alleviates the selection bias and investigates the local pair-wise interactions. Therefore, it is an effective way to link the process, material property, and quality variables in MMPs.

A PLRT from GUIDE performs well for quality "*prediction"* in MMPs but not for "*variation reduction"*. There are two major limitations that prohibit using a PLRT directly in feedforward control for variation reduction: (1) In a MMP, the temporal orders are determined by the design of a manufacturing system. However, the splitting order in PLRTs is prioritized according to the data structure and nonlinear relationships. Therefore, the splitting order in PLRTs may not reveal the same temporal sequence of a MMP. Thus, it is not feasible to select the potential models for the prediction of the final product quality at an intermediate stage based on the data only available in the upstream stages, since the downstream variables may be needed to make the prediction. This limitation results in that a control or adjustment decision cannot be made at an intermediate stage to reduce process variation in a MMP. (2) A PLRT model is usually used to predict a single response. Examples of multiple responses can be found in Segal (1992), Larsen and Speckman (2004), and Lee (2006), but not in a nested structure, i.e., one response becomes a predictor to another response. In a variation reduction problem,

an intermediate quality variable may be a response as well as a predictor to the downstream process. In a typical MMP, multiple variables need to be predicted for quality control purposes. However, it is difficult to evaluate the splitting conditions from multiple trees, which limits the capability to make a control or adjustment decision to achieve optimal performance of multivariate responses.

This chapter develops a unified modeling and control methodology for MMP based on a reconfigured PLRT model. The engineering design knowledge is used to reconfigure the model to an engineering complied, yet statistical equivalent model for feedforward control purposes. Furthermore, the model complexity is reduced by merging the splitting structures while satisfying the specified control accuracy requirement. Finally, a control strategy with an intermediate variable adjustment based on this reconfigured PLRT is proposed to reduce the variation of quality variables at the final stage.

The rest of the chapter is organized as follows. In Section 3.2, we show the overview of proposed methodology in modeling and control. In Section 3.3, we propose the methodology for modeling and reconfiguration of PLRTs. Then we develop the method to reduce model complexity in Section 3.4. Based on the PLRT with reduced model complexity, we develop the feedforward control strategy in Section 3.5. We further use a multistage wafer manufacturing process (MWMP) to illustrate the procedure of modeling and control in Section 3.6. Finally, the conclusion is made in Section 3.7.

3.2 Overview of the Proposed Methodology in Modeling and Control

The proposed method to model and control a MMP with reconfigured PLRT is an engineering knowledge enhanced statistical method, as illustrated in Figure 3.1.

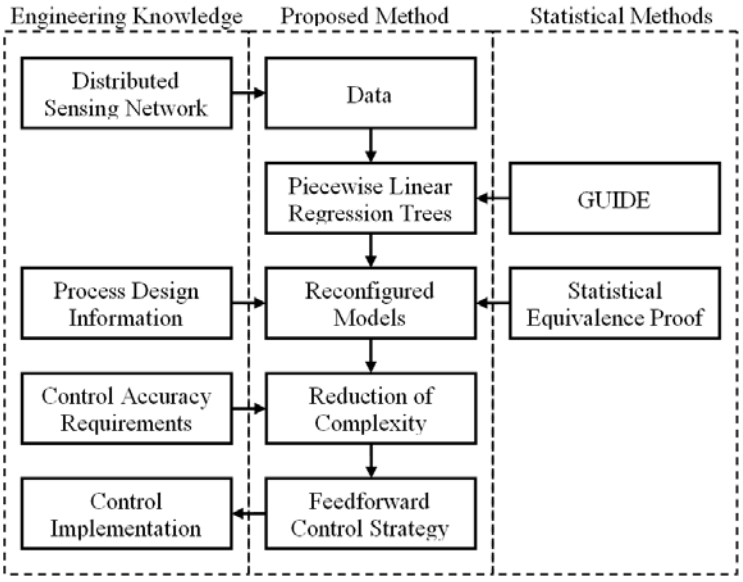


Figure 3.1: Overview of proposed methodology

In Figure 3.1, the observational data of the process, material property, and quality variables are measured from a MMP. Based on these data, PLRTs are estimated by using GUIDE to predict all intermediate and final quality variables. Then the tree models will be reconfigured to an engineering complied structure with a statistically equivalent property. Based on the final quality specifications of the MMP, the reconfigured PLRT model structure is further adjusted to find the simplest model that satisfies the accuracy requirements. In the reconfigured PLRT, a group of potential prediction models are used to predict the final product quality, as the multistage operations move from the upstream stages to the downstream stages. Therefore, a feedforward control strategy with intermediate process variable adjustment is used to take advantages of the temporally ordered layers in predicting quality variables. The control actions are iteratively determined by solving optimization problems with product and process constraints, which are conducted to improve the final product quality in the MMP.

3.3 Engineering-driven Reconfiguration of PLRTs

The engineering-driven reconfiguration ensures the feasibility of PLRTs in a feedforward control strategy. The advantage of PLRTs in prediction accuracy is also preserved in control because the reconfiguration does not re-estimate the local models.

3.3.1 Multistage Manufacturing Process Modeled by PLRTs

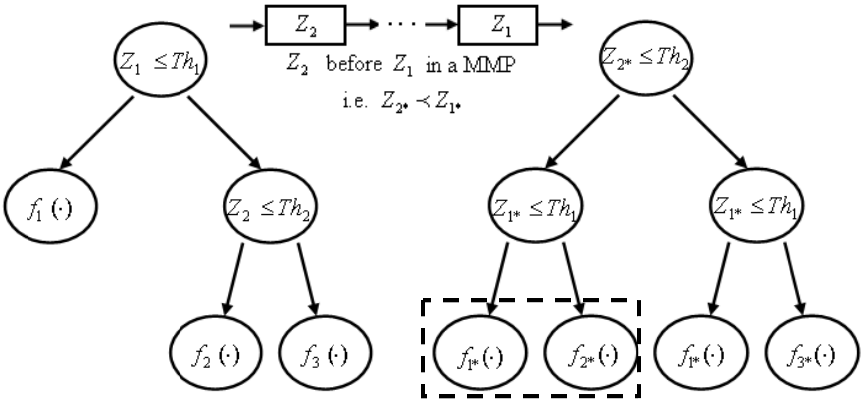
PLRTs model the nonlinear data by partition and local fitting. Figure 3.2 (a) shows an example of a PLRT estimated from GUIDE, which consists of three leaf nodes. In this tree structure, Z_i $(i = 1,2)$ are splitting variables; Th_i $(i = 1,2)$ are splitting boundaries; and f_i $(\cdot)$ $(i = 1,2,3)$ are local regression models. When the splitting condition holds, the tree goes to the left branch. The sample space of the PLRT is illustrated in Figure 3.2 (c), where f_i $(\cdot)$ $(i = 1,2,3)$ are marked in their corresponding sub-regions.

Table 3.1: Variable notations

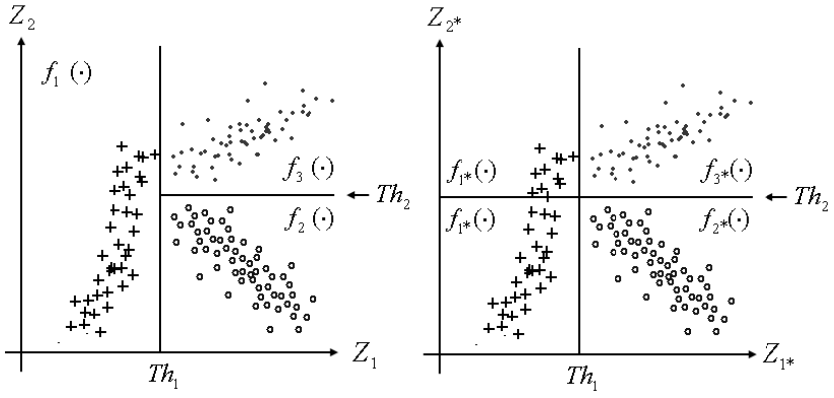
$\mathrm{Y(k)} \in \Re^{m_k \times 1}$	: Quality variables with noise at the k-th stage
$\mathbf{Y}(0)$	: Initial quality vector before entering the manufacturing process
$\mathbf{U}_k \in \Re^{r_k \times 1}$	: Continuous online controllable variables at the k-th stage
u_{lk}	: The l-th variable at the k-th stage, which can be adjusted during the operations at the k-th stage
$\mathbf{X}_k \in \Re^{n_k \times 1}$	: Offline setting variables at the k-th stage
x_{lk}	: The l-th variable at the k-th stage, which can be adjusted between the (k-1)-th stage and the k-th stage
$\mathbf{M} \in \Re^{t \times 1}$	: Material property variables independent of stages

In a typical layout of MMP shown in Figure 2.2 in Chapter 2, a stage is defined as a series of operations applied to a product to complete a manufacturing task. The intermediate quality variables are measured at each stage for modeling. A discrete part or a batch of products is processed. In this MMP, the variables can be classified as quality

variables, process variables, and material property variables. Based on the controllability and variable types, variables are further classified and summarized in Table 3.1.



(a) A PLRT (b) Re-ordered PLRT



(c) Sample space before re-ordering (d) Sample space after re-ordering

Figure 3.2: Re-ordered model from a PLRT at one stage

To model the variable relationship, a PLRT is adopted by conducting regression of the quality variables on their upstream variables. A general form of the model with T leaf nodes and L distinct splitting variables:

$$y = f(\boldsymbol{\eta}) = \sum_{i=1}^{T} f_i(\boldsymbol{\eta}_i) I(g_i(Z_1, ..., Z_L)) \tag{3.1}$$

In this model, y could be any quality variable at any stage; if y is a quality variable at the k-th stage, then $\boldsymbol{\eta} = \{\mathbf{Y}(0), \mathbf{Y}(\mathrm{k}_1), \mathbf{U}_{\mathrm{k}_2}, \mathbf{X}_{\mathrm{k}_2}, \mathbf{M}\}$ ($\mathrm{k}_1 = 1,2,...,\mathrm{k}-1$; $\mathrm{k}_2 = 1,2,...,\mathrm{k}$) represents the known information at the k-th stage; $f_i(\cdot)$ and $\boldsymbol{\eta}_i$ represents the local models and the covariates in the i-th leaf node; $I(\cdot)$ is an indicator function, which is 1 if $g_i(\cdot)$ is non-negative, or 0 otherwise; $g_i(\cdot)$ is the combination of conditions leading to the i-th leaf node; and $Z_1,...,Z_L$ are splitting variables for the tree structure. Furthermore, the $I(g_i(\cdot))$ can be decomposed as a product of the indicator functions of the individual splitting variables, i.e., $I(g_i(Z_1,...,Z_L)) = \prod_{k=1}^{L} I(g_{i,k}(Z_k))$, where $g_{i,k}(\cdot)$ is the splitting condition of the k-th variable for the i-th leaf node. For example, in Figure 3.2 (a), the splitting conditions leading to $f_2(\cdot)$ are $Z_1 > Th_1$ and $Z_2 \le Th_2$, which can be written as $I(g(Z_1, Z_2)) = I(Z_1 - Th_1)I(Th_2 - Z_2)$.

In the PLRT model estimation, there are three important issues to be addressed: splitting variable selection, splitting boundary estimation, and tree structure determination. In this chapter, we follow the procedures in GUIDE, which recursively partitions the sample space, selects the splitting variables by contingence table test, and determines the splitting boundaries by minimizing the prediction errors. When a large tree grows, the 10-fold cross validation error is minimized to prune the tree structure. There are comprehensive discussion on splitting variable selection, splitting boundary estimation and pruning in the literature (Loh, 2002; Kim *et al.*, 2007; Loh *et al.*, 2007), which will not be repeated in this chapter. This chapter uses those methods to estimate a

PLRT model from observational data. This estimated PLRT model will be used as a basis for later model reconfiguration and feedforward control design.

To explain the relationship of nodes in the tree structure, the *layer of nodes* in a tree is defined.

Definition 3.2.1 *The i-th layer of nodes*: The i-th layer of nodes in a tree is a set of nodes with depth i, i.e., the nodes which have (i-1) splits from the root of the tree, including leaf nodes and splitting nodes.

Definition 3.2.1 is illustrated with Figure 3.2 (a). There are three layers because the deepest leaf node from the node is reached by two splittings from the root of the tree: The splitting node of $Z_1 \le Th_1$ is the root node, which forms the first layer of the tree; Leaf node $f_1(\cdot)$ and splitting node of $Z_2 \le Th_2$ form the second layer of the tree; Leaf nodes $f_2(\cdot)$ and $f_3(\cdot)$ form the third layer of the tree.

3.3.2 Reconfiguration of PLRTs

The engineering knowledge of MMPs used for the reconfiguration is the temporal order and the inherent relationship among the variables, i.e., the quality at the current stage is only influenced by the upstream stages rather than the downstream stages. When there is insufficient Markov property of the quality variables, prediction by all upstream variables may also improve the prediction accuracy, comparing to the modeling by only regressing on the quality at last stage.

Assuming there are *L* splitting variables, these splitting variables belong to certain stages of the MMP with temporal order. This chapter uses notations "$\prec$", "$\sim$" or "$\prec\sim$" of variables marked by * in the superscript to describe the temporal order. Table 3.2

summarizes the temporal relationship of these variables, and Z_{i*} ($i = 1,2;$) is used for denoting Z_i in a temporal order. In MMPs, such a kind of temporal order of the quality and process variables at the (k-1)-th stage and the k-th stage can be presented as: $\mathbf{X}_{(k-1)*} \prec\sim \mathbf{U}_{(k-1)*} \prec \mathbf{Y}((k-1)*) \prec \mathbf{X}_{k*} \prec\sim \mathbf{U}_{k*} \prec \mathbf{Y}(k*)$.

Table 3.2: Notations of temporal orders

$Z_{1*} \prec Z_{2*}$: Z_1 is temporally prior to Z_2;
$Z_{1*} \sim Z_{2*}$: Z_1 and Z_2 have the same temporal order;
$Z_{1*} \prec\sim Z_{2*}$: Z_1 is temporally prior or the same as Z_2.

With the temporal order of the splitting variable, the original PLRT is re-ordered into a temporally complied tree, which is defined below for further analysis.

Definition 3.2.2 *Temporally complied tree:* A tree is temporally complied if the splitting variables in the tree is temporally ordered, which is defined by the MMP layout, i.e., if $Z_{i*} \prec\sim Z_{j*}$, then Z_{i*} is in a closer layer or the same layer as the root compared to the location of Z_{j*}.

The reconfigured PLRT should have three appealing properties for the feedforward control purpose: (1) the reconfigured PLRT should be a temporally complied tree; (2) several PLRTs are estimated to predict the intermediate and final quality, which should be combined into a single decision structure; and (3) the reconfigured PLRT should be statistical equivalent to the PLRT models with high prediction accuracy.

The reconfiguration of PLRTs consists of two steps: (1) each PLRT is reconfigured according to the temporal order of the splitting variables, called *re-ordering*; and (2) a group of PLRTs is combined as a reconfigured PLRT called *combining*.

3.3.2.1 Re-ordering

Assuming the splitting order in a PLRT is not consistent with the temporal order as $Z_{1^*} \prec\sim Z_{2^*} \prec\sim ... \prec\sim Z_{L^*}$, the procedure to re-order a PLRT is proposed in the Algorithm 1 in Table 3.3.

Table 3.3: The algorithm for the re-ordering

Algorithm 1.

Step 1. Convert the PLRT to a summation of $f_i(\cdot)$ and $g_i(\cdot)$ as Equation (3.1)

Step 2. Partition the region of $g_i(\cdot)$ w.r.t all splitting variables into the decomposed sub-regions

$g_i^j(\cdot)$ ($j = 1,...,D_i$), i.e., $y = \sum_{i=1}^{T} f_i(\boldsymbol{\eta}_i) I(g_i(Z_1,...,Z_L)) = \sum_{i=1}^{T}\sum_{j=1}^{D_i} f_i(\boldsymbol{\eta}_i) I(g_i^j(Z_1,...,Z_L))$

Step 3. Merge the sub-regions $g_i^j(\cdot)$ and $f_i(\boldsymbol{\eta}_i)$ for Z_i ($i = 1,...,L$) from Z_{L^*} to Z_{1^*}, if the Merge Condition I is satisfied The final re-ordered model $y^* = \sum_{i=1}^{T^*} f_{i^*}(\boldsymbol{\eta}_{i^*}) I(g_{i^*}(Z_{1^*},...,Z_{L^*}))$

Step 4. Formulate the layers into temporal complied tree based on the re-ordered model

In Algorithm 1, all splitting variables Z_i ($\forall i$) are considered in partitioning the regions in Step 2; $g_i^j(\cdot)$ are the decomposed sub-regions of $g_i(\cdot)$, where D_i is the total number of sub-regions considering all possible splits of Z_i ($\forall i$). In Step 3, if the Merge Condition I (defined below) is satisfied, the sub-regions will be merged; otherwise, no further merging is needed.

The Merge Condition I for Z_i in any two decomposed sub-regions j_1 and j_2 in leaf nodes i_1 and i_2 is: $g_{i_1,k}^{j_1}(Z_k) = g_{i_2,k}^{j_2}(Z_k)$ ($\forall k \neq i$) and $f_{i_1}(\boldsymbol{\eta}_{i_1})$ is the same model as $f_{i_2}(\boldsymbol{\eta}_{i_2})$. Here the splitting conditions of the decomposed regions are $I(g_{i_1,i}^{j_1}(Z_i)) \prod_{\forall k \neq i} I(g_{i_1,k}^{j_1}(Z_k))$ and $I(g_{i_2,i}^{j_2}(Z_i)) \prod_{\forall k \neq i} I(g_{i_2,k}^{j_2}(Z_k))$. $f_{i_1}(\boldsymbol{\eta}_{i_1})$ and $f_{i_2}(\boldsymbol{\eta}_{i_2})$ are the

associated local regression models. After the merging process, the splitting condition for the newly merged leaf node is $\prod_{\forall k \neq i} I(g_{i_1,k}^{j_1}(Z_k))$ (or $\prod_{\forall k \neq i} I(g_{i_2,k}^{j_2}(Z_k))$).

To illustrate the Merge Condition I, the tree in Figure 3.2 (a) is re-ordered as an example. Following the procedure of Algorithm 1, there will be four partitioned sub-regions as shown in Figure 3.2 (d) after Step 2. In Step 3, assuming $Z_{2^*} \succ Z_{1^*}$ Z_{1^*} should be merged first. Considering the merge in the dashed rectangular in Figure 3.2 (b), their splitting conditions are $I(Th_1 - Z_{1^*})I(Th_2 - Z_{2^*})$ and $I(Z_{1^*} - Th_1)I(Th_2 - Z_{2^*})$. In this example, $I(g_{1,2}^1(Z_{2^*})) = I(g_{2,2}^1(Z_{2^*})) = I(Th_2 - Z_{2^*})$, but $f_{1^*}(\cdot)$ and $f_{2^*}(\cdot)$ are not the same. Therefore, the Merge Condition I is not satisfied and these two leaf nodes cannot be merged. Once the re-ordered model is obtained, we can formulate Z_{2^*} in the first layer, then Z_{1^*} in the second layer.

Statement 3.1 *Statistical equivalence in re-ordering*: The original PLRT is statistically equivalent to the re-ordered temporally complied tree in prediction, i.e., $y = y^*$.

The proof of Statement 3.1 is in the Appendix. To illustrate the equivalence, Figure 3.2 (c) and (d) are compared. By given a new sample, the local prediction models $f_i(\cdot)$ $(i = 1,2,3)$ in Figure 3.2 (c) and $f_{i^*}(\cdot)$ $(i = 1,2,3)$ in Figure 3.2 (d) are identical, since the re-ordering does not re-estimate the local regression models.

3.3.2.2 Combining

After re-ordering, multiple PLRTs are combined as a single reconfigured tree to predict multiple quality variables. If there are N_1 re-ordered PLRTs, with T_n^* leaf nodes and L_n^*

splitting variables in the *n*-th tree ($n = 1,2,...,\mathrm{N}_1$), the general form of these re-ordered models are denoted as

$$y_n^* = \sum_{i=1}^{T_n^*} f_{i*}^n(\boldsymbol{\eta}_{i*}^n) I(g_{i*}^n(Z_{1*}^n,...,Z_{L_n^*}^n)) \tag{3.2}$$

where all notations are similarly denoted as Equation (3.1) except "*n*" for the *n*-th tree. Furthermore, $Z_{1*},...,Z_{L*}$ are the splitting variables in all these trees, with temporal order $Z_{1*} \prec\sim Z_{2*} \prec\sim ... \prec\sim Z_{L*}$. The procedure to combine the re-ordered models is proposed in the Algorithm 2 in Table 3.4.

Table 3.4: The algorithm for the combining

Algorithm 2.

Step 1. Obtain the re-ordered structure for the models in the form of Equation (3.2)

Step 2. Decompose the $g_i^n(\cdot)$ into $g_i^{n,j}(\cdot)$ using the same approach of Step 2 in Algorithm 1 considering all splitting variables in different PLRTs, i.e.,

$$y_n^* = \sum_{i=1}^{T_n^*} f_{i*}^n(\boldsymbol{\eta}_{i*}^n) I(g_{i*}^n(Z_{1*}^n,...,Z_{L_n^*}^n)) = \sum_{i=1}^{T_n^*} \sum_{j=1}^{D_{n,i}^*} f_{i*}^n(\boldsymbol{\eta}_{i*}^n) I(g_{i*}^{n,j}(Z_{1*},...,Z_{L^*}))$$

Step 3. Merge the decomposed sub-regions $g_i^{n,j}(\cdot)$ using the similar procedure of Step 3 in Algorithm 1 if the Merge Condition II is satisfied The final combined model is

$$y_n^* = \sum_{i=1}^{T^*} f_{i*}^n(\boldsymbol{\eta}_{i*}^n) I(g_{i*}^{comb}(Z_{1^*},...,Z_{L^*}))\ (n = 1,2,...,\mathrm{N}_1)$$

Step 4. Formulate the layers into temporal complied ones in the tree

In Algorithm 2, all splitting variables in these re-ordered trees are considered in the decomposition in Step 2. $D_{n,i}^*$ is the total number of decomposed sub-regions considering all possible splits of Z_{i*} ($\forall i$) from the *i*-th leaf node in the *n*-th tree. In Step 3, if the Merge Condition II is satisfied, the sub-regions will be merged, and a group of N_1 regression models for multiple responses is formed. Otherwise, no further merging is needed.

The Merge Condition II for Z_{i^*} in two decomposed sub-regions j_1 and j_2 in leaf nodes i_1 and i_2 is: $I(g_{i_1^*,k^*}^{n,j_1}(Z_{k^*}^n)) = I(g_{i_2^*,k^*}^{n,j_2}(Z_{k^*}^n))$, ($\forall k^* \neq i^*$) and $f_{i_1^*}^n(\boldsymbol{\eta}_{i_1^*}^n)$ is the same model as $f_{i_2^*}^n(\boldsymbol{\eta}_{i_2^*}^n)$. Here the splitting conditions of the two decomposed sub-regions are $I(g_{i_1^*,i^*}^{n,j_1}(Z_{i^*}^n)) \prod_{\forall k^* \neq i^*} I(g_{i_1^*,k^*}^{n,j_1}(Z_{k^*}^n))$ and $I(g_{i_2^*,i^*}^{n,j_2}(Z_{i^*}^n)) \prod_{\forall k^* \neq i^*} I(g_{i_2^*,k^*}^{n,j_2}(Z_{k^*}^n))$. $f_{i_1^*}^n(\boldsymbol{\eta}_{i_1^*}^n)$ and $f_{i_2^*}^n(\boldsymbol{\eta}_{i_2^*}^n)$ are the associated local models. After the merging process, the splitting condition for the newly merged leaf node is $\prod_{\forall k^* \neq i^*} I(g_{i_1^*,k^*}^{n,j_1}(Z_{k^*}^n))$ (or $\prod_{\forall k^* \neq i^*} I(g_{i_2^*,k^*}^{n,j_2}(Z_{k^*}^n))$).

To illustrate the Merge Condition II, two trees in Figure 3.2 (b) and Figure 3.3 are combined as an example, assuming $Th_1 < Th_3$. In this case, the local models $f_{i^*}(\cdot)$ ($i = 1,2,3$) in Figure 3.2 (b) becomes $f_{i^*}^1(\cdot)$ ($i = 1,2,3$) to distinguish the models in Figure 3.3. There are three distinct splitting variables in these trees: Z_{1^*}, Z_{2^*}, and Z_{3^*}. Following the procedure of Algorithm 2, all possible splits are generated in Step 2. In Step 3, assuming $Z_{2^*} \prec Z_{1^*} \prec Z_{3^*}$, Z_{3^*} should be merged first. Considering the merger of two leaf nodes that are marked by the dashed rectangular in Figure 3.4 (a), the splitting conditions are $I(Th_1 - Z_{1^*})I(Th_2 - Z_{2^*})I(Th_4 - Z_{3^*})$ and $I(Th_1 - Z_{1^*})I(Th_2 - Z_{2^*})I(Z_{3^*} - Th_4)$. In this example, $I(g_{1,1}^{n,1}(Z_{1^*})) = I(g_{2,1}^{n,1}(Z_{1^*})) = I(Th_1 - Z_{1^*})$ for n=1, 2, and $I(g_{1,2}^{1,1}(Z_{2^*})) = I(g_{2,2}^{1,1}(Z_{2^*})) = I(Th_2 - Z_{2^*})$. The local models are also identical. Therefore, the Merge Condition II is satisfied and these two leaf nodes should be merged, shown in Figure 3.4 (b).

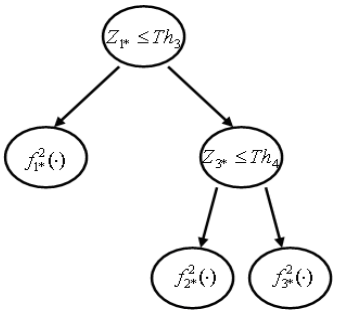
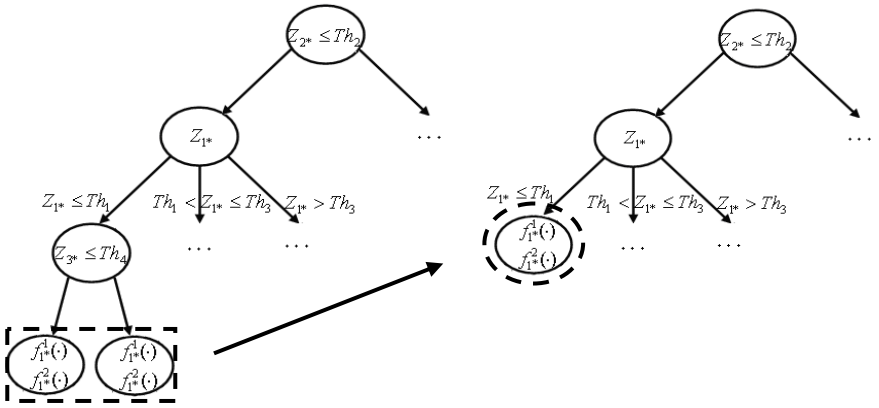


Figure 3.3: Another re-ordered PLRT



(a) Before merge (b) After merge

Figure 3.4: Merging leaf nodes in combining

Statement 3.2 *Statistical equivalence in combining*: A group of re-ordered models from PLRTs is combined into a single statistically equivalent model using Algorithm 2.

The proof of Statement 3.2 is shown in the Appendix. To illustrate the equivalence, the local models in the re-ordered trees (Figure 3.2 (b) and Figure 3.3) are compared with the reconfigured tree (Figure 3.4 (b)). For example, if $Z_{1^*} \le Th_1$, $Z_{2^*} \le Th_2$ and $Z_{3^*} > Th_4$, the local models for prediction are $f_{1^*}^1(\cdot)$ and $f_{1^*}^2(\cdot)$, which are the same as the models with the same splitting conditions, circled by dashed circle in Figure 3.4 (b).

After the reconfiguration, the splitting variables are re-ordered into different layers, which map to the temporal order of the manufacturing stages, as shown in Figure

3.5. The splitting conditions are combined, which lead to different model groups to predict the intermediate and final quality variables stage-by-stage. This reconfigured PLRT is preferred over the original PLRT for the purpose of the feedforward control.

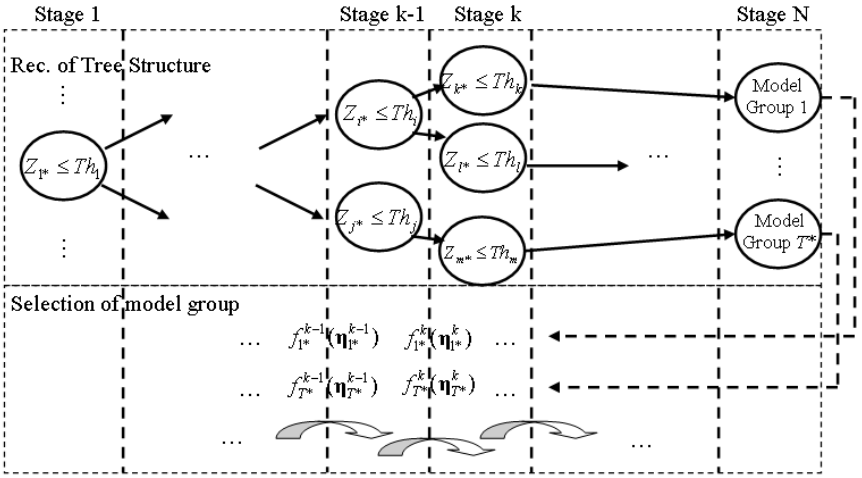


Figure 3.5: Reconfigured PLRT for a MMP

3.4 Reconfigured Model Complexity and Control Accuracy

The PLRTs from GUIDE are pruned by cross validation to minimize the predicted SSE (Loh, 2002). After the reconfiguration, the reconfigured model yields the best prediction accuracy due to the statistical equivalency. However, the reconfigured PLRT may be very complex with many leaf nodes and many potential local models, which increases computational efforts in the control optimization. On the other hand, there is an engineering tolerance for the controlled objectives, which can be further transferred to the needs of the model precision used in the feedforward control. In other words, the model used for control purpose may not have the same level of high precision requirement as the prediction obtained from the original PLRT. Therefore, the model complexity can be reduced, while the model still satisfies the control accuracy requirements. The reduction of the model complexity is achieved by assuming that there are limited numbers of

variables having nonlinear relationship with the response. Detail discussions on how to further simplify the reconfigured PLRT with fewer leaf nodes is provided below.

In a reconfigured PLRT, the control performance can be evaluated by the accumulative errors of all PLRT model errors at different stages. However, different model groups may be selected in control according to the splitting conditions. Thus, it is difficult to estimate the control accuracy for every possible path used in control. In this chapter, the largest prediction variance is proposed to evaluate the control accuracy of this leaf node, shown as follows:

$$\sigma_{k,\mathrm{j}}^2 = \max_{\mathbf{U}_1,\ldots,\mathbf{U}_\mathrm{N},\mathbf{X}_1,\ldots,\mathbf{X}_\mathrm{N}} \mathrm{Var}(\mathbf{Y}(\mathrm{N})_\mathrm{j}) \tag{3.3}$$

$$\text{s.t.} \quad x_{\mathrm{lk}} \in \{x_{\mathrm{lk}}\},\ u_{\mathrm{lk}}^{\mathrm{L}} < u_{\mathrm{lk}} < u_{\mathrm{lk}}^{\mathrm{U}}, \forall \mathrm{l}, \forall \mathrm{k}$$

where $\sigma_{k,\mathrm{j}}^2$ is the maximum prediction variance of the j-th quality variable in the *k*-th leaf node, obtained by enumerating all control actions; $\mathbf{Y}(\mathrm{N})_\mathrm{j}$ is the predicted final quality variable; the optimization constraints are the controllability of the process variables, where $\{x_{\mathrm{lk}}\}$ is the set of all possible settings of x_{lk}, and $u_{\mathrm{lk}}^{\mathrm{L}}, u_{\mathrm{lk}}^{\mathrm{U}}$ represents the lower and upper bound of the feasible range for u_{lk}.

The control accuracy of the overall structure is evaluated by the pooled variance of these leaf nodes. Assuming there are equal numbers of products in different leaf nodes in control, thus, the pooled variance is the average of the control accuracy of all leaf nodes, shown as follows:

$$\sigma_{\mathrm{Rec.,j}}^2 = \frac{1}{T}\sum_{k}^{T} \sigma_{k,\mathrm{j}}^2 \tag{3.4}$$

where T is the number of leaf nodes in the reconfigured PLRT; and $\sigma^2_{\text{Rec.,j}}$ represents the control accuracy for the j-th quality variable. In this way, the control accuracy is evaluated by $\sigma^2_{\text{Rec.,j}}$.

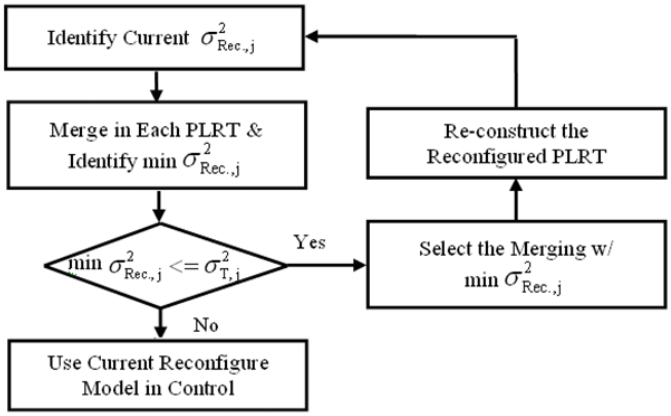


Figure 3.6: The procedure to reduce model complexity

To reduce the model complexity, the leaf nodes should be merged. With less leaf nodes, the prediction performance will be degraded because the PLRTs are pruned to minimize the predicted SSE in the cross validation. There are two issues to be addressed to balance the model complexity and the control accuracy: (1) which leaf nodes should be merged, and (2) when the merging process should be stopped?

The leaf nodes with the least important splitting structure should be merged first because it would result in the smallest decrease in the prediction accuracy. Although the control accuracy is evaluated based on the reconfigured PLRT, the temporarily complied splitting variables no longer provide information on the importance of splitting structure. Nevertheless, the original PLRTs preserve the importance of the splitting variables for prediction in splitting orders from the more significant ones to the less significant ones. Therefore, reducing the number of leaf nodes will merge the nodes in the deepest layer in the original PLRTs. The merging process is stopped when the control accuracy of the

reconfigured PLRT exceeds the pre-determined control accuracy requirement. The merging process is completed in an iterative way shown in Figure 3.6.

In Figure 3.6, the control accuracy of the current reconfigured PLRT $\sigma^2_{\text{Rec.,j}}$ is estimated first. Then different deepest leaf nodes in the original PLRTs are merged once at a time. In this way, a set of new control accuracy estimates $\sigma^2_{\text{Rec.,j}}$ of the final reconfigured PLRTs is obtained. We choose the minimal $\sigma^2_{\text{Rec.,j}}$ and compare it with a pre-determined threshold $\sigma^2_{\text{T,j}}$ for the j-th quality variable. One concludes that the model with minimal $\sigma^2_{\text{Rec.,j}}$ is acceptable if it is smaller than $\sigma^2_{\text{T,j}}$. In this case, we reconstruct the reconfigured PLRT in the next iteration. Otherwise, the control accuracy of the current model does not satisfy the control accuracy requirement, thus the merging should be stopped. After this procedure, the reconfigured model has reached a balance between the model complexity and the control accuracy.

3.5 Feedforward Control Strategy of Reconfigured PLRTs

The engineering-driven reconfiguration has made it possible to develop a feedforward control strategy by actively adjusting the process variables and compensating the quality variable for variation reduction. The overall strategy is shown in Figure 3.7. The basic idea to achieve a feedforward control based on the reconfigured PLRT models is presented below.

At each controllable stage, several potential model groups are determined based on the splitting conditions. If the splitting variables are measured at previous stages or layers, a model group in a leaf is selected when the splitting conditions are satisfied. Otherwise, several branches and leaves may be selected, which form a cluster of potential

model groups. In this case, the splitting conditions are formulated as constraints in the optimization problem.

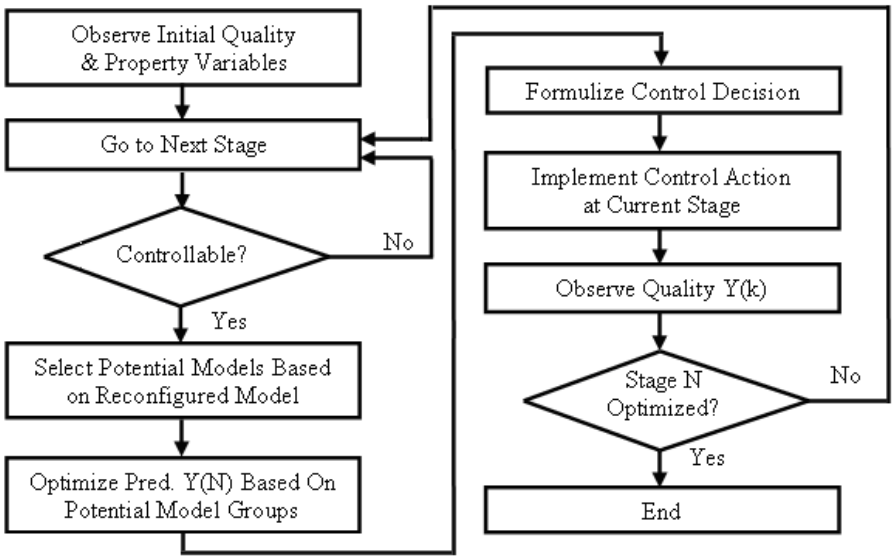


Figure 3.7: The overall feedforward control strategy

The control optimization at the k-th stage is formulated as the-smaller-the-better problem:

$$\min_{u_{\mathrm{li}},x_{\mathrm{li}},\,\mathrm{i}=\mathrm{k},\cdots\mathrm{N}} J(\mathbf{U},\mathbf{X})=\sum_{\mathrm{j}=1}^{\mathrm{m}} \mathrm{c}_{\mathrm{j}}\,\mathrm{E}(\mathbf{Y}(\mathrm{N})_{\mathrm{j}}^{2}) \tag{3.5}$$

$$\begin{aligned} s.t. \quad & \mathbf{Y}(\mathrm{N})_{\mathrm{j}} = f_{\omega}^{\mathrm{j}}(\boldsymbol{\eta}) \\ & h(\mathbf{Y}(\mathrm{s})_{\mathrm{j}}) < H_{\mathrm{js}} \\ & x_{\mathrm{li}} \in \{x_{\mathrm{li}}\} \\ & u_{\mathrm{li}}^{\mathrm{L}} < u_{\mathrm{li}} < u_{\mathrm{li}}^{\mathrm{U}} \\ & I(g_{\omega}(Z_1,\ldots,Z_L)) > 0 \\ & \mathrm{s}=1,2,\cdots,\mathrm{N};\mathrm{i}=\mathrm{k},\cdots,\mathrm{N}. \end{aligned}$$

where the objective function is the weighted summation of the second order moment of m predicted final quality variables; $\mathbf{Y}(\mathrm{N})_{\mathrm{j}}$ is the j-th final quality variable predicted from the k-th stage; c_{j} is the weight of the importance of the j-th quality variable. The decision variables are the process variables from the k-th stage to the N-th stage. In the

constraints, $f_{\omega}^{j}(\cdot)$ is a potential model group for the quality prediction determined by the splitting conditions; $h(\mathbf{Y}(s)_j) < H_{js}$ represents the quality specification for the j-th quality variable at the s -th stage ($s = 1,2,\cdots,N$); $u_{li}^{L} < u_{li} < u_{li}^{U}$ and $x_{li} \in \{x_{li}\}$ ($i = k,\cdots,N$) represent the feasible ranges as described in Equation (3.3). The optimization problem is solved by Iterated Local Search Algorithm (Stutzle, 1998).

3.6 Case Study

A case study in a multistage wafer manufacturing process (MWMP) is conducted to illustrate the procedure of modeling and control based on the reconfigured PLRTs. A comparison study of the feedforward control strategy based on a reconfigured PLRT and regression model groups is conducted to show the effectiveness of the proposed approach.

3.6.1 Wafer Manufacturing Processes

A MWMP is a complex MMP involving chemical and mechanical process to transform a silicon ingot into a wafer with uniform thickness, fine surface roughness, and good overall geometric shape for future processing. The process in this case study consists of five major manufacturing stages as shown in Figure 2.4, including slicing, lapping, chemical vapor deposition (CVD) of polysilicon, CVD of SiO_2, and polishing. Each stage is a combination of multiple operations with quality measured at the end of the stage.

In a MWMP, the overall geometric shape is a critical geometric quality index of a wafer. BOW and WARP of a wafer represent the overall shape of a wafer, which is used

as the quality improvement objective in the case study. In general, smaller absolute values of these variables indicate better quality of the wafer.

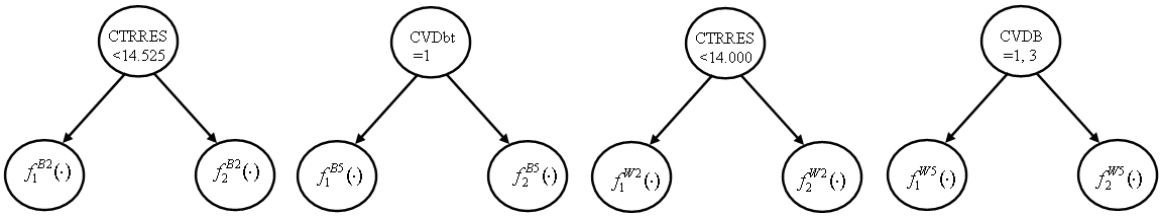
In this case study, observational data of three types of variables (quality, process, and material property) were collected in a real production environment. Those variables are summarized in Table 2.2 in Chapter 2. In this table, the CTRRES represent the position of wafers in an ingot. In the case study, the central thickness of a wafer is measured in each stage, which is used in the selection of settings of downstream process parameters. Therefore, the central thickness of wafer is treated as a predictor rather than quality variables. The initial quality vector $\mathbf{Y}(0)$ in this process is assumed to be a zero vector.

In this process, some intermediate quality specifications of wafers need to be satisfied. For example, the thickness of a wafer in certain lapping batch should be within a specified range; otherwise, the wafer will be broken during the lapping. These intermediate quality specifications are formulated as constraints in the optimization problem. Overall, data of 373 wafers are obtained in production for the case study. The PLRTs are constructed based on the training data set (250 wafers) and the control performance is evaluated based on the testing data set (123 wafers).

3.6.2 PLRT Models of the MWMP

The PLRTs for this MWMP are estimated and shown in Figure 3.8. In Figure 3.8, there are four splitting structures to predict BOW2, BOW5, WARP2, and WARP5, while the models for other quality variables are regression models without splitting structures. In each leaf node, there is a local regression model, where "B" or "W" represents the quality variable BOW or WARP respectively. The PLRTs have explicit interpretations. For

example, material property CTRRES at different segments of ingot yield different prediction models to predict BOW2 (Figure 3.8 (a)). This shows that the prediction of BOW2 is influenced by the material heterogeneity of wafers at the tail and the head of the ingot. Similar interpretations are obtained for BOW5, WARP2, and WARP5.



(a) Model BOW2 (b) Model BOW5 (c) Model WARP2 (d) Model WARP5

Figure 3.8: PLRTs in MWMP

Since GUIDE does not consider the interactions in estimating the local regression models, the regression model of $f^{B3}(\cdot)$, $f^{B4}(\cdot)$, $f^{W3}(\cdot)$ and $f^{W4}(\cdot)$ is re-estimated considering the interactions of predictors to further reduce the predicted SSE.

3.6.3 Reconfiguration of PLRTs

Based on the PLRT from GUIDE, a reconfigured PLRT is obtained in Figure 3.9. In Figure 3.9, the temporal order of the splitting variables is CTRRES $\prec$ CVDB $\prec$ CVDbt , which is re-ordered into different layers of the reconfigured PLRT from the root. In this example, CTRRES is split into three sub-regions in the first layer of the model, which are based on the splittings in the original models to predict BOW2 and WARP2. In the second and the third layer of the model, CVDB and CVDbt are split as the same as the original model. In this way, 12 regression model groups are generated, which will be selected by the splitting conditions. The overall structure clearly represents the sequence

of manufacturing from the root to the leaf nodes, and predicts multiple intermediate and final quality variables.

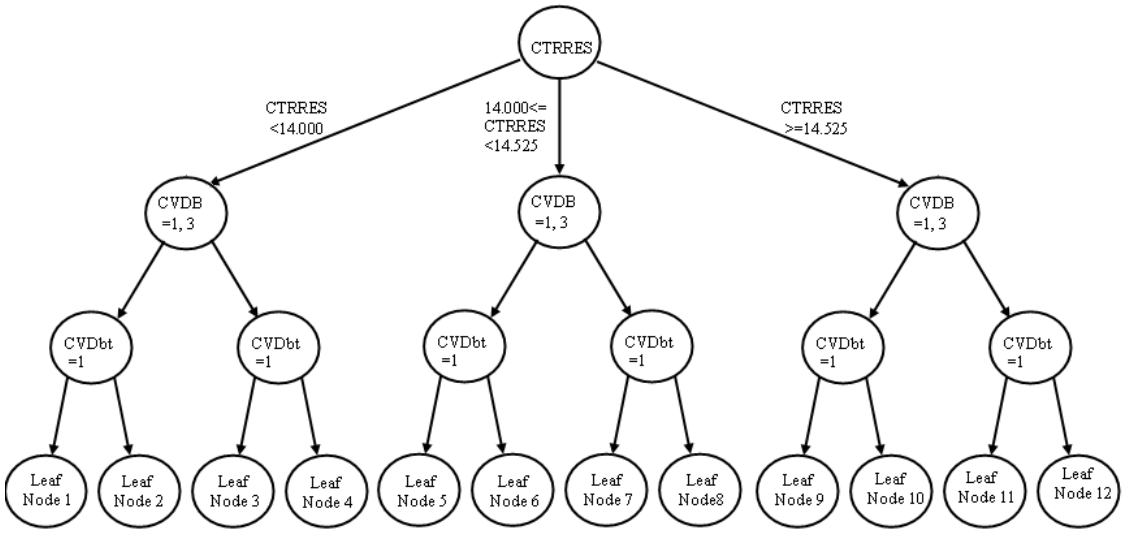
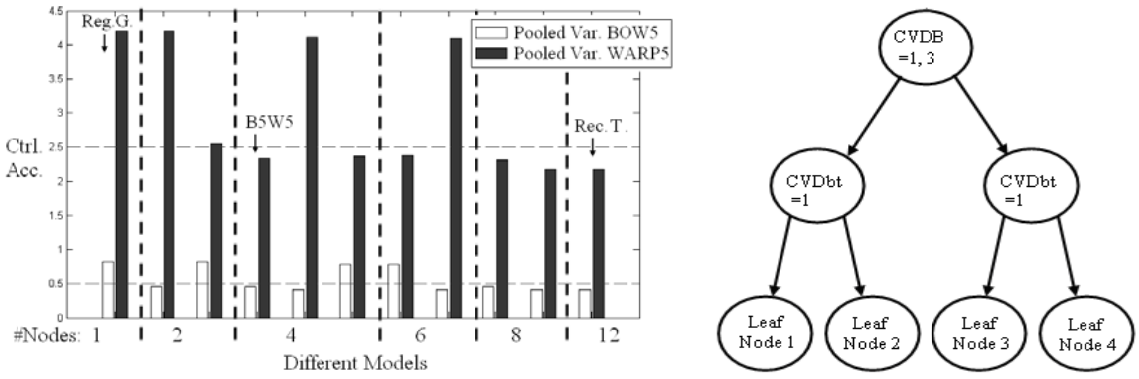


Figure 3.9: Reconfigured PLRT for MWMP

3.6.4 Reduce Model Complexity

To reduce the model complexity, the control accuracy of BOW5 and WARP5 are evaluated in different reconfigured models. Figure 3.10 (a) shows control accuracy of 11 models from regression group (Reg. G.) with the worst control accuracy to the reconfigured PLRT (Rec. T.) with the best control accuracy. The number of nodes is marked for each model. In this figure, the control accuracy varies as different model complexities are adopted. Such an analysis provides guidelines to select a model with appropriate complexity that satisfies the control accuracy requirement. In this case study, the control accuracy requirement of BOW5 and WARP5 are 0.5 and 2.5 (horizontal dashed lines). The model with splits in BOW5 and WARP5 (B5W5) has the minimal number of leaf nodes to satisfy the requirement, which has only two significant splitting

variables and four leaf nodes retained for control optimization, shown in Figure 3.10 (b). By comparing the "Rec. T." model, the model complexity has been significantly reduced.



(a) Control accuracy of different models (b) Final reconfigured PLRT for control

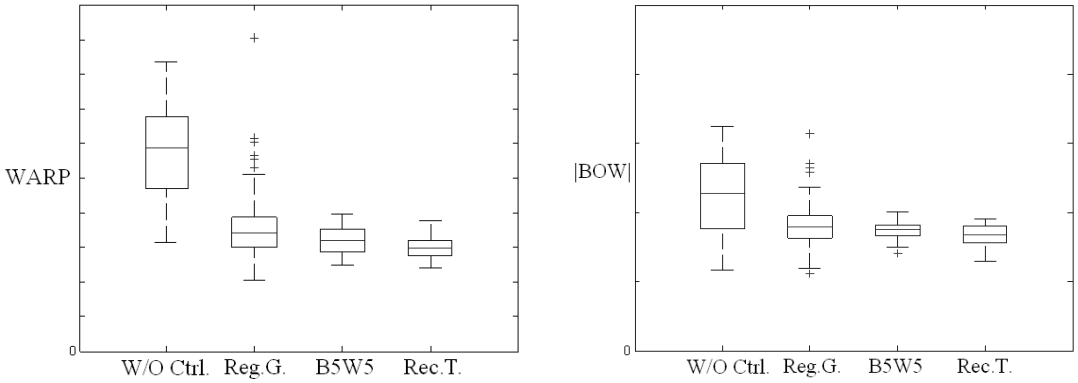
Figure 3.10: Control accuracy and model complexity

3.6.5 Simulation Study of Feedforward Control

To compare the feedforward control strategy, a total of 50 simulation runs were conducted based on three different models: "Reg. G.", "B5W5" and "Rec. T.". In the simulation, the "Reg. G." model is a global regression model without using splitting variables. The "B5W5" and "Rec. T." models use the reconfigured PLRT models for prediction. Without loss of the generality, we set $c_j = 1$ in Equation (3.5).

Figure 3.11 (a) shows the controlled WARP5 in one simulation run. The horizontal axis represents the performance without control and with control based on different models. The control based on the reconfigured PLRTs yields better performance in reducing mean and variance of the final quality than regression group models. Here the control based on the regression group models is the same result in the intermediate feedforward control strategy (Jin and Shi, manuscript). Moreover, there is

no significant increase in mean and variance of the controlled quality when using the "B5W5" model verses the "Rec. T.". This indicates that there is no significant loss in control performance when merging some of the splitting structures. Figure 3.11 (b) shows controlled performance of the absolute value of BOW5 with similar interpretations.



(a) WARP5 (b) |BOW5|

Figure 3.11: Controlled quality performance in a simulation run

Table 3.5: Controlled objective values in simulations

	Reg. G.	B5W5	Rec. T.
Mean	383.80	340.68	307.05
St. Dev.	35.63	7.64	15.20

The values of the optimal objective function of 50 simulation runs are summarized in Figure 3.12. The values of the optimal objective function based on "Reg. G." are larger than those based on reconfigured PLRT in most of the simulation runs, i.e., a better control performance is obtained with the reconfigured PLRT model. The "Rec. T." model has a better controlled performance than the "B5W5" model. However, a more complex model structure leads to a higher demand on computational efforts. The proposed reconfigured PLRT with reduced model complexity has less leaf nodes and

sacrifices the control accuracy, but it still sufficiently meets the control requirements from an engineering perspective.

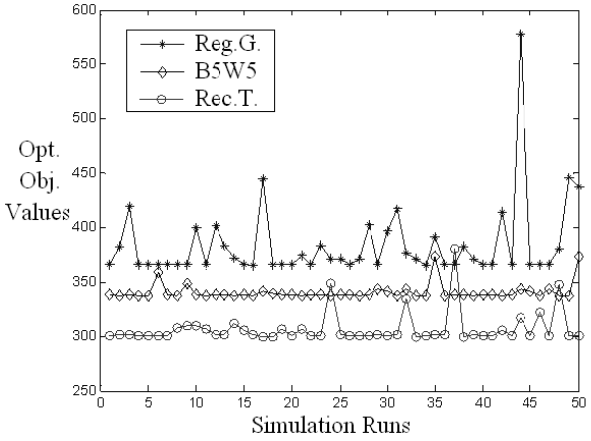


Figure 3.12: Comparison of control performance based on different models

The mean and standard deviation of the optimal values are summarized in Table 3.5. There is an average of 11.24% and 20.00% reduction in objective value for the "B5W5" and "Rec. T." compared to "Reg. G.". The standard deviation of the values of the objective function is also reduced for the proposed "B5W5" model. The study indicates that the reconfigured PLRT is more effective in variation reduction than the standard regression model group based on the proposed control strategy.

3.7 Conclusion

It is a challenging task to model the variations and their propagations in MMPs, especially when the relationships among process parameters and product quality variables are nonlinear. In this case, a PLRT model can be adopted that has high prediction accuracy and explicit interpretation in describing nonlinear data structure. However, it fails to illustrate the temporal order and inherent relationships among variables in a MMP.

This chapter bridges the gap between the needs for advanced models for MMP variation reduction and the limitations of PLRT. An engineering-driven reconfiguration of the PLRT is proposed to convert the original model into an engineering compliant model. The reconfigured PLRT not only has the high prediction accuracy of the original tree structure, but also provides a feasible solution in determining the potential prediction models sequentially as the operations move from the upstream stages to the downstream stages. This sequential model selection procedure enables its capability in active compensation by implementing a feedforward control strategy. The model complexity is also reduced by analyzing the control accuracy of the models. A case study has been conducted in a real MWMP, which demonstrates better control performance by using the reconfigured PLRT model compared to that using a standard regression model.

CHAPTER 4

SEQUENTIAL MEASUREMENT STRATEGY FOR WAFER GEOMETRIC PROFILE ESTIMATION

4.1 Introduction

In semiconductor manufacturing, the geometric shape of wafers is an important index to evaluate wafer quality. For example, the profiles could be used to estimate quality variables defined by the Semiconductor Equipment and Materials International (SEMI) as industrial standards, such as Total Thickness Variation (TTV), Bow and Warp. These variables are not only used for quality measures of the final wafer product, but also for identifying root cause of surface imperfections (Pei *et al.,* 2003; Pei *et al.*, 2004; Zhu and Kao, 2005) during a production. Moreover, the geometric profiles of wafers are modeled for optimal design of process variables in wafer manufacturing processes (Zhao *et al.*, 2011), which requires timely online measurements of the wafer geometric profiles.

In order to provide effective process control of wafer manufacturing processes, it is desirable to quickly obtain wafer geometric profile measurement with adequate accuracy. However, current measuring procedure is time consuming and unable to provide wafer profile information in a timely manner. For example, the existing wafer measurement technology, such as a touching probe type of sensors, takes more than eight hours to measure one typical batch of wafers (e.g. 400 wafers in one production run). Time consuming measurement prohibits the implementation of advanced process

monitoring and diagnosis technologies for quality improvement. Therefore, the objective of this research is to develop an efficient and systematic measurement strategy to reduce the measurement time through sequential sampling and modeling. In this chapter, because of the limitations of evaluating the measurement time on the real sensor system, we propose to minimize a composite index based on the measured sample size and times of model fittings as the efficiency improvement index:

$$Comp.Index = \tau \frac{n_{\text{total}}}{\max(n_{\text{total}})} + (1-\tau)\frac{I_{\text{total}}}{\max(I_{\text{total}})} \quad (4.1)$$

where n_{total} is the total sample size measured for a wafer; I_{total} is the total times of model fittings in the measurement strategy; τ is a weighting coefficient to evaluate the measurement time for each point and the computation time; and $\max(n_{\text{total}})$ and $\max(I_{\text{total}})$ are the maximum of the total sample size and total times of model fittings for a batch of wafers, which are used to normalize the effects of sample size and number of model fittings. When reaching the same accuracy with smaller composite index, we consider the measurement strategy has better efficiency.

In the wafer manufacturing processes, high definition samples of each wafer are measured as geometric profiles. There are different methods to model the geometric profiles from different perspectives in the literature. From the engineering perspective, physical analytical models, such as finite element analysis or partial differential equations, are adopted to model the geometric profiles (Zhang and Kapoor, 1990; Abburi and Dixit, 2006; Ozcelik and Bayramoglu, 2006; Huang and Gao, 2010). A major limitation in these methods is that they require a sophisticated understanding of the profile formation. Another limitation is that these methods are usually used to model a deterministic profile with limited capabilities in modeling the randomness of the profile

errors or the random filed effects. Some other approaches, such as methods in computer graphics, use Spline (Forsey and Bartels, 1988; Lee *et al*., 1997; Sederberg *et al*., 2004), or wavelet analysis (Schroder, 1996; Valette and Prost, 2004) to model the profile data. In most cases, the potential factors to the shape or characteristics are not considered for the profile in modeling.

In this chapter, a Gaussian Process (GP) model is used to characterize the spatial correlated geometric shape of a wafer, including the profile mean, correlated variability and measurement noise. One of the advantages of GP model is that the correlated variability can be further decomposed into the global variability and local variability components with nice interpretations. The former one represents the trend of variation over the whole wafer, while the latter one captures the variation only within a relevant neighborhood to the measurement locations.

In order to implement sampling strategy based on GP models, several efforts have been reported for optimal sampling scheme in a most economic way. In the spatial statistics, researchers have employed the grid spacing determination approaches to reduce the sample size. By maximizing the grid space, the sampling cost will be minimized in an optimal sampling scheme under the constraints of an allowed maximum error variance (Curran and Williamson, 1986; Curran 1988). Others extend the previous work to determine the optimal grid spacing designs for sampling multiple variables by conditional kriging variance based on cross-correlations among variables (McBratney and Webster, 1983a; McBratney and Webster, 1983b; Atkinson *et al.*, 1992; Atkinson *et al.*, 1994). Moreover, relationships between estimation accuracy of the response variable and required sample size are explored and investigated (Wang *et al.*, 2005; Xiao *et al.*, 2005).

However, a major limitation of the aforementioned sampling strategies is that the local spatial variability of the response variable is neglected, which may vary from location to location. To attack these limitations, variable grid spacing approaches are developed based on the local variability (Anderson *et al.*, 2006). In the region with a higher local variability, a smaller grid space is determined, which is equivalent to measuring more samples in the neighborhood, vice verse.

There are different measurement strategies by sequentially allocating the samples based on prior information. One type of strategies is widely used in optimal sensor selection or allocation problem. Another type of strategies is developed in computer experiments (CEs).

In the optimal sensor selection or allocation problem, posterior distributions based on prior measurements are used for sensor location determination to maximize the information gain. When it is difficult to evaluate an exact posterior distribution, Sequential Monte Carlo (SMC) method is used for numerical approximation. The SMC method has shown a powerful ability to solve both the sophisticated statistical problem and engineering applications (Liu and Chen, 1998; Doucet *et al.*, 2000; Doucet *et al.*, 2001). The Bayesian SMC method is also proposed to solve the optimal sensor selection and fusion in target tracking and localization applications (Guo and Wang, 2004). However, the performance of these methods depends on proper parametric form of the Bayesian model, and they are generally computational intensive for posterior calculations.

The sequential design in CE is another stream of sampling strategies, which has been well developed to find the optimum of inputs (Schonlau *et al.*, 1998; Williams *et al.*,

2000; Park *et al.*, 2002; Kleijnen and Beers, 2004; Huang *et al.*, 2006). One of the objectives of sequential design is to reduce the experimental runs to reach the optimal solution, which refers to minimum or maximum of the response. A sequential measurement design strategy is proposed to sequentially allocating more sampling points at the locations with a higher expected improvement (EI) to quickly reach the minimum of the investigated surface (Williams *et al.*, 2000). In their work, a larger expected improvement is defined as the locations with a smaller predicted value or a larger predicted variance for the minimization type of problem.

Other than focusing on minimizing the required experimental runs to obtain the optimal solution, there are other GP-based sequential sampling works, which focus on how to sample sequentially in order to obtain a better model fitting, conditional on the new pair of sample points. These models are usually obtained from the posterior distributions via MCMC. Different thrifty criterions based sequential sampling problem could be found in MacKay, 1992; Cohn, 1996; Muller *et al.*, 2004. Some sequential applications can be shown to approximate static optimal designs, see Seo *et al.* 2000; Gramacy and Lee, 2009.

The optimal sampling schemes and sequential designs provide effective ways in reducing the sample size by solving a set of optimization problems. However, some have limitations in computations, and others may not be applicable for online measurement tasks. For grid spacing determination, the chosen sample locations are not directly associated with the locations with higher local variability within each grid. And for the sequential design, some methods target on optimization objectives, which is not the same as online measurement. Moreover, most of the sampling schemes and existing sequential

measurements involve computational intensive optimization procedure to determine additional samples.

This chapter continues the stream of sequential design in CE to measure samples sequentially, called "sequential measurement strategy", but differs in the following ways. First, the proposed sampling scheme is aimed at considering both the global variation trend and the local variability pattern simultaneously to achieve more accurate prediction ability. Second, the prior engineering knowledge of input-output relationship is taken into account to determine the initial measurement samples. By combining these two aspects, the proposed sequential measurement strategy enhances the wafer quality profile prediction performance with a higher efficiency. Although the proposed framework is similar to the sequential design, the innovation of this chapter lies in two proposed empirical distributions for initial measurements and sequential measurements, which will be discussed in details later.

The rest part of the chapter is organized as follows. The GP-based sequential measurement strategy is described in detail in Section 4.2. A real case study is provided in Section 4.3 to evaluate the proposed measurement strategy for wafer thickness profile estimation in a slicing process. Finally, the conclusion and future work are summarized in Section 4.4.

4.2 GP Model based Sequential Measurement Strategy

4.2.1 Overview of the Sequential Measurement Strategy

The framework of the proposed sequential measurement strategy is shown in Figure 4.1.

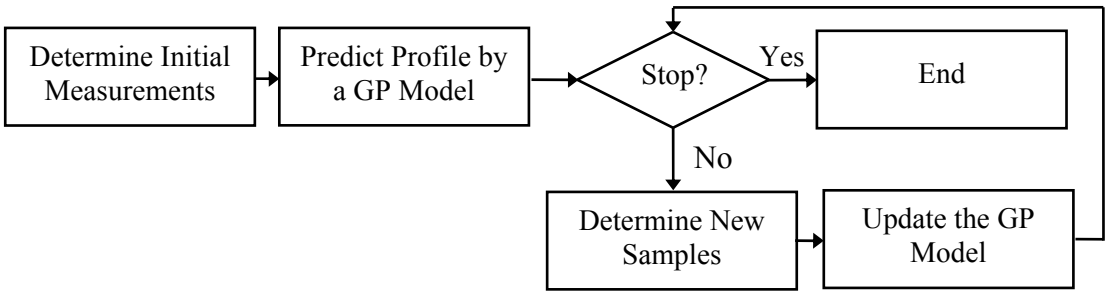


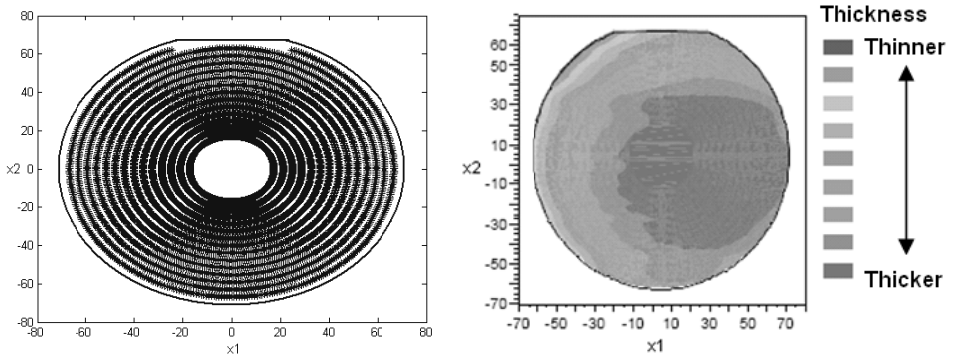
Figure 4.1: A framework of sequential measurement strategy

In the proposed methodology, all measurement locations are determined by sampling empirical distributions. The empirical distributions are the estimated probability density functions evaluated at discrete potential measurement locations. The sequential measurement strategy starts to sample an empirical distribution obtained from the engineering knowledge, then fits a GP model based on the initial measurements. In the estimation, the measured locations are partitioned into a training sample set and a testing sample set. A GP model is estimated based on the training sample set and the model accuracy is evaluated based on the testing sample set. If the stopping rule is satisfied, the iterative measurement procedure stops; otherwise, additional samples are measured to further improve the estimation performance. In this approach, the magnitude of gradient and the predicted Mean Squared Error (MSE) from the previous GP model are used to determine the sequential measurements. By iteratively taking the samples and re-fitting the model, the GP models are expected to better approximate the true wafer profile closely.

4.2.2 Measurement Locations and Data Format

The gauge used in this wafer measurement study is a touching probe type of sensor, which has minimal distance of allowable movement and the maximum measuring range. The specification of the gauge defines a potential measurement zone, denoted as D for

the set of potential measurement locations. In a continuous measuring scheme, there is a position calibration mechanism such that all wafers are measured at the same locations. For each potential measurement location, there are several quality features to be measured. In the wafer example, wafer thickness, flatness, and local warp are interested quality features. Each of these quality features forms a highly spatially correlated data profile, called a "wafer geometric profile".



(a) Potential measurement locations (b) Thickness profile

Figure 4.2: Potential measurement points and measurement result

Figure 4.2 shows an example of the potential measurement locations for a wafer and the measurements of wafer thickness profile. In Figure 4.2 (a), the solid curve represents the edge of the wafer, and the inner rings formed by stars represent the potential measurement locations in *D*. The line segment at the top of a wafer is the reference edge. The total number of potential measurement locations is usually a large number in practice, which is denoted as n_D. For example, n_D is larger than 5000 for a 6-inch wafer. In Figure 4.2 (b), the thickness readings of wafers are taken at the potential measurement locations of Figure 4.2 (a), which form the wafer thickness profile. The grey scale represents the thickness of the wafer in different ranges. And the thicknesses of unmeasured locations are estimated by interpolating the measured points. It is clear

that there is thickness variation in the wafer as the left up corner is thinner. Similar format of data can be obtained from the measurement gauge for other geometric profiles.

4.2.3 Determination of Initial Measurement Samples

The determination of the measurement samples is equivalent to select a subset of the overall potential measurement locations. When there is no prior knowledge regarding the profile distribution, there are two typical ways to determine the initial samples. One is to use random measurement strategy, i.e., to measure the samples with randomly selected locations; the other way is to view the initial measurement samples as a design in CE and then incorporate Latin Hypercube Design (LHD) or uniform design with space filling criterions to define sample locations (Santner *et al.*, 2003).

The random measurement approach is straightforward and easy to be implemented, but a major limitation of this approach is that any two chosen initial samples may be too close to each other. Since the wafer profile is spatially correlated, the samples within the neighborhood not only have high correlation and hence contribute little information towards better fitting, but also may lead to singularity when inversing the correlation matrix in model estimation (Santner *et al.*, 2003).

The space filling design in CE has advantages over the random sampling strategy by increasing the pairwise distance in an initial design. However, some typical forms may be difficult to implement in the regions with irregular shapes, such as the circles of the potential measurement points for a wafer. Other typical forms of space filling design choose the design points mainly based on the distance among the input variables, which ignore the relationship between response (profile) and input variables (locations). In this case, the design may be inefficient, especially when the input-output relationship is

available from engineering knowledge. More advanced techniques may use this input-output relationship in design with a more complicated way (Santner *et al.*, 2003). Therefore, this chapter proposes a computationally efficient approach by sampling a weighted empirical distribution from engineering knowledge.

To efficiently determine the measurement samples, the local variability of a wafer profile is first defined as:

Definition 4.2.1 *Local variability of a wafer profile*: the local variability of a wafer profile at location $\mathbf{x}$ is the sample variance of the profile values at the measured locations within a neighborhood region $\mathfrak{N}(\mathbf{x})$:

$$\sigma_Y^2(\mathbf{x}) = \frac{1}{n_{\mathbf{x}}-1}\sum_{\mathbf{x}'\in\mathfrak{N}(\mathbf{x})}(Y(\mathbf{x}') - \bar{Y}(\mathbf{x}'))^2 \tag{4.2}$$

where $\mathbf{x}'$ is a location within the neighborhood of $\mathbf{x}$; $\bar{Y}(\mathbf{x}')$ is the sample mean of the profile values in the neighborhood; and $n_{\mathbf{x}}$ is sample size in that neighborhood. The neighborhood $\mathfrak{N}(\mathbf{x})$ is defined as the *k*-th nearest neighborhood. Generally speaking, *k* is chosen to be at least 25 in a usual case so that a good estimation of variance could be guaranteed.

To determine efficient initial measurement samples, the regions with potentially a higher local variability should have more samples taken in these regions, as shown in the grid spacing determination approach (Anderson *et al.*, 2006). In the wafer manufacturing process, the ideal wafer profile has uniform thickness. The observed thickness deviate from the ideal profiles, which have a set of potential root cause factors, denoted as $U(\mathbf{x}) = \{u_1(\mathbf{x}),...,u_q(\mathbf{x})\}$. Here, $u_i(\mathbf{x})\,(i = 1,2,\cdots,q)$ are the potential factors associated with wafer locations, such as the contact span at $\mathbf{x}$. The local variability of the profile

has a proportional relationship of its potential factors from the engineering perception and verified by the data, shown as:

$$\sigma_Y^2(\mathbf{x}) \propto \frac{\prod_{i=1}^{q_1} u_i^{t_i}(\mathbf{x})}{\prod_{j=q_1+1}^{q} u_j^{t_j}(\mathbf{x})} \tag{4.3}$$

where $\prod_{i=1}^{q_1} u_i^{t_i}(\mathbf{x})$ is the product of factors proportional to the local variability of the profile; $\prod_{j=q_1+1}^{q} u_j^{t_j}(\mathbf{x})$ is the product of factors inverse proportional to local variability of the profiles; there are q_1 proportional factors and $(q - q_1)$ inverse proportional factors; and t_i and t_j are parameters in the power for each factors.

Table 4.1: The procedure to determine the initial measurement locations

Step 1. Obtain the proportional relationship of the local variability of the wafer profile with its potential factors in Equation (4.3)
Step 2. Estimate the empirical distribution for the initial measurement using Equation (4.4)
Step 3. Determine the sample size n_0 and allocate the sample sizes to circles, which is proportional to the summation of the probability of the points on that circle
Step 4. Sample the points from the outmost circle to the innermost circles. For each circle, points are sampled for G times, and the samples are selected to have max-min distances to the samples on the circles outside
Step 5. Measure the wafer profile at the locations determined in Step 4

Based on the proportional relationship in Equation (4.3), the initial measurement points are determined by sampling an empirical distribution defined as

$$\Pr(\mathbf{x}) = \frac{1}{c_1} \frac{\prod_{i=1}^{q_1} u_i^{t_i}(\mathbf{x})}{\prod_{j=q_1+1}^{q} u_j^{t_j}(\mathbf{x})} \tag{4.4}$$

where c_1 is the corresponding normalizing constant.

By sampling the empirical distribution defined in Equation (4.4), the sample locations with larger local variability will have a higher probability to be selected as the initial measurements. A detailed procedure is summarized in Table 4.1. In this

procedure, we use stratified sampling and a max-min criterion to determine initial measurements from the outmost circle to the innermost circles. The max-min criterion is

$$\max_{\forall \mathbf{x}_1 \in \mathrm{Cir}_1, \mathbf{x}_2 \in \mathrm{Cir}_2} \min \left\| \mathbf{x}_1 - \mathbf{x}_2 \right\|_2 \tag{4.5}$$

where $\mathrm{Cir}_1$and Cir_2 are the sets of locations of the outer circles and the inner circles in Step 4 of Table 4.1.

4.2.4 GP Models for Wafer Geometric Profiles

Based on the measurements, a GP model is adopted to model a wafer geometric profile as

$$Y(\mathbf{x}) = \mathbf{f}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta} + Z(\mathbf{x}) + \varepsilon \tag{4.6}$$

where $\mathbf{f}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}$ represents the mean part of the wafer profile; in general, the basis functions $\mathbf{f}^{\mathrm{T}}(\mathbf{x}) = \left[f_1(\cdot), \cdots, f_p(\cdot)\right]$ are known; $\boldsymbol{\beta} = \left[\beta_1, \cdots, \beta_p\right]^{\mathrm{T}}$ is the regression coefficient vector; $Z(\mathbf{x})$ is a Gaussian process with mean 0 and covariance function $\sigma_Z^2\psi$; σ_Z^2 is the variance of the covariance function, which represents the wafer profile fluctuation caused by manufacturing error; and ε is the uncorrelated noise term follows normal distribution $NID(0, \sigma_\epsilon^2)$, which represents the measurement noise. Note that the correlation function applied is a commonly used anisotropy Gaussian correlation function:

$$\psi\left(\mathbf{x}_j, \mathbf{x}_k\right) = \exp\left(-\sum_{i=1}^{p} \phi_i (\mathrm{x}_{ij} - \mathrm{x}_{ik})^2\right) \tag{4.7}$$

where $\phi_i$is the scale parameter associated with the i-th predictor; $\boldsymbol{\Phi} = \left[\phi_1, \cdots, \phi_p\right]$; and p is the dimension of the input variables. In the wafer profile estimation problem, $\mathbf{x}_j$ is the j-th location on the wafer with coordinate$(\mathrm{x}_{1j}, \mathrm{x}_{2j})$, and p=2. To be more specific, x_{1j} is the axis parallel to the reference edge of a wafer, and x_{2j} is the axis perpendicular to the reference edge of a wafer. The origin is at the geometric center of the wafer.

In the wafer profile estimation problem, we use an ordinary kriging model to fit the wafer geometric profile:

$$Y(\mathbf{x}) = \beta_0 + Z(\mathbf{x}) \tag{4.8}$$

where β_0 is the constant mean part. This simplification is based on the fact that (1) the GP model with a constant mean part is adequate to model the wafer profile; and (2) the measurement noise of the wafer profile ε is negligible compared with the profile accuracy requirement.

This model is obtained in the following way. We partition the measured samples into a training sample set $\{\mathbf{x}_i, Y_i\}_{i=1}^{n^{Tr}}$, and a testing sample set $\{\mathbf{x}_i, Y_i\}_{i=n^{Tr}+1}^{n^{Tr}+n^{Te}}$. Based on the training sample set, the predicted profile at an unobserved location $\mathbf{x}$ is obtained by the ordinary kriging predictor as:

$$\hat{\boldsymbol{Y}}(\mathbf{x}) = \hat{\beta}_0 + \boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\left(\boldsymbol{Y} - \hat{\beta}_0\mathbf{1}\right) \tag{4.9}$$

where $\mathbf{1}$ is a $n^{Tr} \times 1$ vector with all elements equal to 1; $\boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}} = [\psi(\mathbf{x} - \mathbf{x}_1)\ \psi(\mathbf{x} - \mathbf{x}_2)\ \ldots\ \psi(\mathbf{x} - \mathbf{x}_{n^{Tr}})]$; $\boldsymbol{\Psi}$ is a matrix with elements $\psi(\mathbf{x}_j - \mathbf{x}_k)$ in the row j and column k; $\mathbf{Y} = \left[Y_1\ Y_2\ \ldots\ Y_{n^{Tr}}\right]^{\mathrm{T}}$; and $\hat{\beta}_0 = \mathbf{1}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\mathbf{Y}/\mathbf{1}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\mathbf{1}$. The $\hat{\boldsymbol{Y}}(\mathbf{x})$ is the best linear unbiased estimator which interpolates all the measured locations (Santner *et al.*, 2003).

In the parameter estimation, the scale parameter $\boldsymbol{\Phi}$ is estimated by maximum likelihood estimation (MLE), denoted as $\hat{\boldsymbol{\Phi}}$. Then $\hat{\boldsymbol{\Phi}}$ is plugged in Equation (4.9) to calculate $\hat{\beta}_0$ (Santner *et al.*, 2003). In this way, a predicted profile is obtained by changing $\mathbf{x}$ in Equation (4.9). When there are additional samples collected in new iterations, the unknown parameters will be re-estimated, and new profile at interested locations can be predicted with the updated ordinary kriging model.

4.2.5 Determination of Sequential Samples

The samples are proposed to be measured sequentially, so that the samples collected at later measurement iterations can be appropriately selected based on the prior information from the GP model. If the stopping rule is not satisfied, the measurements of the (*i*+1)-th iteration is required based on the GP model in the *i*-th iteration. We propose to sample an empirical distribution, weighted by the magnitude of gradient and predicted Mean Squared Error (MSE) of the GP model.

Denote the predicted GP response as $\hat{Y}^i(\mathbf{x})$ in the *i*-th iteration, the gradient of the predicted GP as $d\hat{Y}^i(\mathbf{x})$ with the magnitude $|d\hat{Y}^i(\mathbf{x})|$, and the MSE at any location $\mathbf{x}$ as $Err_i(\mathbf{x})$, then we have $d\hat{Y}^i(\mathbf{x}) = \frac{\partial^p \hat{Y}^i(\mathbf{x})}{\partial \mathrm{x}_1 \partial \mathrm{x}_2 \ldots \partial \mathrm{x}_p}$, and $Err_i(\mathbf{x}) = \hat{\sigma}_Z^2\Big(1 + (\mathbf{1}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{\psi}(\mathbf{x}) - 1)^{\mathrm{T}}(\mathbf{1}\boldsymbol{\Psi}^{-1}\mathbf{1})^{-1}(\mathbf{1}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{\psi}(\mathbf{x}) - 1) - \boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{\psi}(\mathbf{x})\Big)$ (Lophaven *et al.*, 2002). Then the samples of the (*i*+1)-th iteration are sampled from the following empirical distribution:

$$\Pr(\mathbf{x}) = \frac{1}{\mathrm{c}_2}\left\{\lambda\left(\frac{|d\hat{Y}^i(\mathbf{x})|}{\mathrm{c}_3}\right) + (1-\lambda)\left(\frac{Err_i(\mathbf{x})}{\mathrm{c}_4}\right)\right\} \tag{4.10}$$

where λ is a weighting coefficient, which is a tuning parameter; c_2 is a normalizing constant for the distribution; c_3 and c_4 are the maximum values of the magnitude of gradient and prediction error, respectively, which are used to standardize the magnitude of gradient and prediction error. In Equation (4.10), the first part represents the area with large fluctuation, and the second part represents the area with larger prediction uncertainty. More samples in these two types of local areas should be measured to reduce the prediction error. Note that the prediction accuracy in a local region can be improved by taking additional measurements. This is because the ordinary kriging model

interpolates these extra measurements. Recall that when two sampled locations are closer to each other, their correlation may become higher. In addition, when larger sample size is affordable, the maximum distance between any two sampled locations could be reduced. In this way, higher prediction accuracy could be achieved, when there are more samples measured in the neighborhood of that location.

In practice, the distance between measurements should be larger than a minimal distance to avoid singularity problem when computing the inverse of the correlation matrix. In other words, the new samples from sampling locations will not be measured if they are too close to previous ones.

4.2.6 Stopping Rule

The sequential measurement strategy takes samples sequentially until a stopping rule is satisfied to achieve the required estimation accuracy. In most cases, the root mean sum of prediction error (RMSPE) of a profile can be adopted to evaluate the overall profile prediction accuracy, which is defined as:

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y(\mathbf{x}_i) - \hat{Y}(\mathbf{x}_i)\right)^2} \tag{4.11}$$

where $\hat{Y}(\mathbf{x}_i)$ is the predicted profile at the location $\mathbf{x}_i$ from the estimated GP model; and n is the number of measurements compared.

It is ideal to evaluate the RMSPE of samples, which are not used in modeling. In order to estimate the RMSPE of the overall wafer profile, we compute the testing error based on the testing sample set collected in each measurement iteration to determine if the measurement stops. The measurement will stop if

$$RMSPE_i^{test} \leq \sqrt{Th_{\sigma^2}} \tag{4.12}$$

where Th_{σ^2} is a pre-determined estimation variance, representing the profile accuracy requirement. In this chapter, it is determined from the tolerance of the quality variables or quality profiles. $RMSPE_i^{test}$ is the root mean sum of the testing sample set in the *i*-th measurement iteration.

4.2.7 Parameter Estimation

In the proposed method, there are several parameters to be determined: the initial sample size n_0, the sequential sample size n_i, and the weighting coefficient λ in Equation (4.10). In this chapter, these parameters are selected before the sequential measurement strategy is implemented online, based on a "golden" profile. The "golden" profile is regarded as a representative profile to a batch of the profiles. In the wafer example, the wafer profiles from the *same batch* are assumed to follow the same distribution due to the similarity of the process conditions. A "golden" profile is selected from one of the representative wafers, where the measurements at all possible potential measurement locations are obtained. The parameters are determined when estimating of the "golden" profile by the sequential measurement strategy.

The initial sample size n_0 is firstly determined by manipulating n_0 and comparing RMSPEs in the "golden" profile. More specifically, we draw n_0 samples using Equation (4.4) from D in the "golden" profile for N times, denoted as $[\mathbf{x}^{n_0}, Y^{n_0}]^1, [\mathbf{x}^{n_0}, Y^{n_0}]^2, \cdots, [\mathbf{x}^{n_0}, Y^{n_0}]^N$. Based on these samples, N GP models are estimated and their RMSPEs of the unmeasured samples are calculated. We accept the initial sample size as the minimal sample size with $M_{RMSPE} < T\sqrt{Th_{\sigma^2}}$, where M_{RMSPE} is the median of the RMSPEs of N GP models. T is a properly selected constant to have a reliable initial GP model for additional samples. If T is large, n_0 will be small. The

estimated initial GP model may have large variation in estimation, and the additional samples may not be reliable for quick approximation of the geometric profile. If T is small, n_0 will be large. It may take much time to measure many samples to ensure the unnecessary initial accuracy.

After n_0 is determined, n_i and λ are determined to minimize the composite index defined in Equation (4.1). Here, we assume the additional sample sizes n_i are the same in all measurement iterations. Following the sequential measurement strategy, we estimate the composite index for different combinations of n_i and λ based on the "golden" profile. In the strategy, we applied the same Th_{σ^2}. Therefore, the combination of n_i and λ with the smallest composite index yields the best measurement efficiency, and it will be selected as the parameters for the sequential measurement strategy.

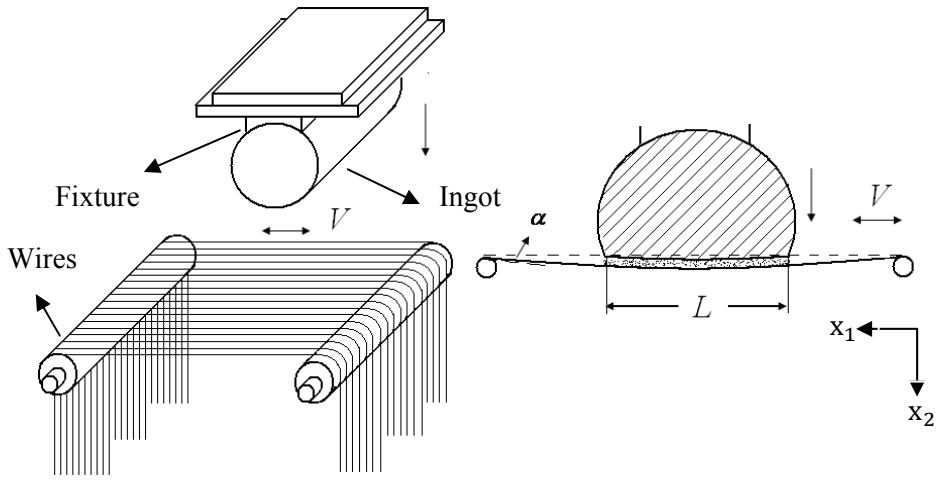
4.3 Case Study

A real case study is conducted to predict the wafer thickness profiles in a slicing process. Detail procedures are provided in this section to illustrate the effectiveness of the proposed sequential measurement strategy.

4.3.1 Wafer Slicing Processes

A slicing process is used to transform the silicon ingot into wafers with rough surface and non-uniform thickness. Figure 4.3 (a) shows a set up of wafer slicing process. The ingot is mounted to a fixture and pressed against multiple tensioned and equally spaced wires. The wires are moving back-and-forth with the given speed *V*, while the slurry sprinkles onto the cutting edge. Figure 4.3 (b) is an illustration of cutting edge and the contact span interaction. In the slicing process, the ingot is pressed against the wires such that

there is a bow angle α formed between the wire and the horizontal level. The length that silicon material has contact with the wire is called contact span, defined as L. During the slicing process, the slurry thin film is formed between the wire and ingot, in which the abrasives remove the silicon material.



(a) An ingot in slicing processes (b) Cutting edge and contact span

Figure 4.3: Slicing processes

The sliced wafer determines the initial geometric quality in a wafer production. The profile is significant for wafer monitoring and root cause diagnosis in slicing processes as well as downstream stages. One important wafer quality feature is the thickness profile, which represents the thickness over the wafer disks. The thickness profile data set is further used to estimate the TTV of the wafers, i.e., the difference between the maximum and the minimum of the thickness profile.

In this case study, a 6-inch ingot with over 400 wafers is sliced in a HCT wire saw system. Several steps are taken into actions to ensure the representativeness of the wafer profiles collected as the normal production: (1) the ingot is sliced with another 6-inch ingot with the same technology specification and very similar length, as the normal

production. The wafers in one ingot are measured for the case study. (2) The system is checked for the wires, guide wear and other maintenance check points to avoid abnormal machine statuses. And (3) the slurry characteristic is measured and checked to ensure a satisfied slicing efficiency.

After the slicing process, the thickness profiles are measured using a gauge with a touching probe. First, the wafers are loaded to a conveyer belt sequentially. Each wafer will pass through a measurement area, where the readings of thickness at all possible locations in D are recorded. The measurement time for each wafer is over 60 seconds, and it will take over 8 hours to measure all 400 wafers. Therefore, 71 wafers are selected from the whole ingots for this case study. For each wafer, the wafer thickness profile is stored in a three column matrix, where the first and the second columns represent the coordinates of the measurement locations, and the third column represents the wafer thickness.

4.3.2 Parameter Determination in the Case Study

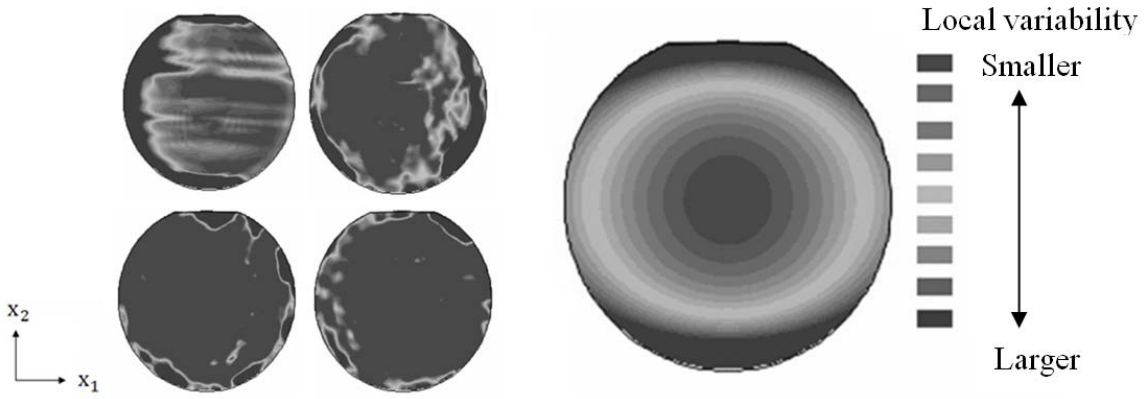
The profile data of the 71 wafers are used to evaluate the sequential measurement strategy. The measurement strategy will be used to select a subset of the data on each profile to fit GP models, which mimic the measurement procedure in practice. The thickness at each location in D and the TTV for each wafer will be predicted by the final GP model when the stopping rule is satisfied. These predicted values will be treated as real measurements if the accuracy requirement is satisfied. Since both the thickness and the TTV of each wafer are also measured, the predicted thickness and TTV by GP models will be compared with the measured thickness and TTV to evaluated measurement

strategies. For these 71 wafers, one wafer is selected as the "golden" profile and the rest 70 wafers are used for evaluation.

There are many potential factors for thickness variation, such as slicing speed V, contact span L and bow angle α, as marked in Figure 4.3 (b) (Zhu and Kao, 2005). From engineering knowledge and data collected, a partial proportional relationship is obtained among those variables, which shows a larger variation of thickness profile when the contact span is shorter or the location is nearer to the edge of a wafer, denoted as

$$\sigma_Y^2(\mathbf{x}) \propto \frac{r^t(\mathbf{x})}{L^t(\mathbf{x})} \tag{4.13}$$

where r is the radius from the location $\mathbf{x}$ to the center of the wafer; L is the contact span of the location $\mathbf{x}$; and this case study assumes $t_i = t_j = t$. The unknown parameter t is estimated from the local variability of the "golden" profile by MLE, which is $\hat{t} = 1.97$.



(a) Local variability (b) Fitted proportional relationship

Figure 4.4: Local variability (nearest 25 points) and fitted proportional relationship

Figure 4.4 (a) shows different local variability patterns estimated from the wafer profiles by Equation (4.2). It is clear that there is a large variation at the location where the contact span is short or the radius is large. Based on these common characteristics,

we use the proportional relationship in Equation (4.13) to capture these variation patterns, and use the "golden" profile to calibrate the parameter *t*, as shown in Figure 4.4 (b).

Following the procedure in Section 4.2.7, we determine the parameters based on the "golden" profile. In this case study, we choose *T*=3 to determine the initial sample size n_0. In this case, $n_0 = 100$. Then we obtain the composite index versus the λ and n_i by analyzing the "golden" profile. In this case study, we choose $\tau = 0.5$ by assuming the measurement time for a batch of samples and model fitting time for these samples are comparable in iterations. Thus, when $n_i = 70$ and $\lambda = 0.8$, the composite index is minimized. These values will be used for the simulation in the case study.

4.3.3 Performance Analysis and Comparison

After the parameters are determined, a series of GP models is estimated and the samples are determined in iterations. Figure 4.5 shows the intermediate results of sequential measurement procedure for the thickness profile prediction. The initial sampling distribution is weighted by radius and contact span, shown in Figure 4.5 (a). Based on the initial sample distribution, $n_0 = 100$ initial samples are measured, whose locations are shown in Figure 4.5 (b), marked as stars. In iterations, we partition all of the measured data into a training sample set (75%) and a testing sample set (25%). A GP model is estimated based on the training sample set (75 samples for the initial model fitting). In this model, the mean of the thickness profile is removed before the modeling. Figure 4.5 (c) shows the measured thickness profile (solid lines) and the estimated profile by the GP model ("+" lines), and the GP model is:

$$\hat{Y}(\mathbf{x}) = -0.3011 + Z(\mathbf{x}) \tag{4.14}$$

where $\widehat{\boldsymbol{\Phi}} = [5.95 \quad 2.50]$; $\sigma_Z^2 = 2.79$; and $RMSPE_0^{test}$ is 0.8662 [micron]. From Figure 4.5 (c), the GP model provides a nice approximation of the overall wafer profile. However, a larger prediction error can be observed at some locations (marked by a dashed circle), the prediction is not accurate enough. Additional samples may be needed to further reduce the prediction errors.

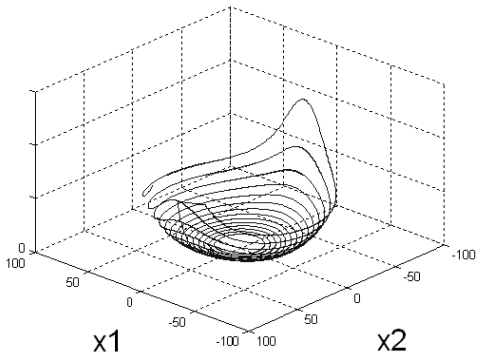
When additional measurements are needed, the magnitude of gradient and predicted MSE are estimated based on the GP model in Equation (4.14), shown in Figure 4.5 (d) and (e), respectively. Finally, the empirical distribution for sequential samples is obtained in Figure 4.5 (f). Additional $n_i = 70$ points are measured by sampling this empirical distribution. During the sampling, the Euclidean distance of any two measurement points is calculated. If the distance is smaller than a pre-determined threshold, the new samples will not be measured such that any two samples are not too close to each other. In this case study, the distance threshold is 3 [mm].

Once additional samples are determined, all measured 170 samples are randomly partitioned into a training sample set (75%) and a testing sample set (25%) again. The GP model is updated based on the training sample set as:

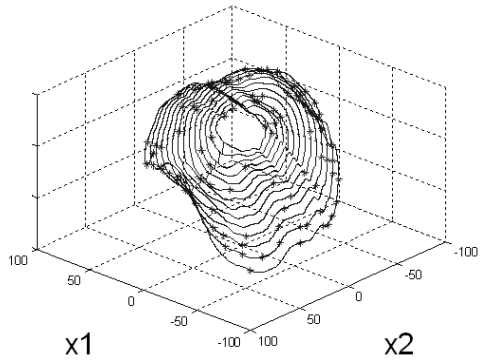
$$\hat{Y}(\mathbf{x}) = -0.4808 + Z(\mathbf{x}) \tag{4.15}$$

where $\widehat{\boldsymbol{\Phi}} = [4.20 \quad 5.00]$; $\sigma_Z^2 = 4.96$; and $RMSPE_0^{test}$ is 0.8389 [micron]. In this way, the GP model is improved as the sample size increases.

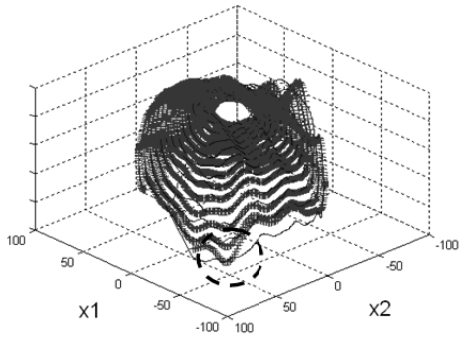
The sequential samples are measured iteratively until the accuracy of the estimated quality variables satisfies the stopping rule. In this chapter, $Th_{\sigma^2} = 0.04$, i.e., the requirement in the standard deviation is 0.2 [micron]. This accuracy requirement is an engineering specification of design in wafer manufacturing processes.



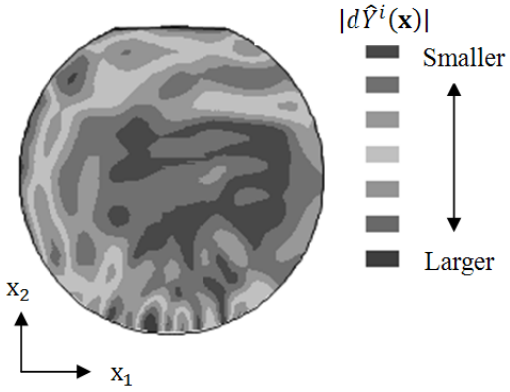
(a) Initial empirical distribution



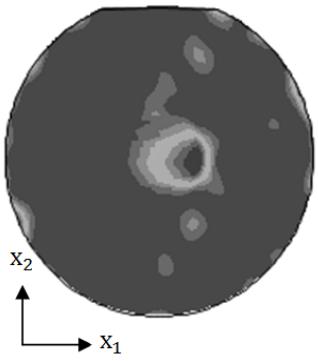
(b) Initial samples



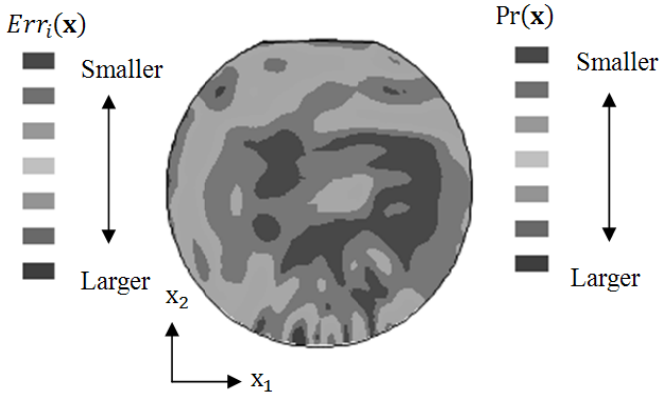
(c) GP model and true profile



(d) Magnitude of the gradient



(e) Predicted MSE



(f) Empirical distribution of seq. sample

Figure 4.5: Intermediate results of sequential measurement strategy

Once the stopping rule is satisfied, the TTV can be estimated as:

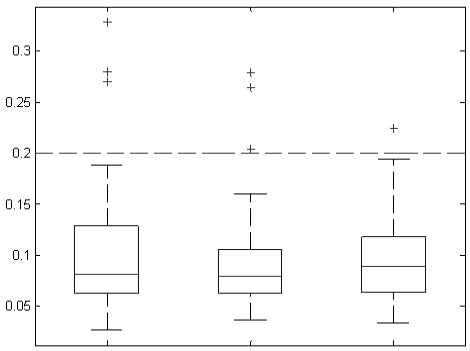
$$\widehat{TTV} = \max\left(\hat{Y}(\mathbf{x})\right) - \min\left(\hat{Y}(\mathbf{x})\right) \tag{4.16}$$

To evaluate the sequential measurement performance, the thickness profiles of 70 wafers are measured based on three different measurement strategies: random measurement strategy (denoted as "Rand."), sequential measurement strategy with space filling initial measurements (denoted as "Space-seq."), and sequential measurement strategy with initial measurements from engineering knowledge (denoted as "Eng.-seq."). In the random measurement strategy, the measurement locations are randomly selected following a discrete uniform distribution. The Euclidean distances of samples are calculated to reject samples in close neighborhood. The sampling process of random measurement is also completed in a sequential way with the same sample size and stopping rule employed.

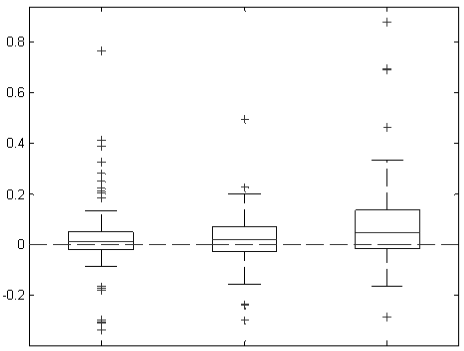
The "Space-seq." measurement strategy is different from the "Eng.-seq." measurement strategy in initial measurements. In the "Eng.-seq." measurement strategy, an Engineering driven initial empirical distribution is used to determine the initial samples. In the "Space-seq." measurement strategy, n_0 space filling initial measurements are measured. Then additional samples are measured using the same way as the "Eng.-seq." measurement strategy. Because some of the typical designs, such as LHD, may not be directly used in an irregular region, the initial measurements of the space filling design are determined in the following way: (1) the sample size is allocated to circles, which is proportional to the radiuses of the circles; (2) the samples on the outmost circle are selected as equally spaced samples; (3) the samples on the second outmost circle are also equally spaced, but a max-min criterion is applied to maximize the minimal distance to the samples on the circles outside; and (4) repeat Step 3 for the circles with smaller radiuses, until all samples are selected.

Based on these three measurement strategies, the results are summarized in Figure 4.6. Here the RMSPE refers to the RMSPE of all unmeasured locations, i.e., the locations not selected by measurement strategies. The RMSPE of the unmeasured locations are used to quantify how well the GP models approximate the profile. Figure 4.6 (a) and (b) represent the box-plots of RMSPE and $\widehat{TTV}$ deviation for 70 sliced wafers using those three strategies. The $\widehat{TTV}$ deviation refers to the deviation between the calculated $\widehat{TTV}$ based on the final GP models and the measured TTV of the wafers. Since the stopping rule sets the same standards in the estimation accuracy of the profile, we have comparable accuracy performance for three strategies in the RMSPE and $\widehat{TTV}$ deviation. However, to achieve the comparable estimation accuracy of the profile, both "Space-seq." measurement strategy and "Eng.-seq." measurement strategy use less samples, as shown in Figure 4.6 (c), and they have smaller composite indexes as shown in Figure 4.6 (d).

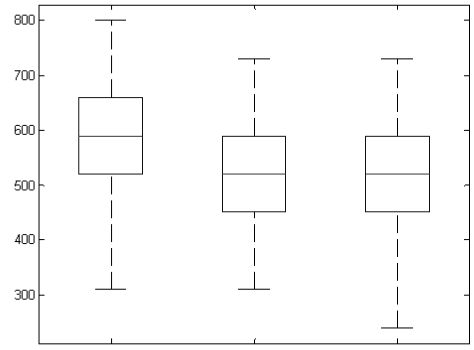
Moreover, the RMSPEs with different sample sizes are compared in Figure 4.6 (e). It is clear that the RMSPEs of both "Space-seq." measurement strategy and "Eng.-seq." measurement strategy yield better estimation performance than the random measurement strategy. The "Eng.-seq." measurement strategy has better estimation performance when sample size is small, but quickly converges to the similar performance as the "Space-seq." measurement strategy does. This result indicates that the initial empirical distribution provides useful information to obtain a reliable initial GP model for sequential measurements. The sequential measurement strategy performs well, even if the initial engineering knowledge is not available and the space filling initial measurements are used instead.



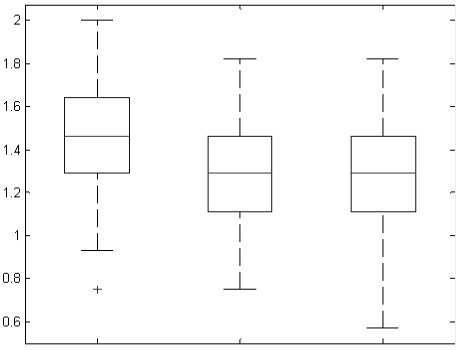
(a) RMSPE



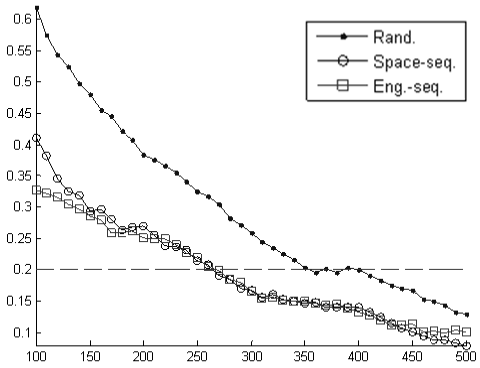
(b) $\widehat{TTV}$ deviation



(c) Sample size



(d) Composite index



(e) RMSPE V.S. sample size

Figure 4.6: Performance measure for three measurement strategies

4.4 Conclusion

Wafer geometric profiles are important quality features in semiconductor manufacturing. In most cases, the measurements of wafer profile are not available during the production, since it is time consuming to measure profiles of a large batch of wafers.

This chapter proposes an efficient sequential measurement strategy to approximate the thickness profile by estimated GP models. New empirical distributions are proposed to determine measurement locations, include both the initial distribution from engineering knowledge, and the sequential measurement distribution from the estimated GP models. In this chapter, the case study indicates that proposed sequential measurement strategy requires smaller sample size to achieve comparable estimation accuracy than the random measurement strategy. Moreover, the initial empirical distribution contributes in obtaining a reliable initial GP model, when comparing with the space filling measurement strategy.

In the GP model estimation, the computation complexity is intensive when the training sample size becomes larger, and the inversion of the covariance matrix may easily become ill-conditioned. In future research, computationally more efficient meta-models will be studied to develop new measurement strategies.

CHAPTER 5

MULTISTAGE MULTIMODE PROCESS MONITORING BASED ON A PIECEWISE LINEAR REGRESSION TREE CONSIDERING MODELING UNCERTAINTY

5.1 Introduction

A multistage manufacturing process (MMP) consists of multiple operations at different stages to produce one product. The variation of the product quality is introduced by the operations at the current stage as well as those propagated from upstream stages. The accumulation of variation may result in nonconforming products at the end of the MMP. Therefore, a quick detection of changes in a multistage process is important for quality assurance and improvement.

The monitoring of a multistage process is a challenging problem, because the variables and stages are interrelated. The output of a stage becomes the input of the next stage. The conventional statistical process control (SPC) cannot be directly applied in this case, because it monitors the final product quality without consideration of the inter-stage relationships. Thus, it is difficult to identify the stages with assignable causes. More discussions on the topic can be found in the review papers of Lowry and Montgomery (1995), and Woodall and Montgomery (1999).

Considering the variable relationship, a regression based risk-adjusted approaches (Hawkins, 1991, 1993; Shu *et al*., 2004a) and cause-selecting methods (Zhang, 1985, 1992; Shu *et al*., 2004b) are developed. In principle, both types of approaches use regression models to predict the quality variables at the current stage, but differ in the

following way: The risk adjustment method regresses the quality variables on the process and upstream quality variables. It monitors the residuals and the covariates, thus distinguishing the process change at the current stage or that from the upstream stages. The cause-selecting method predicts the quality variables based only on the quality variables at the previous stage. By monitoring the residuals, a process change is identified for the current stage.

Beyond the statistical model based adjustment, multistage process monitoring is also developed based on engineering models (Xiang and Tsung, 2008). In this approach, a state space model is estimated using EM algorithm to model the stream of variation (Jin and Shi, 1999; Shi, 2006). The one-step-ahead prediction error of quality observations is monitored using a "group EWMA chart."

The existing approaches for multistage process monitoring are successful to quickly detect the process change by assuming only one operational mode or one baseline model under normal conditions. However, this may not be true in a complex manufacturing process. A process may involve different incoming material properties, production flows, machines, process settings, etc. Due to the active compensation capability of the process, the manufacture produces conforming product, indicating an in-control process, but the variation propagation differs in path to realize the conforming final product under the normal conditions. The multimode shows different clusters of data at each stage, with multiple variation propagation patterns among stages using different baseline models.

In recent years, process monitoring problem with multiple clusters under normal conditions becomes important. In general, these clusters are identified first and then the

monitoring is carried out within each cluster. When multiple operational modes are observed in the variable space, Hwang and Han (1999) developed a hierarchical clustering and a super PCA model for real time process monitoring. Similarly, when the time-ordered clusters are observed, Harnish *et al.* (2009) used a modified agglomerative clustering approach to identify multiple change points along the time index; Jobe and Pokojovy (2009) proposed a distance based clustering method in the transformed space, and charting individual clusters. Furthermore, Zhang and Albin (2007) proposed a scale-based clustering with dummy dimension to identify the number of clusters.

The aforementioned process monitoring approaches with multimode provide great opportunities when the clusters have already been identified in either variable space or time domain. However, these approaches focus on monitoring variables with clusters at a single stage and ignore the variable relationship in a MMP. When the variables are interrelated in a MMP, monitoring all variables simultaneously may not effectively detect process changes at certain stages. Moreover, multiple operational modes essentially determine different propagation patterns among stages, which is not modeled nor monitored in the existing approaches. Therefore, there is a gap between current methodologies and the monitoring of a multistage multimode process (MMOP).

A MMOP is commonly encountered in practice. For example, in a wafer manufacturing process, wafers are lapped to improve thickness uniformity, coated with thin films in the chemical vapor deposition (CVD) process, and polished to achieve mirror-like surfaces. Figure 5.1 shows the real measurements of the wafer thickness from two ingots after lapping, CVD, and polishing operations. The left box plot at each stage refers to Ingot 1, and the right box plot at each stage refers to Ingot 2. The circles and

crosses represent the means of the thickness for these two ingots. In this wafer manufacturing process, the lapped wafers have different thickness for those two ingots, and the polishing process adjusts the process setting variables, such as pressure, rotation speed, or time, to compensate incoming thickness variation and achieve conforming wafers in the end. In this case, the thickness of wafers evolves in different paths as the operations move on, which forms a MMOP.

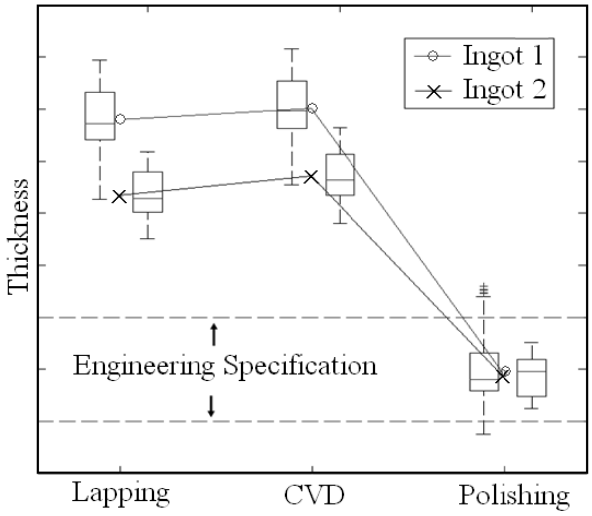


Figure 5.1: A multistage multimode wafer manufacturing process

In this chapter, we propose a piecewise linear regression tree (PLRT) based risk adjustment method to address the monitoring problem in a MMOP. By partitioning the variable space into several sub-regions, local regression models of PLRTs are used to represent different propagation patterns related to the quality, process, and material property variables stage-by-stage (Loh, 2002; Kim *et al*., 2007; Jin and Shi, 2011). The objective is to monitor the mean shift of the residuals for these local models. The proposed approach is based on the following assumptions: (1) the multimode operations are revealed by multiple baseline models in variable spaces; (2) the variation propagation under different operational modes is well approximated by piecewise linear models; and

(3) during the process monitoring, the operational modes remain stable, and there are no changes in modes nor baseline model structures.

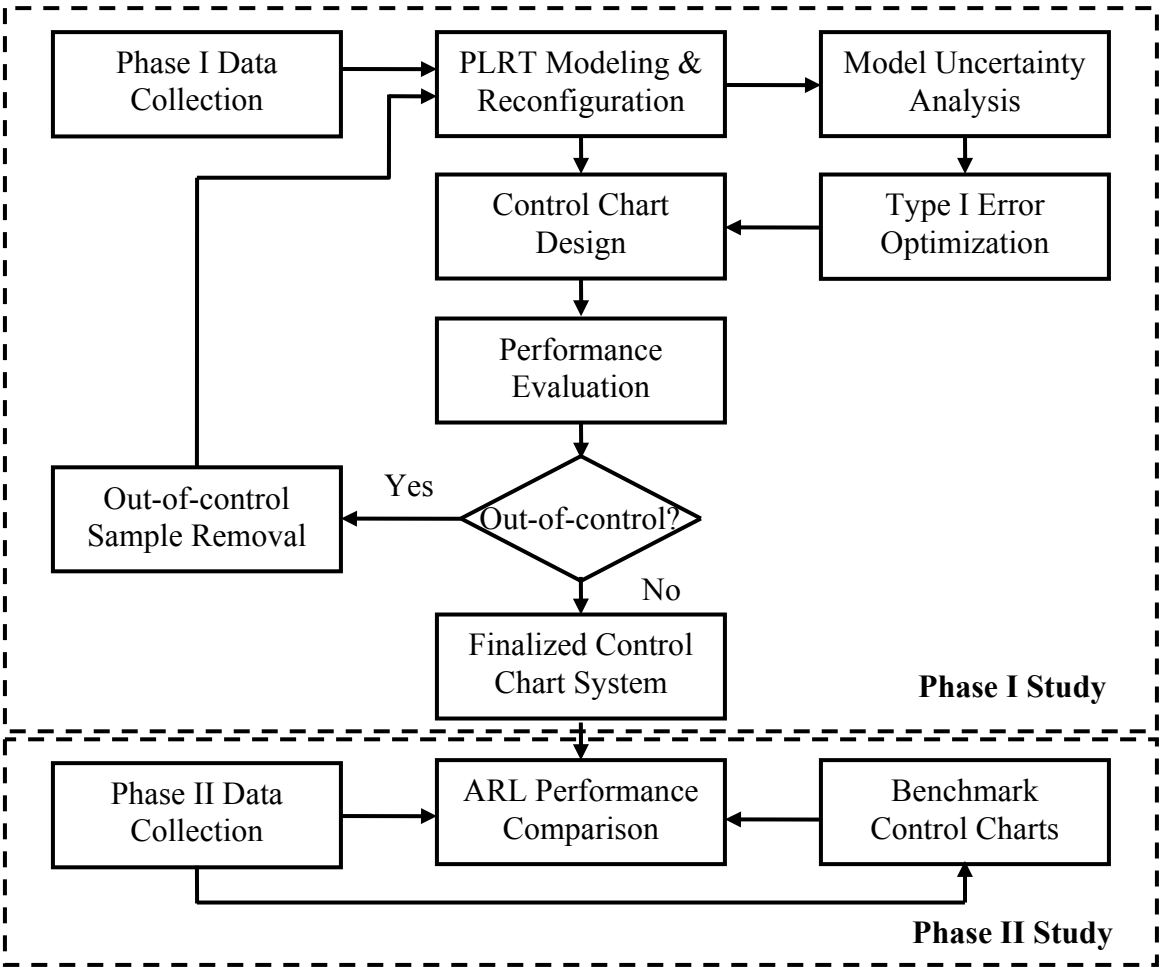


Figure 5.2: An overview of proposed method

An overview of the proposed method is shown in Figure 5.2. In this framework, we model the MMOP by the PLRTs and reconfigure the trees into engineering compliant model based on the Phase I data (Jin and Shi, 2011). In the process monitoring, we use the Shewhart charts to monitor the residuals of every local models in the PLRTs, which form a control chart system. We study the impact of the model uncertainty to the run length distribution, and further optimize the control chart system. We remove out-of-control samples with assignable causes iteratively, until no more out-of-control samples

show up in the finalized control chart system. At this moment, the control chart system is used to monitor the future samples, and the control chart performance is compared with other benchmark control charts.

The rest part of the chapter is organized as follows. After the introduction, Section 5.2 discusses the modeling procedure of the PLRTs for MMOPs. Section 5.3 presents the development of the control chart system for risk adjustment based on local models. In Section 5.4, we analyze the run length distribution considering modeling uncertainty for the control chart system optimization. Section 5.5 presents the case studies of simulation models and a real example in a multistage wafer manufacturing process to demonstrate the control chart performances. Finally, the conclusion is provided in Section 5.6.

5.2 Piecewise Linear Regression Tree for Multistage Multimode Process Modeling

In a MMP with N stages, the operations of a product is completed stage-by-stage, shown in Figure 2.2 in Chapter 2. Typical variables to describe a MMP are the product quality variables, process setting variables, and material property variables, defined in Table 3.1 in Chapter 3. In this process, some of the quality variables are measured repeatively at the end of each stage.

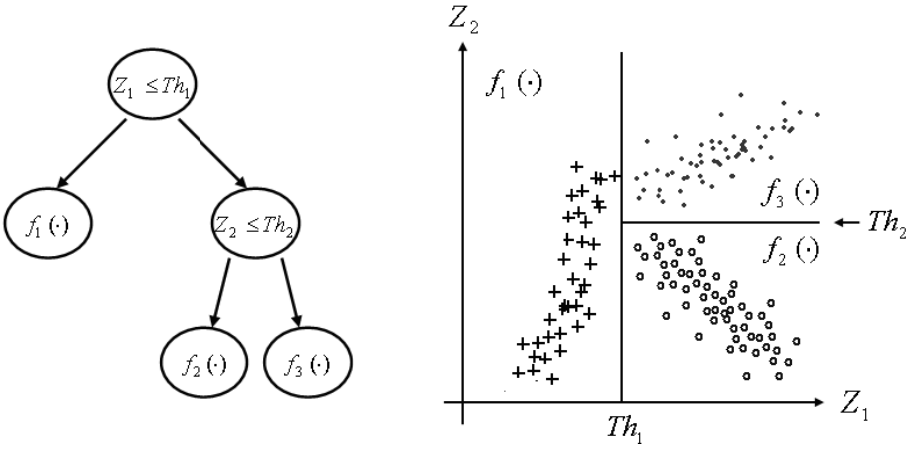
The discovery of multimode in a multistage manufacturing process is a challenging task. This is not only because both number of modes and mode conditions are unknown, but also because the variables are interrelated to form different baseline models. The proposed PLRT is assumed to have a good match of the MMOP in the following three aspects: (1) the number of leaf nodes of a PLRT reveals the number of modes in a multistage process; (2) the splitting conditions for each leaf nodes represent

the mode conditions; and (3) the local regression models in different leaf nodes of a PLRT are the baseline models of different modes.

By using PLRTs, we identify the number of modes by the number of leaf nodes, quantify the mode conditions by the splitting conditions, and determine the model structures and parameters in the local models. Figure 5.3 shows an example of a PLRT and its partitioned variable space (redrawn from Jin and Shi, 2011). There are three leaf nodes in Figure 5.3(a) and corresponding sub-regions in their variable space in Figure 5.3(b). Two splitting variables Z_i $(i = 1,2)$ and corresponding splitting boundaries Th_i $(i = 1,2)$ quantify these leaf nodes. The selection of model goes to the left branch if the splitting condition is satisfied. In each leaf node, a local regression model $f_i(\cdot)$ $(i = 1,2,3)$ predicts the response.

Multimode in a MMP is usually defined from an engineering perspective, either by different raw material, operational conditions, or production flows. In this chapter, we use a data driven approach to identify multiple baseline models to describe variable relationships. In this way, we monitor the process changes from the baseline conditions characterized by the statistical models. After the PLRT is constructed, we may compare the operational modes with splitting variables to explain the scenarios. For example, in Figure 5.3, we treat $f_i(\cdot)$ $(i = 1,2,3)$ as three baseline models in the variation propagation. When $Z_1 < Th_1$, the mode with baseline model $f_1(\cdot)$ represents the current scenario of production.There are different types of methods to construct a tree model, such as Classification and Regression Tree (CART) (Breiman *et al.*, 1984), Bayesian Tree (Chipman *et al.*, 1998, 2002; Dennison *et al.*, 2002) and Smoothed and Unsmoothed

Piecewise-polynomial Regression Trees (SUPPORT) (Chaudhuri *et al*., 1994). This chapter uses Generalized, Unbiased Interaction Detection and Estimation (GUIDE) (Loh, 2002; Kim *et al*., 2007) because of its advantages in splitting, prediction, selection bias alleviation and interaction detection.



(a) A PLRT (b) Partitioned variable space with local models

Figure 5.3: A PLRT and its splitting variable space

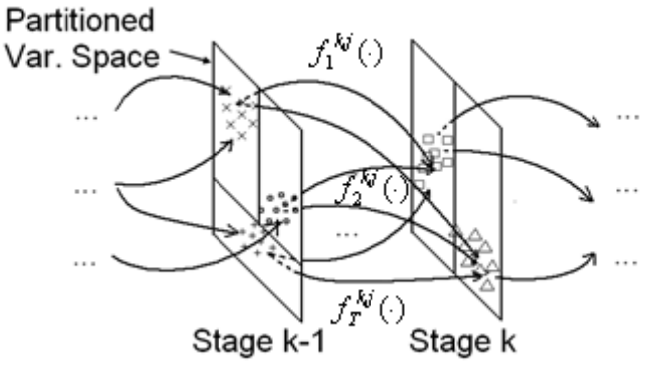


Figure 5.4: Multiple baseline models in variation propagation

To model a MMOP, we adopt the same procedure for modeling and reconfiguration as that used by Jin and Shi (2011). The i-th local model $f_i^{kj}(\cdot)$ is identified to predict the $\mathbf{Y}(k)_{j,t}$, the j-th quality variable at the k-th stage for the sample index t as

$$\mathbf{Y}(k)_{j,t} = \beta_{i,0}^{k,j} + (\boldsymbol{\beta}_{i,1}^{k,j})^{\mathrm{T}} \boldsymbol{\eta}_{k,t} + e_{i,t}^{k,j} \tag{5.1}$$

where the model is estimated based on the centralized data; $\beta_{i,0}^{k,j}$ and $\boldsymbol{\beta}_{i,1}^{k,j}$ are regression coefficients; $\boldsymbol{\eta}_{k,t}$ is the covariates vector for the *t*-th sample with dimension $\mathrm{p}_k \times 1$, where p_k is the number of the covariates; $e_{i,t}^{k,j}$ are the residual errors in $f_i^{kj}(\cdot)$; and *t* is the sample index. $e_{i,t}^{k,j}$ follows an i.i.d normal distribution with mean 0 and variance $\sigma_{e_{i,t}^{k,j}}^2$. Figure 5.4 shows multiple baseline models in the variation propagation after the reconfiguration where the baseline models are represented as arrows. In each stage, the variable space is partitioned into different regions with different data clusters, such as circles, crosses, and squares in Figure 5.4. To predict the quality variables at the k-th stage, there may be different baseline models $f_i(\cdot)$ $(i = 1,2,...,T)$ representing different variation propagation patterns. In this way, a PLRT quantifies the multimode of variation propagation among multiple stages.

5.3 Design of a PTO Control Chart System

After we obtain the PLRTs, we propose a risk adjustment type of control chart system to monitor the MMOP. We call the control chart system as "**P**iecewise linear regression **T**ree based control chart system with **O**ptimized type I error" (PTO). We discuss the structure of the PTO control chart system in this section.

The regression based risk adjustment approaches detect a mean shift in a MMP by charting the residuals (Hawkins, 1993). When a PLRT is constructed to model the variation propagation, multiple local regression models are developed. Therefore, a

group of Shewhart control charts is constructed for the residuals of each local model. The residuals and their estimated variance after the risk adjustment can be computed as

$$\begin{cases} \hat{e}_{i,t}^{k,j} = \mathbf{Y}(k)_{j,t} - \hat{\beta}_{i,0}^{k,j} - (\hat{\boldsymbol{\beta}}_{i,1}^{k,j})^{\mathrm{T}} \boldsymbol{\eta}_{k,t} \\ \hat{\sigma}^2_{e_{i,t}^{k,j}} = \mathrm{var}(\hat{e}_{i,t}^{k,j}) \end{cases} \quad if \;\; I(g_i^{k,j}(Z_1,...,Z_L)) = 1 \qquad (5.2)$$

where $g_i^{k,j}(Z_1,...,Z_L)$ is the splitting condition based on splitting variables $Z_1,...,Z_L$; L is the number of splitting variables; and $I(x)$ is an indicator function, which is

$$I(x) = \begin{cases} 1 & if \; x \geq 0 \\ 0 & if \; x < 0 \end{cases} \qquad (5.3)$$

In Equation (5.2), the coefficients of the regression model $\beta_{i,0}^{k,j}$, $\boldsymbol{\beta}_{i,1}^{k,j}$ and $\sigma^2_{e_i^{k,j}}$ are unknown in most cases, and $\hat{\beta}_{i,0}^{k,j}$, $\hat{\boldsymbol{\beta}}_{i,1}^{k,j}$ and $\hat{\sigma}^2_{e_{i,t}^{k,j}}$ are the corresponding estimates based on $\mathrm{m_i}$ training samples, i.e., $t = 1,2,...,\mathrm{m_i}$.

Without loss of generality, we choose the sample size for process monitoring as 1. Thus, we monitor the standardized residual $\dfrac{\hat{e}_{i,t}^{k,j}}{\hat{\sigma}_{e_{i,t}^{k,j}}}$ ($i = 1,2,...T_{k,j}$) and compare it with the control limits as

$$\begin{cases} UCL = z_{\alpha/2} \\ CL = 0 \\ LCL = -z_{\alpha/2} \end{cases} \quad if \;\; I(g_i^{k,j}(Z_1,...,Z_L)) = 1 \qquad (5.4)$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$-th quantile of a standard normal distribution. When the control chart for the i-th model signals, we conclude that there is a mean shift in the residuals, i.e., $E(e_{i,t}^{k,j}) = \Delta_i^{k,j} \neq 0$.

When using the PTO control chart system to monitor a process, only one local model will be selected as the risk adjustment model for a sample based on the splitting conditions. Therefore, we have different concepts of run length in both the individual control chart and in the control chart system. Here we refer "a run in a control chart" as a sample that enters one individual control chart, given that this sample satisfies the splitting conditions. And we refer "a run in a control chart system" as a sample that enters the control chart system, without specifying the splitting conditions.

Two concepts of runs refer to monitoring at an individual control chart level and a control chart system level. Taking a PLRT with two leaf nodes in a MMOP as an example, the probability that a new sample satisfies the splitting conditions to the first and the second leaf node is q_1 and q_2, respectively. When there are m runs in the control chart system, there are $\mathrm{m}\,q_1$ and $\mathrm{m}\,q_2$ expected runs in these two control charts. Therefore, the conversion of the run length in a control chart to that in a control chart system depends on the probability to select the control chart in the PLRT.

5.4 Model Uncertainty Analysis for Control Chart System Optimization

In the regression based risk adjustment methods, the regression parameter estimation has an impact on the monitoring performance (Shu *et al.*, 2004). This is because that the parameters are estimated from a training data set, involving uncertainties from sensing noise, sampling strategy, or natural variability of the process. The uncertainties of the estimates will impact the control chart performance, such as excessive false alarms. Similar to the risk adjustment methods based on a single regression model, the PLRT based control chart system also suffers performance degradation considering modeling uncertainty, but in a more complicated way.

There are two types of modeling uncertainties involved to change the Average Run Length (ARL) performance in a PLRT based control chart system: (1) different training data sets may result in different partitions of the sub-regions in the variable space, including both the splitting variable selection and splitting boundary estimation; and (2) parameter estimation has uncertainty within a sub-region.

Moreover, the two types of modeling uncertainties are inter-related with each other, because the PLRT partitions the sample space and fits the local models recursively. Given a regression model, the uncertainty of the parameter estimates transfers to the uncertainty of contingence table tests in splitting variable selection and MSE estimates of splitting boundaries (Loh, 2002). This newly partitioned variable space further changes the parameter estimates in each sub-region. The relationship of the two types of modeling uncertainties makes the assessment of a PLRT uncertainty a very challenging problem.

Considering the two types of modeling uncertainties, the control chart performance is degraded, not only because of the parameter estimates in each baseline model, but also the selection of baseline model for risk adjustment. In the Phase II study, a new sample near the splitting boundaries may be misclassified for the risk adjustment based on an incorrect local model due to the estimation uncertainty of splitting boundaries. From a risk adjustment point of view, the variance of residuals in local models may be quite different, which results in quite different performance of the ARL.

In order to tackle the mentioned problem, we transfer the modeling uncertainty to run length distribution by conditioning the splitting uncertainty. That is, we first analyze the run length performance considering the parameter estimation uncertainty in baseline

models, conditioning on a baseline model. Then, we try to integrate the splitting uncertainty into the analysis. This idea is motivated by Bayesian Tree (Chipman *et al.*, 1998, 2002), which provides a probability framework for both tree structure and parameter estimates.

In this chapter, our objective is to monitor the mean shift in baseline model residuals, and the operational modes are assumed to be stable. Therefore, we assume that the splitting structure, including the splitting variables and splitting orders will not be changed for a stable MMP, but the splitting boundary may be changed due to the data uncertainty. Therefore, we consider the effects of uncertainties from both the splitting boundary estimation and the baseline model parameter estimation.

5.4.1 Run Length Distribution Considering Parameter Estimation

Conditioning on a baseline model, e.g., the *i*-th model, the analysis of run length distribution with parameter estimates is similar to that of the run length performance of regression model based control charts (Shu *et al.*, 2004). In their paper, the effects of parameter estimation are investigated to monitor the standardized residuals when there is a mean shift in residuals or a covariate. In this chapter, we follow the same procedure for the analysis, and further extend their scheme from a single covariate to multiple covariates.

Based on Equation (5.1), the regression coefficients follow normal distributions as follows:

$$\hat{\beta}_{i,0}^{k,j} = \frac{\sum_{t=1}^{m_i} \mathbf{Y}(k)_{j,t}}{m_i} \sim N(\beta_{i,0}^{k,j}, \frac{\sigma_{e_{i,t}^{k,j}}^2}{m_i}) \tag{5.5}$$

$$\hat{\boldsymbol{\beta}}_{i,1}^{k,j} \sim MVN(\boldsymbol{\beta}_{i,1}^{k,j}, (\boldsymbol{\eta}_k \boldsymbol{\eta}_k^{\mathrm{T}})^{-1} \sigma_{e_{i,t}^{k,j}}^2)$$

where m_i is the training sample size for the *i*-th local model, and $\boldsymbol{\eta}_k$ is the data matrix of covariates with dimension $p_k \times m_i$.

Based on Equation (5.1) and Equation (5.2), a future observation with a mean shift $E(e_{i,t}^{k,j}) = \Delta_i^{k,j} = a\sigma_{e_{i,t}^{k,j}}$ in the residuals has $\hat{e}_{i,t}^{k,j} = (\beta_{i,0}^{k,j} - \hat{\beta}_{i,0}^{k,j}) + (\boldsymbol{\beta}_{i,1}^{k,j} - \hat{\boldsymbol{\beta}}_{i,1}^{k,j})^{\mathrm{T}} \boldsymbol{\eta}_{k,t} + e_{i,t}^{k,j}$, and the residual is a normal variable given the baseline model parameter estimates, i.e.,

$$\hat{e}_{i,t}^{k,j} \mid \hat{\beta}_{i,0}^{k,j}, \hat{\boldsymbol{\beta}}_{i,1}^{k,j} \sim N(E(\hat{e}_{i,t}^{k,j}), \hat{\sigma}^2_{e_{i,t}^{k,j}}) \tag{5.6}$$

where $E(\hat{e}_{i,t}^{k,j}) = \Delta_i^{k,j} + (\beta_{i,0}^{k,j} - \hat{\beta}_{i,0}^{k,j}) + (\boldsymbol{\beta}_{i,1}^{k,j} - \hat{\boldsymbol{\beta}}_{i,1}^{k,j})^{\mathrm{T}} \boldsymbol{\mu}_{\boldsymbol{\eta}_k}$ and $\hat{\sigma}^2_{e_{i,t}^{k,j}} = (\boldsymbol{\beta}_{i,1}^{k,j} - \hat{\boldsymbol{\beta}}_{i,1}^{k,j})^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\eta}_k} (\boldsymbol{\beta}_{i,1}^{k,j} - \hat{\boldsymbol{\beta}}_{i,1}^{k,j}) + \sigma^2_{e_{i,t}^{k,j}}$. Here $\boldsymbol{\mu}_{\boldsymbol{\eta}_k} = E(\boldsymbol{\eta}_{k,t})$, the expectation of the covariates with dimension $p_k \times 1$, and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_k} = Cov(\boldsymbol{\eta}_{k,t})$, the covariance of the covariates with dimension $p_k \times p_k$.

Therefore, the monitoring statistics in a Shewhart chart is

$$\frac{\hat{e}_{i,t}^{k,j}}{\hat{\sigma}_{e_{i,t}^{k,j}}} = \frac{1}{W}\left[Z_{\hat{e}_{i,t}^{k,j}} \sqrt{\mathbf{Z}_{i,A}^{k,j} (\mathbf{Z}_{i,A}^{k,j})^{\mathrm{T}} + 1} + a - \frac{Z_{i,0}^{k,j}}{\sqrt{m_i}} - Z_{i,1}^{k,j}\right] \tag{5.7}$$

where $W = \dfrac{\hat{\sigma}_{e_{i,t}^{k,j}}}{\sigma_{e_{i,t}^{k,j}}}$; $Z_{\hat{e}_{i,t}^{k,j}}$ and $Z_{i,0}^{k,j}$ are standardized normal random variables; $Z_{i,1}^{k,j}$ is a normal random variable with mean as zero and variance as $\boldsymbol{\mu}_{\boldsymbol{\eta}_k}^{\mathrm{T}} (\boldsymbol{\eta}_k \boldsymbol{\eta}_k^{\mathrm{T}})^{-1} \boldsymbol{\mu}_{\boldsymbol{\eta}_k}$; and $\mathbf{Z}_{i,A}^{k,j}$ is a $1 \times p_k$ normal random vector with mean as zeros and covariance matrix as $\mathbf{A}^{\mathrm{T}} (\boldsymbol{\eta}_k \boldsymbol{\eta}_k^{\mathrm{T}})^{-1} \mathbf{A}$. Here $\mathbf{A}$ is a $p_k \times p_k$ matrix, such that $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_k} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$. Since $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_k}$ is a positive semi-definite matrix, $\mathbf{A} = \mathbf{D}\mathbf{V}^{1/2}\mathbf{D}^{\mathrm{T}}$, where $\mathbf{D}$ is the matrix with eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_k}$ as

columns, and $\mathbf{V}^{1/2}$ is the matrix with the square roots of eigenvalues as diagonal elements. The detailed derivation and definitions can be found in the Appendix.

Because we use a Shewhart chart to monitor the residuals, and the samples are independent, the probability that the run length in a control chart equals t_0 is denoted as $q_{t_0,i}$, which is

$$q_{t_0,i} \triangleq \Pr(t = t_0 \mid Z_{i,0}^{k,j}, Z_{i,1}^{k,j}, \mathbf{Z}_{i,A}^{k,j}, W, i)$$

$$= [\Phi(\frac{z_{\alpha/2}W - (a - Z_{i,0}^{k,j} / \sqrt{\mathrm{m_i}} - Z_{i,1}^{k,j})}{\sqrt{\mathbf{Z}_{i,A}^{k,j}(\mathbf{Z}_{i,A}^{k,j})^{\mathrm{T}} + 1}}) - \Phi(\frac{-z_{\alpha/2}W - (a - Z_{i,0}^{k,j} / \sqrt{\mathrm{m_i}} - Z_{i,1}^{k,j})}{\sqrt{\mathbf{Z}_{i,A}^{k,j}(\mathbf{Z}_{i,A}^{k,j})^{\mathrm{T}} + 1}})]^{t_0 - 1} \cdot$$

$$[1 - \Phi(\frac{z_{\alpha/2}W - (a - Z_{i,0}^{k,j} / \sqrt{\mathrm{m_i}} - Z_{i,1}^{k,j})}{\sqrt{\mathbf{Z}_{i,A}^{k,j}(\mathbf{Z}_{i,A}^{k,j})^{\mathrm{T}} + 1}}) + \Phi(\frac{-z_{\alpha/2}W - (a - Z_{i,0}^{k,j} / \sqrt{\mathrm{m_i}} - Z_{i,1}^{k,j})}{\sqrt{\mathbf{Z}_{i,A}^{k,j}(\mathbf{Z}_{i,A}^{k,j})^{\mathrm{T}} + 1}})] \qquad (5.8)$$

where $\Phi(x)$ is the accumulative density function of a standard normal random variable.

5.4.2 Run Length Distribution Considering Splitting Uncertainty

After the run length distribution is analyzed conditioning on the *i*-th local model, we further analyze the impact of the splitting uncertainty to the run length distribution. We discuss the splitting uncertainty based on the types of the splitting variables, i.e., continuous variables and categorical variables. Furthermore, the conditions of continuous variables can be classified into two categories: the variable smaller than the splitting boundaries or the variable larger than the splitting boundaries. In this chapter, we denote $l \in Set$ A, if Z_l is a continuous variable smaller than the splitting boundaries; $l \in Set$ B, if Z_l is a continuous variable larger than the splitting boundaries; and $l \in Set$ C, if Z_l is a categorical variable.

Assuming the splitting order is $Z_1, Z_2, ..., Z_L$, i.e., given a splitting variable Z_l, all variables split prior to Z_l have a smaller l, the corresponding splitting boundaries for $l \in Set$ A or $l \in Set$ B are

$$\hat{Th}_l \mid g_{i,1}^{k,j}(Z_1), ..., g_{i,l-1}^{k,j}(Z_{l-1}) = Th_l + \tilde{Th}_l, l = 1, ..., L \tag{5.9}$$

where $\hat{Th}_l$ is the splitting boundary estimates from GUIDE, conditioning on the previous splits; Th_l is the true boundary for the baseline models; $\tilde{Th}_l$ is the estimation error; and $g_{i,l}^{k,j}(\cdot)$ is the splitting condition of the splitting variable.

Here, we further assume $E(\hat{Th}_l \mid g_{i,1}^{k,j}(Z_1), ..., g_{i,l-1}^{k,j}(Z_{l-1})) = Th_l$.

When Z_l is observed and Th_l is estimated from the data, for $l \in Set$ A, $g_{i,l}^{k,j}(Z_l) = Th_l - Z_l$, and the probability that a new sample satisfies the condition is denoted as p_l^-, which is

$$p_l^- \hat{=} \Pr(Z_l < \hat{Th}_l \mid g_{i,1}^{k,j}(Z_1), ... g_{i,l-1}^{k,j}(Z_{l-1})) = \Pr(Z_l < Th_l + \tilde{Th}_l \mid g_{i,1}^{k,j}(Z_1), ... g_{i,l-1}^{k,j}(Z_{l-1}))$$

$$= \Pr(\tilde{Th}_l > -g_{i,l}^{k,j}(Z_l) \mid g_{i,1}^{k,j}(Z_1), ..., g_{i,l-1}^{k,j}(Z_{l-1})) \tag{5.10}$$

For $l \in Set$ B, $g_{i,l}^{k,j}(Z_l) = Z_l - Th_l$, and the probability that a new sample satisfies the condition is denoted as p_l^+, which is

$$p_l^+ \hat{=} \Pr(Z_l > \hat{Th}_l \mid g_{i,1}^{k,j}(Z_1), ... g_{i,l-1}^{k,j}(Z_{l-1})) = \Pr(Z_l > Th_l + \tilde{Th}_l \mid g_{i,1}^{k,j}(Z_1), ... g_{i,l-1}^{k,j}(Z_{l-1}))$$

$$= \Pr(\tilde{Th}_l < g_{i,l}^{k,j}(Z_l) \mid g_{i,1}^{k,j}(Z_1), ..., g_{i,l-1}^{k,j}(Z_{l-1})) \tag{5.11}$$

For $l \in Set$ C, the splitting conditions are set operations, i.e., $Z_l \in C_l$, where C_l is a subset of its all possible values of Z_l as $\{c_{ls}\}_{s=1}^{S_l}$. When there is no estimation

uncertainty of C_l , we treat C_l as a fix subset of $\{c_{ls}\}_{s=1}^{S_l}$. When there is estimation uncertainty of C_l , denoted as $\hat{C}_l$, when Z_l is observed, then the probability that a new sample satisfies the condition is denoted as v_l , which is

$$v_l \triangleq \Pr(I(g_{i,l}^{k,j}(Z_l))>0 \mid g_{i,1}^{k,j}(Z_1),...,g_{i,l-1}^{k,j}(Z_{l-1})) = \sum_{\forall c_{ls}} \Pr(Z_l = c_{ls}, c_{ls} \in C_l \mid g_{i,1}^{k,j}(Z_1),...,g_{i,l-1}^{k,j}(Z_{l-1}))$$

$$= \sum_{\forall c_{ls}} \Pr(Z_l = c_{ls}) \Pr(c_{ls} \in C_l \mid Z_l = c_{ls}, g_{i,1}^{k,j}(Z_1),..., g_{i,l-1}^{k,j}(Z_{l-1})) \quad (5.12)$$

where $\Pr(Z_l = c_{ls})$ is the probability that the splitting value of a new sample is c_{ls} ; $\Pr(c_{ls} \in C_l \mid Z_l = c_{ls}, g_{i,1}^{k,j}(Z_1),...,g_{i,l-1}^{k,j}(Z_{l-1}))$ is conditional probability that c_{ls} belongs to the true splitting conditions, given that a new sample's splitting value is c_{ls} , and the previous splitting $g_{i,1}^{k,j}(Z_1),...,g_{i,l-1}^{k,j}(Z_{l-1})$.

Based on a PLRT with splitting uncertainty, the *i*-th baseline model is selected, when the splitting condition is satisfied, i.e., $I(g_i^{k,j}(Z_1,...,Z_L)) = \prod_{l=1}^{L} I(g_{i,l}^{k,j}(Z_l)) = 1$. Therefore, the probability to select the *i*-th baseline model is the product from Equation (5.10) to Equation (5.12), denoted as q_i , based on the types of variables and conditions:

$$q_i \triangleq \Pr(\prod_{l=1}^{L} I(g_{i,l}^{k,j}(Z_l)) = 1) = \Pr(g_{i,1}^{k,j}(Z_1) > 0,..., g_{i,L}^{k,j}(Z_L) > 0)$$

$$= \Pr(g_{i,1}^{k,j}(Z_1) > 0) \Pr(g_{i,2}^{k,j}(Z_1) > 0 \mid g_{i,2}^{k,j}(Z_1) > 0)...\Pr(g_{i,L}^{k,j}(Z_L) > 0 \mid g_{i,1}^{k,j}(Z_1) > 0,...,g_{i,L-1}^{k,j}(Z_{L-1}) > 0)$$

$$= \prod_{l \in Set\ \mathrm{A}} p_l^{+} \prod_{l \in Set\ \mathrm{B}} p_l^{-} \prod_{l \in Set\ \mathrm{C}} v_l \quad (5.13)$$

By integrating the parameter estimation uncertainty and splitting uncertainty based on Equation (5.8) and Equation (5.13), we have the run length distribution considering both types of uncertainties:

$$\Pr(t = t_0 \mid Z_{i,0}^{k,j}, Z_{i,1}^{k,j}, \mathbf{Z}_{i,A}^{k,j}, W) = \sum_i q_{t_0,i} q_i \tag{5.14}$$

where $q_{t_0,i}$ and q_i are denoted in Equation (5.8) and Equation (5.13), respectively. Here the run length refers to that in a control chart.

5.4.3 Estimation of Modeling Uncertainty

In order to evaluate the run length distribution of the control charts, there are several unknown random variable distributions to estimate, such as W, $\widetilde{Th}_l$ and c_{ls}. In this chapter, we use the cross validation to estimate the distributions of these random variables, then evaluate the run length performance based on these distributions. We discuss the random variables one by one.

First, we obtain the cross validation residual variances, and use the pooled variance as the variance of residuals, i.e., $\sigma^2_{e_{i,t}^{k,j}} = \frac{1}{M}\sum_{m=1}^{M} \hat{\sigma}^{2\ (-m)}_{e_{i,t}^{k,j}}$, where M is the number of cross validations; and $\hat{\sigma}^{2\ (-m)}_{e_{i,t}^{k,j}}$ is the estimate of residual variance in the *m*-th cross validation. Due to the recursive splitting procedure, this variance estimate is conditioning on the previous splits and models. Therefore, we can obtain an empirical distribution of W by calculating the following:

$$W^{(-m)} = \frac{\hat{\sigma}^{(-m)}_{e_{i,t}^{k,j}}}{\sigma_{e_{i,t}^{k,j}}} = \frac{\hat{\sigma}^{(-m)}_{e_{i,t}^{k,j}} \sqrt{M}}{\sqrt{\sum_{m=1}^{M} \hat{\sigma}^{2\ (-m)}_{e_{i,t}^{k,j}}}} \tag{5.15}$$

and the empirical distribution of W is denoted as $h_i^W(w)$.

Second, since $E(\hat{Th}_l \mid g_{i,1}^{k,j}(Z_1),...,g_{i,l-1}^{k,j}(Z_{l-1})) = Th_l$, we have $Th_l = \frac{1}{M}\sum_{m=1}^{M}\hat{Th}_l^{(-m)}$, where $\hat{Th}_l^{(-m)}$ is the estimate of splitting boundary in the m-th cross validation. We further calculate the difference between the $\hat{Th}_l$ and Th_l as:

$$\widetilde{Th}_l^{(-m)} = \hat{Th}_l^{(-m)} - Th_l \tag{5.16}$$

Based on Equation (5.16), we can estimate the probability p_l^+ and p_l^- from the empirical distribution of $\widetilde{Th}_l^{(-m)}$.

Third, for the splitting condition $Z_l \in C_l$, we need to estimate $\Pr(Z_l = c_{ls})$ and $\Pr(c_{ls} \in C_l \mid Z_l = c_{ls}, g_{i,1}^{k,j}(Z_1),...,g_{i,l-1}^{k,j}(Z_{l-1}))$ for Equation (5.12). $\Pr(Z_l = c_{ls})$ can be estimated from the empirical distribution of the training samples. $\Pr(c_{ls} \in C_l \mid Z_l = c_{ls}, g_{i,1}^{k,j}(Z_1),...,g_{i,l-1}^{k,j}(Z_{l-1}))$ is estimated using the empirical distribution from the M-fold cross validation as

$$\Pr(c_{ls} \in C_l \mid Z_l = c_{ls}, g_{i,1}^{k,j}(Z_1),..., g_{i,l-1}^{k,j}(Z_{l-1})) = \frac{\sum_{\mathrm{i}=1}^{M} \mathrm{m}_{c_{ls}\in C_l,\mathrm{i}}}{\sum_{\mathrm{i}=1}^{M} \mathrm{m}_{Z_l = c_{ls},\mathrm{i}}} \tag{5.17}$$

where $\mathrm{m}_{Z_l = c_{ls},\mathrm{i}}$ is the sample size when samples' splitting values equal to c_{ls} in the i-th iteration of cross validation; and $\mathrm{m}_{c_{ls}\in C_l,\mathrm{i}}$ is the sample size when c_{ls} is an element in C_l in the i-th iteration of cross validation, i.e., c_{ls} satisfies the splitting condition $Z_l \in C_l$.

By substituting the distributions of W, $\widetilde{Th}_l$ and c_{ls} into Equation (5.14), we obtain the run length distribution considering both estimation uncertainties and splitting uncertainties.

5.4.4 Optimization of the Control Chart System

By using multiple control charts to monitor a system, there may be excessive false alarm by using the Type I error α without Bonferroni correction. The Bonferroni charts assume the control charts have similar run length performance and divide the overall Type I error by the number of control charts. The control limits are shown as follows, based on the charting system proposed in Equation (5.4):

$$\begin{cases} UCL = z_{\alpha/(2D)} \\ CL = 0 \\ LCL = -z_{\alpha/(2D)} \end{cases} \quad if \;\; I(g_i^{k,j}(Z_1,...,Z_L)) = 1 \tag{5.18}$$

where $\alpha/(2D)$ is the Type I error for each control chart; and D is the number of quality responses in the PLRTs for risk adjustment.

In the Bonferroni correction of multiple control charts, the total Type I error of the overall charting system is equally allocated to each control chart. This allocation of the Type I error is conservative. Moreover, the uncertainty analysis of the PLRTs may indicate different type of model estimation uncertainty in risk adjustment. Thus, equal assignment of the overall Type I error may result in sub-optimal performance on the overall ARL.

In order to improve the ARL performance of the control chart system, we consider the modeling uncertainty when designing the control chart. In a multistage manufacturing process, one considers the optimization of the control chart systems by

minimizing the out-of-control ARL given that the in-control ARL is larger than a pre-determined threshold (Wu *et. al*, 2004). The Type-I errors for all the control charts are the decision variables in this optimization problem. In this framework, Shewhart charts are used to monitor the quality variables without risk-adjustment. We extend the formulation to the optimization of the control chart system for MMOPs with risk adjustment by solving the following optimization problem:

$$\begin{aligned} &\min_{\alpha_n, n=1,\ldots,T} \ \mathrm{ARL}_1^{\mathrm{all}} \\ &s.t. \ \mathrm{ARL}_0^{\mathrm{all}} \geq \gamma \end{aligned} \tag{5.19}$$

where $\mathrm{ARL}_0^{\mathrm{all}}$ and $\mathrm{ARL}_1^{\mathrm{all}}$ are the ARL of the overall control chart system, where the run length refers to that in a control chart system; α_n is the Type-I error for the control charts in the *n*-th control chart, when there are total of T control charts in a PLRT; and γ is the minimal ARL of the overall control chart system, which indicates the upper bound of the Type-I error of the control chart system.

For the in-control ARL of the control chart system $\mathrm{ARL}_0^{\mathrm{all}}$, i.e., $a = 0$ in Equation (5.7), the ARL for the *n*-th control chart is:

$$\mathrm{ARL}_0^n = E(t) = \sum_{t_0} t_0 \sum_i q_{t_0,i} q_i \tag{5.20}$$

To convert the above run length in a control chart to the run length in a control chart system, we integrate the probability to select the control chart:

$$\mathrm{ARL}_0^{n^*} = E(t^*) = \frac{E(t)}{q_n} \tag{5.21}$$

where t^* is the run length in the control chart system.

Because of the Shewhart type of independence for the individual control chart and control chart system, the probability P_0 that the control chart system has a false alarm is

$$P_0 = 1 - \prod_{n=1}^{T} (1 - 1/\mathrm{ARL}_0^{n^*}) \tag{5.22}$$

and

$$\mathrm{ARL}_0^{\mathrm{all}} = 1/P_0 \tag{5.23}$$

Similarly, for the out-of-control ARL of the control chart system $\mathrm{ARL}_1^{\mathrm{all}}$, i.e., $a \neq 0$ in Equation (5.7), the ARL for a group of control charts in the *l*-th PLRT is:

$$\mathrm{ARL}_1^n = E(t) = \sum_{t_0} t_0 \sum_i q_{t_0,i} q_i \tag{5.24}$$

After conversion of the run length from a control chart to a control chart system:

$$\mathrm{ARL}_1^{n^*} = E(t^*) = \frac{E(t)}{q_n} \tag{5.25}$$

where t^* is the run length in the control chart system.

Then the probability P_1 that the control chart system has a miss-detection is

$$P_1 = 1 - \prod_{n=1}^{T} (1/\mathrm{ARL}_1^{n^*}) \tag{5.26}$$

and

$$\mathrm{ARL}_1^{\mathrm{all}} = 1/(1 - P_1) = \prod_{n=1}^{T} \mathrm{ARL}_1^{n^*} \tag{5.27}$$

By substituting from Equation (5.23) to Equation (5.27) into Equation (5.19), we may solve for α_n using a similar way as in Wu *et. al* (2004). In this chapter, the overall optimization procedure is illustrated in Figure 5.5.

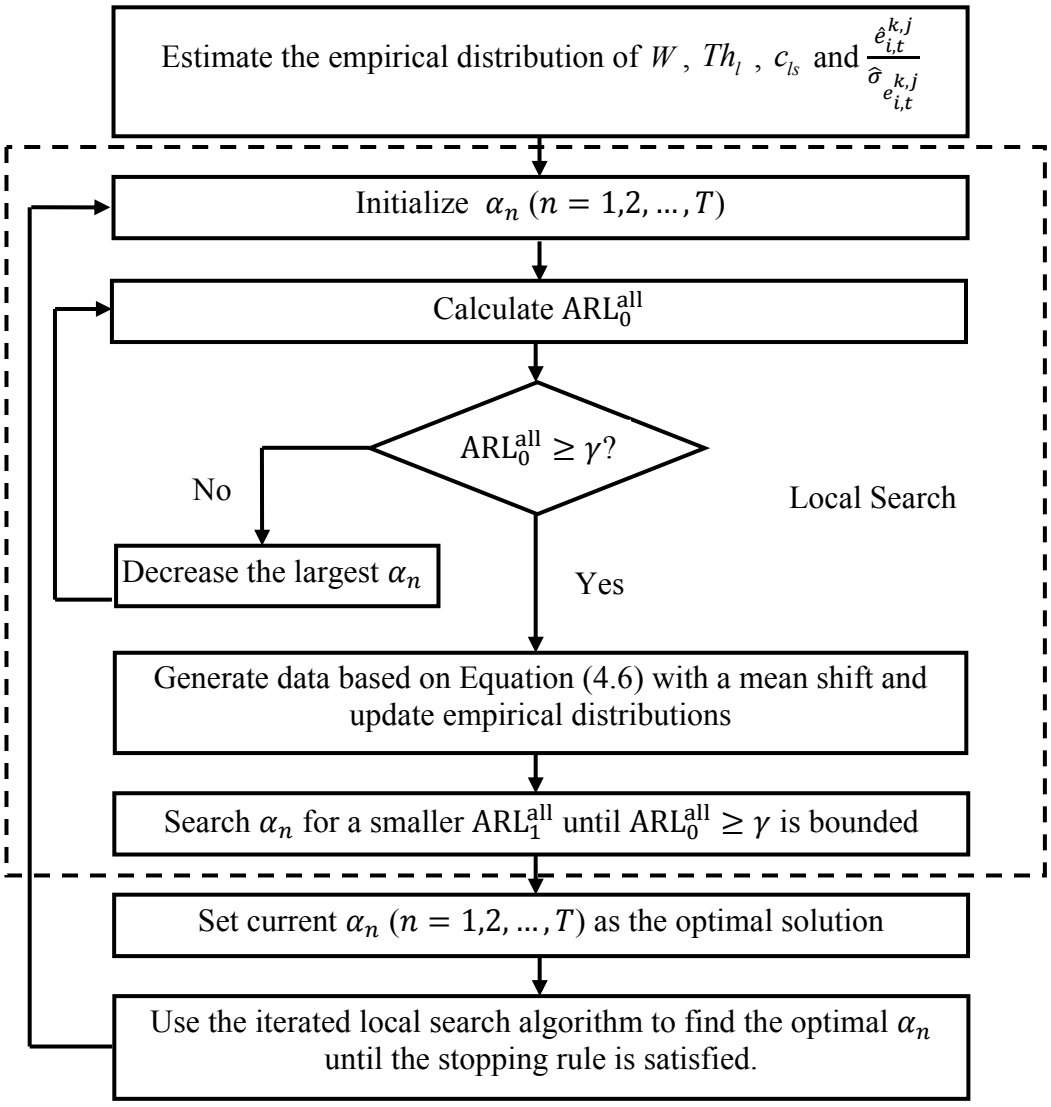


Figure 5.5: Optimization flow chart

It should be pointed out that the proposed method is easy to be implemented. A software package, named "PTOMonitor", has been developed following the procedure shown in Figure 5.2. Once a training set of samples is collected, the PTOMonitor will automatically construct the control chart system with optimized Type I errors.

5.5 Case Study

To evaluate the ARL performance of the proposed method, we use two data sets: one of the data sets is generated from simulation models, and the other data set is collected from the real production in wafer manufacturing processes. We compare the performance with other two types of benchmark control chart systems: (1) the conventional risk adjustment control chart system based on global regression models with the Bonferroni correction, denoted as "GB", and (2) the PLRT based risk adjustment control chart system with the Bonferroni correction, denoted as "PTB".

In the GB control chart system, the quality variables are predicted by global regression models of their upstream variables, without considering multimode. We monitor the residuals of regression models, where the Type I error is allocated to each Shewhart chart using the Bonferroni correction.

In the PTB control chart system, the quality variables are predicted by PLRTs, and the standardized residuals are similarly monitored as a PTO control chart system. Similar to the Bonferroni correction, the overall Type I error is equally allocated to each Shewhart chart, but further improved by solving the optimization in Equation (5.19) with an additional constraint such that the Type I errors for individual control charts are the same. This can be regarded as an improvement of the Bonferroni correction. The proposed PTO control chart system has the optimized Type I error by considering the modeling uncertainty. A comparison of the three types of control chart systems is provided in Sections 5.5.1 and 5.5.2.

5.5.1 Performance Comparison based on Simulation Models

The simulation models are developed based on a two-stage manufacturing process illustrated in Figure 5.6, where $\mathbf{Y}(k)$ ($k = 1,2$) are the quality variables at the k-th stage, where $\mathbf{Y}(1) = \begin{bmatrix}\mathbf{Y}(1)_1 & \mathbf{Y}(1)_2\end{bmatrix}^{\mathrm{T}} = \begin{bmatrix}y_{11} & y_{12}\end{bmatrix}^{\mathrm{T}}$, $\mathbf{Y}(2) = \begin{bmatrix}\mathbf{Y}(2)_1 & \mathbf{Y}(2)_2\end{bmatrix}^{\mathrm{T}} = \begin{bmatrix}y_{21} & y_{22}\end{bmatrix}^{\mathrm{T}}$; $\mathbf{X}_k$ ($k = 1,2.$) are the process variables, and $\mathbf{X}_1 = \begin{bmatrix}x_{11} & x_{12}\end{bmatrix}^{\mathrm{T}}$, $\mathbf{X}_2 = \begin{bmatrix}x_{21} & x_{22}\end{bmatrix}^{\mathrm{T}}$; and M is the material property variable.

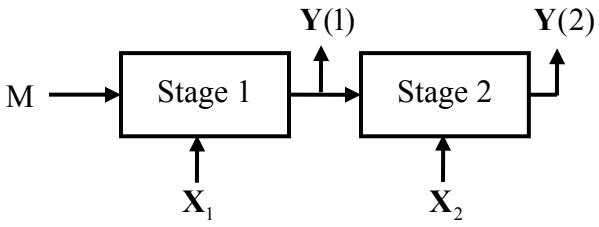
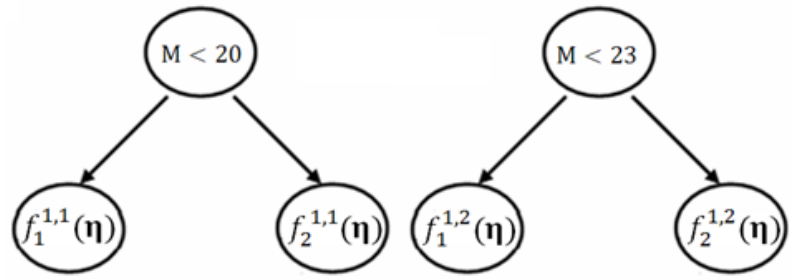
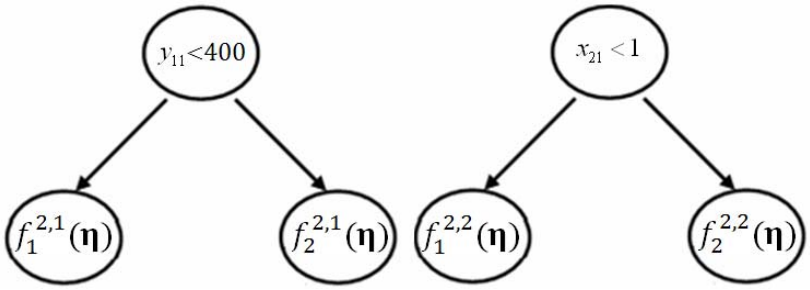


Figure 5.6: A two-stage manufacturing process for simulation models



(a) Model for y_{11} (b) Model for y_{12}



(c) Model for y_{21} (d) Model for y_{22}

Figure 5.7: Multimode structure for simulation models

The multimode structures for the quality variables are shown in Figure 5.7, where $f_i^{k,j}(\boldsymbol{\eta})$ is the *i*-th local model to predict the *j*-th quality variable at the *k*-th stage. Taking the model for y_{11} as an example, the simulation model $f_1^{1,1}(\boldsymbol{\eta})$ is $y_{11} = 10 + x_{11} - 2x_{12} + 20\mathrm{M} + e_1^{1,1}$, where $e_1^{1,1} \sim N(0,1)$, and the simulation model $f_2^{1,1}(\boldsymbol{\eta})$ is $y_{11} = 390 + x_{11} - 2x_{12} + \mathrm{M} + e_2^{1,1}$, where $e_2^{1,1} \sim N(0,1)$. The probability to select the first and the second local model is 0.3 and 0.7, respectively.

Based on these simulation models, we generate 300 training samples, and then use them to estimate global regression models and PLRT models. For example, the estimated global regression model for y_{11} is $\hat{y}_{11} = -512.88 + 0.92x_{11} - 1.15x_{12} + 81.33\mathrm{M} - 1.78\mathrm{M}^2 - 0.04x_{12}\mathrm{M}$, where the mean square error (MSE) of residuals is 13.23. When we consider the multimode to estimate the PLRT models, there is a splitting of material variable M at 20.01, where the estimated probability to select the first and the second local model is 0.30 and 0.70, respectively. The estimated $f_1^{1,1}(\boldsymbol{\eta})$ model is $\hat{y}_{11} = 5.12 + 0.95x_{11} - 2.00x_{12} + 20.26\mathrm{M}$, where the MSE of residuals is 0.99; and the estimated $f_2^{1,1}(\boldsymbol{\eta})$ is $\hat{y}_{11} = 389.30 + 1.00x_{11} - 2.00x_{12} + 1.03\mathrm{M}$, where the MSE of residuals is 1.02. To compare the modeling performance, we calculate ratio of the standard errors of the model residuals for the PLRT models and the global regression models in Table 5.1, where $\hat{\sigma}_{e^{k,j}}$ is the standard error of the global model to predict $\mathbf{Y}(k)_j$, and $\hat{\sigma}_{e_i^{k,j}}$ is the standard error of the *i*-th local model to predict $\mathbf{Y}(k)_j$ in the PLRTs. In Table 5.1, a smaller ratio of the standard errors indicates that the PLRTs are more accurate considering the multimode in the MMOP.

Based on the estimated models, we follow the proposed monitoring procedure and construct the control chart systems for GB, PTB, and PTO control chart systems. In order to optimize the Type I errors in the PTO control chart system, we assume that there are one local standard deviation of shifts in residuals in one of the eight local models with equal probability. The optimized Type I errors are summarized in Table 5.2, while the Type I errors for all control charts in GB and PTB control chart systems are 0.68×10^{-3} and 0.72×10^{-3}, respectively. We expect a smaller $\text{ARL}_1^{\text{all}}$ for the PTO control chart systems than that of the other two benchmark control chart systems, if there is a mean shift in the residuals of the local models with larger Type I error allocated, vice verse. Taking the control chart for the residuals of $f_1^{1,1}(\boldsymbol{\eta})$ as an example, the optimized Type I error is 1.46×10^{-3}, which is larger than the Type I errors in the GB and PTB control chart systems. Thus, the control limits becomes tighter and it is easier to detect the mean shift of the residuals of $f_1^{1,1}(\boldsymbol{\eta})$, with smaller $\text{ARL}_1^{\text{all}}$.

After the optimization of the Type I errors, we compare the ARL performance of these three control chart systems under different faulty scenarios in the Phase II study, which is shown in Table 5.3. The $\text{ARL}_0^{\text{all}}$ from the simulation for three control chart systems are 369.08, 368.00 and 367.47, respectively. The $\text{ARL}_1^{\text{all}}$ and the standard errors in the parenthesis are summarized. It is clear that the $\text{ARL}_1^{\text{all}}$ based on PLRT models has a better performance than that based on global regression models, which ignore the inherent multimode structures of the simulation models. Moreover, the PTO control chart system has a comparable $\text{ARL}_1^{\text{all}}$ as the PTB control chart system in some of the faulty scenarios, such as the mean shift in all eight models, but performs much better than

some other faulty scenarios, such as the mean shift in $f_1^{1,1}(\boldsymbol{\eta})$ or mean shifts in $f_2^{1,2}(\boldsymbol{\eta})$, $f_1^{2,2}(\boldsymbol{\eta})$ and $f_2^{2,2}(\boldsymbol{\eta})$.

Table 5.1: Ratios of the standard errors of simulation models

	$\frac{\hat{\sigma}_{e_1^{1,1}}}{\hat{\sigma}_{e^{1,1}}}$	$\frac{\hat{\sigma}_{e_2^{1,1}}}{\hat{\sigma}_{e^{1,1}}}$	$\frac{\hat{\sigma}_{e_1^{1,2}}}{\hat{\sigma}_{e^{1,2}}}$	$\frac{\hat{\sigma}_{e_2^{1,2}}}{\hat{\sigma}_{e^{1,2}}}$	$\frac{\hat{\sigma}_{e_1^{2,1}}}{\hat{\sigma}_{e^{2,1}}}$	$\frac{\hat{\sigma}_{e_2^{2,1}}}{\hat{\sigma}_{e^{2,1}}}$	$\frac{\hat{\sigma}_{e_1^{2,2}}}{\hat{\sigma}_{e^{2,2}}}$	$\frac{\hat{\sigma}_{e_2^{2,2}}}{\hat{\sigma}_{e^{2,2}}}$
Ratio	0.27	0.28	0.18	0.20	0.10	0.12	0.05	0.08

Table 5.2: Designed PTO control chart system for the simulation models

Control charts	Type I errors of PTO $\times 10^{-3}$	Control limits of standardized residuals
$f_1^{1,1}(\boldsymbol{\eta})$	1.46	-3.18 3.18
$f_2^{1,1}(\boldsymbol{\eta})$	0.55	-3.46 3.46
$f_1^{1,2}(\boldsymbol{\eta})$	0.61	-3.43 3.43
$f_2^{1,2}(\boldsymbol{\eta})$	1.20	-3.24 3.24
$f_1^{2,1}(\boldsymbol{\eta})$	0.41	-3.53 3.53
$f_2^{2,1}(\boldsymbol{\eta})$	0.61	-3.43 3.43
$f_1^{2,2}(\boldsymbol{\eta})$	1.19	-3.24 3.24
$f_2^{2,2}(\boldsymbol{\eta})$	0.77	-3.36 3.36

Table 5.3: ARL_1^{all} performance based on simulation models ($ARL_0^{all} \geq 370$)

Mean shift locations	Mean shift magnitude	ARL_1^{all}(GB) 369.08(4.63) (ARL_0^{all})	ARL_1^{all}(PTB) 368.00(4.19) (ARL_0^{all})	ARL_1^{all}(PTO) 367.47(4.13) (ARL_0^{all})
$f_1^{1,1}(\boldsymbol{\eta})$	σ	356.95(4.00)	189.42(1.68)	135.98(1.37)
$f_1^{1,1}(\boldsymbol{\eta})$	2σ	316.52(5.37)	34.35(0.54)	25.59(0.42)
$f_1^{1,1}(\boldsymbol{\eta})$	3σ	252.52(4.38)	9.35(0.16)	7.71(0.13)
$f_2^{2,2}(\boldsymbol{\eta})$	σ	368.78(6.04)	282.46(5.36)	282.89(5.65)
$f_2^{2,2}(\boldsymbol{\eta})$	2σ	364.33(5.21)	78.86(0.90)	75.85(1.05)
$f_2^{2,2}(\boldsymbol{\eta})$	3σ	359.86(5.02)	16.32(0.21)	15.86(0.19)
$f_1^{1,1}(\boldsymbol{\eta}), f_2^{1,2}(\boldsymbol{\eta}), f_1^{2,1}(\boldsymbol{\eta})$	σ	349.50(5.08)	91.34(1.27)	75.01(1.14)
$f_1^{1,1}(\boldsymbol{\eta}), f_2^{1,2}(\boldsymbol{\eta}), f_1^{2,1}(\boldsymbol{\eta})$	2σ	291.39(3.91)	12.02(0.18)	10.97(0.15)
$f_1^{1,1}(\boldsymbol{\eta}), f_2^{1,2}(\boldsymbol{\eta}), f_1^{2,1}(\boldsymbol{\eta})$	3σ	216.50(3.63)	3.34(0.03)	3.21(0.04)
$f_2^{1,2}(\boldsymbol{\eta}), f_1^{2,2}(\boldsymbol{\eta}), f_2^{2,2}(\boldsymbol{\eta})$	σ	359.80(5.70)	80.19(1.09)	59.75(0.86)
$f_2^{1,2}(\boldsymbol{\eta}), f_1^{2,2}(\boldsymbol{\eta}), f_2^{2,2}(\boldsymbol{\eta})$	2σ	340.41(4.50)	10.32(0.11)	8.42(0.09)
$f_2^{1,2}(\boldsymbol{\eta}), f_1^{2,2}(\boldsymbol{\eta}), f_2^{2,2}(\boldsymbol{\eta})$	3σ	298.19(4.09)	2.60(0.02)	2.33(0.01)
All eight models	σ	305.86(3.89)	28.18(0.32)	26.22(0.33)
All eight models	2σ	191.36(2.75)	3.42(0.05)	3.35(0.05)
All eight models	3σ	105.01(1.42)	1.23(0.01)	1.23(0.01)

5.5.2 Performance Comparison in Wafer Manufacturing Processes

We further demonstrate the effectiveness of the proposed PTO control chart system in a wafer manufacturing process. In this process, there are multiple operations to transform an ingot into wafers with thin film deposited. These operations include slicing, lapping, and CVD. The wafer manufacturing process is a very complex process, involving chemical and mechanical interactions of the wafers. The heterogeneity of the material property and different process conditions may introduce multimode under the normal conditions. Therefore, we will monitor the process by setting up the proposed PTO control chart system.

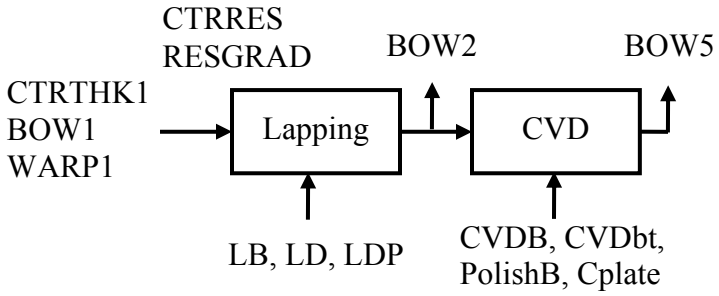
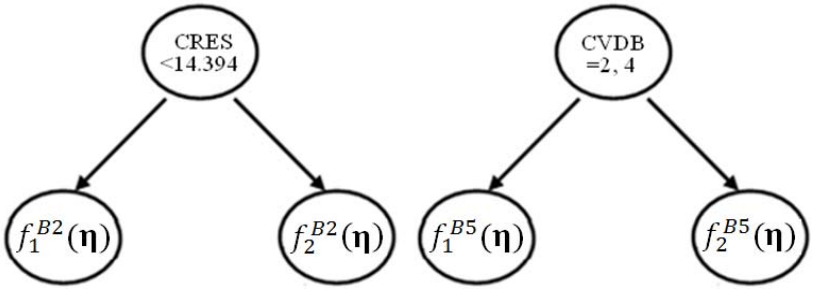


Figure 5.8: A two-stage wafer manufacturing process



(a) Model for BOW2 (b) Model for BOW5

Figure 5.9: Estimated multimode structures for the wafer manufacturing process

In this process, the objective is to monitor a geometric quality variable BOW2 and BOW5, which represents the bending orientation of the overall shape, due to substrate

geometric shape after the lapping process, and the residual stress of the thin films deposited in the CVD process. A two-stage process is shown in Figure 5.8. In this case study, quality, process, and material property variables were collected in a real production environment. The descriptions of those variables are summarized in Table 2.2 in Chapter 2.

Table 5.4: Ratios of the standard errors of models in wafer manufacturing process

	$\frac{\hat{\sigma}_{e_1^{B2}}}{\hat{\sigma}_{e^{B2}}}$	$\frac{\hat{\sigma}_{e_2^{B2}}}{\hat{\sigma}_{e^{B2}}}$	$\frac{\hat{\sigma}_{e_1^{B5}}}{\hat{\sigma}_{e^{B5}}}$	$\frac{\hat{\sigma}_{e_2^{B5}}}{\hat{\sigma}_{e^{B5}}}$
Ratio	0.81	0.79	0.80	0.51

Table 5.5: PTO control chart system for wafer manufacturing processes

Control charts	Type I errors of PTO $\times 10^{-3}$	Control limits of standardized residuals
$f_1^{B2}(\boldsymbol{\eta})$	0.72	-3.38 3.38
$f_2^{B2}(\boldsymbol{\eta})$	1.69	-3.14 3.14
$f_1^{B5}(\boldsymbol{\eta})$	0.49	-3.49 3.49
$f_2^{B5}(\boldsymbol{\eta})$	0.85	-3.34 3.34

Table 5.6: ARL_1^{all} performance for wafer manufacturing processes ($ARL_0^{all} \geq 370$)

Mean shift location	Mean shift magnitude	ARL_1(GB) 369.30(5.23) (ARL_0^{all})	ARL_1(PTB) 369.75(5.11) (ARL_0^{all})	ARL_1(PTO) 371.68(4.43) (ARL_0^{all})
$f_1^{B2}(\boldsymbol{\eta})$	σ	150.53(1.90)	123.24(1.79)	125.47(1.70)
$f_1^{B2}(\boldsymbol{\eta})$	2σ	29.54(0.47)	17.30(0.21)	17.79(0.23)
$f_1^{B2}(\boldsymbol{\eta})$	3σ	7.91(0.09)	4.78(0.05)	4.87(0.05)
$f_2^{B2}(\boldsymbol{\eta})$	σ	180.01(2.24)	116.93(1.89)	88.38(1.29)
$f_2^{B2}(\boldsymbol{\eta})$	2σ	40.44(0.56)	21.63(0.34)	16.00(0.23)
$f_2^{B2}(\boldsymbol{\eta})$	3σ	11.18(0.11)	6.20(0.06)	5.06(0.04)
$f_1^{B2}(\boldsymbol{\eta}), f_1^{B5}(\boldsymbol{\eta})$	σ	102.19(1.32)	71.03(1.04)	77.03(1.39)
$f_1^{B2}(\boldsymbol{\eta}), f_1^{B5}(\boldsymbol{\eta})$	2σ	17.45(0.21)	9.61(0.12)	10.62(0.14)
$f_1^{B2}(\boldsymbol{\eta}), f_1^{B5}(\boldsymbol{\eta})$	3σ	4.74(0.04)	2.91(0.04)	3.07(0.04)
$f_2^{B2}(\boldsymbol{\eta}), f_2^{B5}(\boldsymbol{\eta})$	σ	148.35(2.54)	57.30(0.64)	48.31(0.56)
$f_2^{B2}(\boldsymbol{\eta}), f_2^{B5}(\boldsymbol{\eta})$	2σ	31.72(0.46)	9.44(0.12)	8.10(0.11)
$f_2^{B2}(\boldsymbol{\eta}), f_2^{B5}(\boldsymbol{\eta})$	3σ	8.94(0.13)	2.94(0.03)	2.68(0.03)
All four	σ	72.18(1.10)	34.74(0.65)	32.30(0.53)
All four	2σ	11.69(0.13)	4.96(0.07)	4.79(0.06)
All four	3σ	3.25(0.05)	1.64(0.02)	1.60(0.02)

Overall, there are 373 samples collected during the manufacturing process. By iteratively removing the out-of-control samples in the Phase I study, as shown in Figure 5.2, 362 samples are used to construct the control chart systems. Similar to the performance study in Section 5.5.1, we first obtain the global regression models and the PLRT models, then optimize the control chart systems and compare the ARL performance. To estimate the PLRT models considering the multimode in wafer manufacturing process, all upstream quality, process and material property variables of BOW2 and BOW5 are treated as their predictors, respectively. By incorporating potential variables for multimode conditions in Loh's method (2002), the variables are selected and the estimated multimode structure is shown in Figure 5.9, where f_i^{B2} and f_i^{B5} (i=1, 2.) are the i-th local model to predict BOW2 and BOW5. The ratios of the standard errors of the model residuals for the PLRT models and the global regression models are summarized in Table 5.4, where $\hat{\sigma}_{e^{B2}}$ and $\hat{\sigma}_{e^{B5}}$ are the standard errors of the global models to predict BOW2 and BOW5, and $\hat{\sigma}_{e_i^{B2}}$ and $\hat{\sigma}_{e_i^{B5}}$ are the standard errors of the i-th local model to predict BOW2 and BOW5 in the PLRTs. The PLRTs have better modeling performance than the global regression models.

Based on the estimated models, we construct the PTO control chart systems following the procedure shown in Figure 5.2. The optimized Type I errors are summarized in Table 5.5, while the Type I errors for all control charts in GB and PTB control chart systems are 1.35×10^{-3} and 0.77×10^{-3}, respectively. The ARL performance of these three control chart systems under different faulty scenarios in the Phase II study is shown in Table 5.6. The $\mathrm{ARL}_0^{\mathrm{all}}$ from the simulation for three control chart systems are 369.30, 369.75 and 371.68, respectively. The $\mathrm{ARL}_1^{\mathrm{all}}$ and the standard

errors in the parenthesis are summarized. Similar to the result in Table 5.3, the ARL_1^{all} based on PLRT models has much better performance than that based on global regression models. Moreover, the PTO control chart system has better ARL_1^{all} than the PTB control chart system, such as the mean shifts in $f_2^{B2}(\boldsymbol{\eta})$. This is because the optimized Type I error of $f_2^{B2}(\boldsymbol{\eta})$ is larger in this MWMP. The PTO control chart system has capability to adjust the Type I error according to the modeling uncertainty.

5.6 Conclusions

A MMP may have multiple operational modes due to its complex nature and different variation propagation patterns. The process under different modes represents normal production conditions. However, existing process monitoring methods usually assume only one baseline model, or ignore the inter-relationship of variables when they are clustered in time space or variable space. Therefore, these methods may not be effective in a MMOP.

In this chapter, we proposed to use a PLRT to capture the variable relations in a MMOP, where we identify the number of operational modes by the number of leaf nodes, the mode conditions by the splitting conditions, and the baseline models by the local regression models. We set up a risk adjustment type of Shewhart control chart system to monitor the residuals of local models in order to detect process mean shifts in the residuals. Considering the modeling uncertainty, we study the run length distribution and optimize the control chart systems based on the modeling uncertainty. The proposed method has shown a better ARL performance than the risk adjustment based on global regression models or PLRT based risk adjustment with the Bonferroni correction, in both the simulation case and the real example in the wafer manufacturing processes.

CHAPTER 6

CONCLUSIONS AND FUTURE RESEARCH

6.1 Summary and Original Contributions

In wafer manufacturing processes, the availability of massive observational data provides opportunities to the advancement of quality control research, while it poses the challenges including the high dimensionality and heterogeneity of the data, and effectiveness in complex manufacturing process modeling. This thesis contributes to the quality control research by developing a unified variation modeling, analysis and control methodology for MWMPs, which includes the following aspects.

1. *An intermediate feedforward control strategy was developed for variation reduction by using intermediate quality responses to adjust control actions and analyzing the impact of measurement noise.* This method uses a group of regression models to capture the stage-to-stage variation in a MWMP. The intermediate feedforward control strategy adjusts and updates the control actions based on the online measurements of intermediate wafer quality. The proposed approach is evaluated on a MWMP that transforms an ingot into polished wafers. The proposed approach provides better control performance than the feedforward control based on a single regression model.
2. *A reconfigured PLRT for MMP control was developed to model the nonlinear data structures by reconfiguring the PLRTs to engineering complied models and reducing tree complexity.* This method proposes a methodology of feedforward

control based on piecewise linear models to model the nonlinear data structure. An engineering-driven reconfiguration method for piecewise linear regression trees is proposed. The model complexity is further reduced by merging the leaf nodes with the constraint of the control accuracy requirement. A case study indicates the proposed method has better control performance than that based on a group of regression models.

3. *A sequential measurement strategy was developed to measure the wafer geometric profile data more quickly and more efficiently by integrating the engineering driven sampling distribution.* This method proposes a sequential measurement strategy to reduce the number of samples measured in wafers, yet provide adequate accuracy for the quality feature estimation. In the proposed approach, initial samples are measured first, then a Gaussian process model is fitted to estimate the true profile of a wafer. The profile prediction and its uncertainty serve as guidelines to determine the measurement locations for the next sampling iteration. The measurement stops when the prediction error of the testing sample set satisfies the accuracy requirement. A case study indicates that the proposed methods take fewer samples than the random measurement strategy to model the wafer thickness profile data in slicing processes, while achieving comparable modeling accuracy.
4. *A monitoring method for a MMOP was developed to detect the mean shift of the residuals by using PLRT models, analyzing modeling uncertainty and optimizing the control chart performance.* This method uses a PLRT to capture the variable relations in a MMOP, where it identifies the number of operational modes by the

number of leaf nodes, the mode conditions by the splitting conditions, and the baseline models by the local regression models. A risk adjustment type of Shewhart control chart system is developed to monitor the residuals of local models in order to detect process mean shifts in the residuals. Considering the modeling uncertainty, the run length distribution is studied and the control chart system is optimized. The proposed method has shown a better ARL performance than the risk adjustment based on global regression models or PLRT based risk adjustment with the Bonferroni correction, in both the simulation case and the real example in the wafer manufacturing processes.

6.2 Future Research

There are several potential topics to be explored for further development of the variation modeling, analysis and control methodology. Here are several examples.

1. *Modeling considering data uncertainty and various model structures*. In Chapter 3 of the thesis, the PLRT models are constructed by assuming that the data uncertainty is negligible. The local models are also assumed to be linear regression models, and the manufacturing system is a static system. However, the data uncertainty is commonly encountered in manufacturing environments, which can be considered in the variation modeling efforts. Besides, the models may also need to have various forms with local generalized regression models or dynamic models to model different types of quality responses or system dynamics.
2. *Advancement of control and monitoring methodology*. In Chapter 3 of the thesis, the intermediate feedforward control strategy is developed based the PLRT models. However, the intermediate feedforward control strategy uses online

quality measurements, where measurement noise may not be negligible. The PLRT models also have modeling uncertainties. Therefore, it is necessary to study the impact of the measurement noise and modeling uncertainty to the control performance. In Chapter 5 of the thesis, the mean shifts of the residuals are monitored based on the PLRT models. The monitoring of the variance-covariance change and mode condition change can be further studied to better monitor the MMOP.

3. *Other applications of proposed methodology*. The increasing complexity and more abundant data in manufacturing process make the proposed methodology meet the challenges in this area. The variation modeling, analysis and control methodology can also be explored in other applications, such as engineered surface modeling and improvements, and quality control in nano-manufacturing.

APPENDIX

A.1 Proof of Statement 3.1

The temporal order of the splitting variables is assumed as $Z_{1*} \prec= Z_{2*} \prec= \ldots \prec= Z_{L*}$. In the decomposition of the sub-regions of $g_i(\cdot)$ into $g_i^j(\cdot)$

$$y = \sum_{i=1}^{T} f_i(\boldsymbol{\eta}_i) I(g_i(Z_1,...,Z_L)) = \sum_{i=1}^{T}\sum_{j=1}^{D_i} f_i(\boldsymbol{\eta}_i) I(g_i^j(Z_1,...,Z_L)) \quad \text{(A.1)}$$

Since the decomposed sub-regions involve all splitting variables, the temporally complied variables can be substituted into $g_i^j(\cdot)$.

$$y = \sum_{i=1}^{T}\sum_{j=1}^{D_i} f_i(\boldsymbol{\eta}_i) I(g_i^j(Z_{1*},...,Z_{L*})) \quad \text{(A.2)}$$

Since the splitting variables are temporally complied, the tree can be re-arranged into a temporally complied tree. Based on this tree, the merge of sub-regions follows the reverse temporal order. After the merge, sub-region $g_{i*}^j(\cdot)$ is the *j*-th region defined by a subset of $\{Z_{i*}\}$ for $f_i(\cdot)$, and there are T^* leaf nodes left:

$$y = \sum_{i^*=1}^{T^*} f_{i*}(\boldsymbol{\eta}_{i*}) I(g_{i*}(Z_{1*},...,Z_{L*})) = y^* \quad \text{(A.3)}$$

The original PLRT is statistically equivalent as the re-ordered model in prediction. □

A.2 Proof of Statement 3.2

Without loss of the generality, consider the case when there are two re-ordered models to be combined together, which are $y_1^* = \sum_{i=1}^{T_1^*} f_{i*}^1(\boldsymbol{\eta}_{i*}^1) I(g_{i*}^1(Z_{1*}^1,...,Z_{L_1^*}^1))$ and

$y_2^* = \sum_{i=1}^{T_2^*} f_{i*}^2(\boldsymbol{\eta}_{i*}^2) I(g_{i*}^2(Z_{1*}^2,...,Z_{L_2^*}^2))$. If $g_i^n(\cdot)$ is decomposed by all possible splits of the splitting variables in both models, then the first model is

$$y_1^* = \sum_{i=1}^{T_1^*} f_{i*}^1(\boldsymbol{\eta}_{i*}^1) I(g_{i*}^1(Z_{1*}^1,...,Z_{L_1^*}^1)) = \sum_{i=1}^{T_1^*}\sum_{j=1}^{D_{1,i}^*} f_{i*}^1(\boldsymbol{\eta}_{i*}^1) I(g_{i*}^{1,j}(Z_{1*}^1,...,Z_{L_1^*}^1, Z_{1*}^2,...,Z_{L_2^*}^2))$$

$$= \sum_{i=1}^{T_1^*}\sum_{j=1}^{D_{1,i}^*} f_{i*}^1(\boldsymbol{\eta}_{i*}^1) I(g_{i*}^{1,j}(Z_1,...,Z_{L^*})) \quad \text{(A.4)}$$

Similarly, the second model is:

$$y_2^* = \sum_{i=1}^{T_2^*}\sum_{j=1}^{D_{2,i}^*} f_{i*}^2(\boldsymbol{\eta}_{i*}^2) I(g_{i*}^{2,j}(Z_1,...,Z_{L^*})) \quad \text{(A.5)}$$

Since all possible splits of the splitting variables in both models are considered,

$$g_{i*}^{1,j}(Z_1,...,Z_{L^*}) = g_{i*}^{2,j}(Z_1,...,Z_{L^*}) \quad \text{(A.6)}$$

By following the procedure in Step 3 of Algorithm 2, these two models can be presented as:

$$y_1^* = \sum_{i*}^{T^*} f_{i*}^1(\boldsymbol{\eta}_{i*}^1) I(g_{i*}^{comb}(Z_1,...,Z_{L^*})) \quad \text{(A.7)}$$

and

$$y_2^* = \sum_{i*}^{T^*} f_{i*}^2(\boldsymbol{\eta}_{i*}^2) I(g_{i*}^{comb}(Z_1,...,Z_{L^*})) \quad \text{(A.8)}$$

where $g_{i*}^{comb}(Z_1,...,Z_{L^*})$ in both models are the same. Therefore, the combined model is the same as the original two re-ordered models in prediction. □

A.3 Derivation of Equation (5.7)

Based on Equation (5.2), we have

$$\frac{\hat{e}_{i,t}^{k,j}}{\hat{\sigma}_{e_{i,t}^{k,j}}} = \frac{\sigma_{e_{i,t}^{k,j}}}{\hat{\sigma}_{e_{i,t}^{k,j}}}[\frac{\hat{e}_{i,t}^{k,j} - E(\hat{e}_{i,t}^{k,j})}{\sigma_{e_{i,t}^{k,j}}} + \frac{E(\hat{e}_{i,t}^{k,j})}{\sigma_{e_{i,t}^{k,j}}}] \tag{A.9}$$

By substituting $E(\hat{e}_{i,t}^{k,j}) = \Delta_i^{k,j} + (\beta_{i,0}^{k,j} - \hat{\beta}_{i,0}^{k,j}) + (\boldsymbol{\beta}_{i,1}^{k,j} - \hat{\boldsymbol{\beta}}_{i,1}^{k,j})^{\mathrm{T}} \boldsymbol{\mu}_{\boldsymbol{\eta}_k}$ and $\hat{\sigma}^2_{e_{i,t}^{k,j}} = (\boldsymbol{\beta}_{i,1}^{k,j} - \hat{\boldsymbol{\beta}}_{i,1}^{k,j})^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\eta}_k} (\boldsymbol{\beta}_{i,1}^{k,j} - \hat{\boldsymbol{\beta}}_{i,1}^{k,j}) + \sigma^2_{e_{i,t}^{k,j}}$ from Equation (5.6) into Equation (A.9), we have

$$\frac{\hat{e}_{i,t}^{k,j}}{\hat{\sigma}_{e_{i,t}^{k,j}}} = \frac{\sigma_{e_{i,t}^{k,j}}}{\hat{\sigma}_{e_{i,t}^{k,j}}}[\frac{\hat{e}_{i,t}^{k,j} - E(\hat{e}_{i,t}^{k,j})}{\hat{\sigma}_{e_{i,t}^{k,j}}} \frac{\sqrt{(\boldsymbol{\beta}_{i,1}^{k,j} - \hat{\boldsymbol{\beta}}_{i,1}^{k,j})^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\eta}_k} (\boldsymbol{\beta}_{i,1}^{k,j} - \hat{\boldsymbol{\beta}}_{i,1}^{k,j}) + \sigma^2_{e_{i,t}^{k,j}}}}{\sigma_{e_{i,t}^{k,j}}} +$$

$$\frac{\Delta_i^{k,j} + (\beta_{i,0}^{k,j} - \hat{\beta}_{i,0}^{k,j}) + (\boldsymbol{\beta}_{i,1}^{k,j} - \hat{\boldsymbol{\beta}}_{i,1}^{k,j})^{\mathrm{T}} \boldsymbol{\mu}_{\boldsymbol{\eta}_k}}{\sigma_{e_{i,t}^{k,j}}}] \tag{A.10}$$

Define $Z_{i,0}^{k,j} = \frac{\hat{\beta}_{i,0}^{k,j} - \beta_{i,0}^{k,j}}{\sigma_{e_{i,t}^{k,j}} / \sqrt{\mathrm{m}}}$, $Z_{i,1}^{k,j} = \frac{(\hat{\boldsymbol{\beta}}_{i,1}^{k,j} - \boldsymbol{\beta}_{i,1}^{k,j})^{\mathrm{T}} \boldsymbol{\mu}_{\boldsymbol{\eta}_k}}{\sigma_{e_{i,t}^{k,j}}}$, $Z_{i,\mathbf{A}}^{k,j} = \frac{(\hat{\boldsymbol{\beta}}_{i,1}^{k,j} - \boldsymbol{\beta}_{i,1}^{k,j})^{\mathrm{T}} \mathbf{A}}{\sigma_{e_{i,t}^{k,j}}}$ and $Z_{\hat{e}_{i,t}^{k,j}} = \frac{\hat{e}_{i,t}^{k,j} - E(\hat{e}_{i,t}^{k,j})}{\hat{\sigma}_{e_{i,t}^{k,j}}}$, $Z_{i,0}^{k,j}$ and $Z_{\hat{e}_{i,t}^{k,j}}$ are standard normal random variables, $Z_{i,\mathbf{A}}^{k,j}$ is a multivariate normal random vector with mean as zeros and covariance as $\mathbf{A}^{\mathrm{T}}(\boldsymbol{\eta}_k \boldsymbol{\eta}_k^{\mathrm{T}})^{-1}\mathbf{A}$, and $Z_{i,1}^{k,j}$ is a normal random variable with zero mean and variance as $\boldsymbol{\mu}_{\boldsymbol{\eta}_k}^{\mathrm{T}}(\boldsymbol{\eta}_k \boldsymbol{\eta}_k^{\mathrm{T}})^{-1}\boldsymbol{\mu}_{\boldsymbol{\eta}_k}$ based on Equation (5.11) and Equation (5.12). Here, $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_k} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$, therefore,

$$\frac{\hat{e}_{i,t}^{k,j}}{\hat{\sigma}_{e_{i,t}^{k,j}}} = \frac{1}{W}[Z_{\hat{e}_{i,t}^{k,j}} \sqrt{\mathbf{Z}_{i,A}^{k,j} (\mathbf{Z}_{i,A}^{k,j})^{\mathrm{T}} + 1} + a - \frac{Z_{i,0}^{k,j}}{\sqrt{\mathrm{m_i}}} - Z_{i,1}^{k,j}] \qquad \square$$

REFERENCES

Abburi, N. R. and Dixit, U. S. (2006) A knowledge-based system for the prediction of surface roughness in turning process, *Robotics and Computer-Integrated Manufacturing*, **22**, 363-372.

Agrawal, R., Lawless, J. F. and Mackay, R. J. (1999) Analysis of variation transmission in manufacturing processes - part II, *Journal of Quality Technology*, **31**, 143-154.

Anderson, A. B., Wang, G. and Gertner, G. (2006) Local variability based sampling for mapping a soil erosion cover factor by cosimulation with Landsat^{TM} images, *International Journal of Remote Sensing*, **27**, 2423-2447.

Atkinson, P. M., Webster, R. and Curran, P. J. (1992) Cokriging with ground-based radiometry, *Remote Sensing of Environment*, **41**, 45-60.

Atkinson, P. M., Webster, R. and Curran, P. J. (1994) Cokriging with airborne MASS imagery, *Remote Sensing of Environment*, **50**, 335-345.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*, Wadsworth, Belmont, CA.

Chaudhuri, P., Huang, M. C., Loh, W. Y. and Yao, R. (1994) Piecewise-polynomial regression trees, *Statistical Sinica*, **4**, 143-167.

Chipman, H., George, E. and McCulloch, R. (1998) Bayesian CART model search (with discussion), *Journal of the American Statistical Association*, **93**, 935-960.

Chipman, H., George, E. and McCulloch, R. (2002) Bayesian treed models, *Machine Learning*, **48**, 299-320.

Cohn, D. A. (1996) Neural network exploration using optimal experimental design, *Neural networks*, **9**, 1071-1083.

Curran, P. J. and Williamson, H. D. (1986) Sample size for ground and remotely sensed data, *Remote Sensing of Environment*, **20**, 31-41.

Curran, P. J. (1988) The variogram in remote sensing: an introduction, *Remote Sensing of Environment*, **24**, 493-507.

Denison, D., Adams, N., Holmes, C. and Hand, D. (2002) Bayesian partition modeling, *Computational Statistics and Data Analysis*, **38**, 475-485.

Djurdjanovic, D. and Zhu, J. (2005) Stream of variation based error compensation strategy in multistation manufacturing processes, in *Proceedings of the 2005 ASME International Mechanical Engineering Congress and Exposition*, paper IMECE2005-81550, 314-319, Orlando, FL.

Doucet, A., Godsill, S. J. and Andrieu, C. (2000) On sequential simulation based methods for Bayesian filtering, *Statistics and Computing*, **10**, 197-208.

Doucet, A., de Freitas, J. F. G. and Gordon, N. (2001) *Sequential Monte Carlo in Practice*, Cambridge University Press, Cambridge, U.K..

Harnish, P., Nelson, B. and Runger, G. (2009) Process partitions from time-ordered clusters, *IIE Transactions*, **41**, 3-17.

Forsey, D. and Bartels, R. (1988) Hierarchical B-spline refinement, *Computer Graphics*, **22**, 205-212.

Gramacy, R. B. and Lee, H. K. H. (2009) Adaptive design and analysis of supercomputer experiment, *Technometric*, **51**, 130-145.

Guo, D., Wang, X. (2004) Dynamic sensor collaboration via sequential monte carlo, *IEEE journal on selected areas in communications*, **22**, 1037-1047.

Harnish, P., Nelson, B. and Runger, G. (2009) Process partitions from time-ordered clusters, *IIE Transaction*s, **41**, 3-17.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning*, Springer, New York.

Hauck, D. J., Runger, G. C. and Montgomery, D. C. (1999) Multivariate statistical process monitoring and diagnosis with grouped regression-adjusted variables, *Communications in Statistics-Simulation and Computation*, **28**, 309-328.

Hawkins, D. M. (1991) Multivariate quality control based on regression-adjusted variables, *Technometrics*, **33**, 61-75.

Hawkins, D. M. (1993) Regression adjustment for variables in multivariate quality control, *Journal of Quality Technology*, **25**, 170-182.

Huang, D., Allen, T. T., Notz, W. I. and Miller, R. A. (2006) Sequential kriging optimization using multiple-fidelity evaluations, *Structural and Multidisciplinary Optimization*, **32**, 369-382.

Huang X. and Gao Y. (2010) A discrete system model for form error control in surface grinding, *International Journal of Machine Tools and Manufacture*, **50**, 219-230.

Hwang, D. H. and Han, C. (1999) Real-time monitoring for a process with multiple operating modes, *Control Engineering Practice*, **7**, 891-902.

Intellektik, F., Informatik, F. and Stutzle, T. (1999) Iterated local search for the quadratic assignment problem, Technical report, aida-99-03, FG Intellektik.

Izquierdo, L. E., Shi, J., Hu, S. J. and Wampler, C. W. (2007) Feedforward control of multistage assembly processes using programmable tooling, *Transactions of NAMRI/SME*, **35**, 295-302.

Jin, J. and Ding, Y. (2004) Online automatic process control using observable noise factors for discrete part manufacturing, *IIE Transactions*, **36**, 899-911.

Jin, J. and Shi, J. (1999) State space modeling of sheet metal assembly for dimensional control, *ASME Transactions*, *Journal of Manufacturing Science and Engineering*, **121**, 756-762.

Jin, R. and Shi, J., Intermediate feedforward control strategy in wafer manufacturing process, manuscript.

Jin, R., Chang, C. J. and Shi, J. (2011) Sequential sensing strategy of wafer profiles using Gaussian process model, *IIE Transactions*, in-press.

Jin, R. and Shi, J. (2011) Reconfigured piecewise linear regression tree for multistage manufacturing process control, *IIE Transactions*, in-press.

Jin, R., Liu, K. and Shi, J., Multistage multimode process monitoring based on a piecewise linear regression tree considering modeling uncertainty, manuscript.

Jobe, J. M. and Pokojovy, M. (2009) A multistep, cluster-based multivariate chart for retrospective monitoring of individuals, *Journal of Quality Technology*, **41**, 323-339.

Jiao, Y. and Djurdjanovic, D. (2010) Joint allocation of measurement points and controllable tooling machines in multistage manufacturing processes, *IIE Transactions*, **42**, 703-720.

Joseph, R. (2003) Robust parameter design with feed-forward control, *Technometrics*, **45**, 284-292.

Jobe, J. M. and Pokojovy, M. (2009) A multistep, cluster-based multivariate chart for retrospective monitoring of individuals, *Journal of Quality Technology*, **41**, 323-339.

Kim, H., Loh, W. Y., Shih, Y. and Chaudhuri, P. (2007) Visualizable and interpretable regression models with good prediction power, *IIE Transactions*, **39**, 565-579.

Kleijnen, J. P. C. and Beers W. C. M. van (2004) Application-driven sequential designs for simulation experiments: kriging metamodelling, *Journal of the Operational Research Society*, **0**, 1-8.

Larsen, D. R., and Speckman, P. L. (2004) Multivariate regression trees for analysis of abundance data, *Biometrics*, **60**, 543-549.

Lawless, J. F., Mackay, R. J. and Robinson, J. A. (1999) Analysis of variation transmission in manufacturing processes - part I. *Journal of Quality Technology*, **31**, 131-142.

Lee, S. K. (2006) On classification and regression trees for multiple responses and its application, *Journal of Classification*, **23**, 123-141.

Lee, S., Wolberg, G. and Shin, S. Y. (1997) Scattered data interpolation with multilevel B-splines, *IEEE Transaction on Visualization and Computer Graphics*, **3**, 228-244.

Liu, J. S., Chen, R. (1998) Sequential Monte Carlo methods for dynamic systems, *Journal of the American Statistical Association*, **93**, 1032-1044.

Loh, W. Y. (2002) Regression trees with unbiased variable selection and interaction detection, *Statistical Sinica*, **12**, 361-386.

Loh, W. Y., Chen, C. and Zheng, W. (2007) Extrapolation errors in linear model trees, *ACM Transactions on Knowledge Discovery in Data*, **1**, 1-17.

Loh, W. Y. (2007) Regression by parts: fitting visually interpretable models with GUIDE in *Handbook of Data Visualization*, Chen, C., Hardle, W. and Unwin, A., Springer-Verlag, Berlin, Germany, pp. 447-468.

Lophaven, S. N., Nielsen, H. B. and Søndergaard, J. (2002), *manual of DACE*, a matlab kriging tool box.

Lourenco, H.R., Martin, O.C. and Stutzle, T. (2002) Iterated local search, *Handbook of Metaheuristics*, *International Series in Operations Research & Management Science*, **57**, 321-353.

Lowry, C. A. and Montgomery, D. C. (1995) A review of multivariate control charts, *IIE Transactions*, **27**, 800-810.

Mackay, D. J. C. (1992) Information-based objective functions for active data selection, *Neural Computation*, **4**, 589-603.

Mantripragada, R. and Whitney, D. E. (1999) Modeling and controlling variation propagation in mechanical assemblies using State Transition Models, *IEEE Transactions on Robotics and Automation*, **115**, 124-140.

McBratney, A. B. and Webster, R. (1983a) How many observations are needed for regional estimation of soil properties? *Journal of Soil Science*, **135**, 177-183.

McBratney, A. B. and Webster, R. (1983b) Optimal interpolation and isarithmic mapping of soil properties V. Co-regionalization and multiple sampling strategy, *Journal of Soil Science*, **34**, 137-162.

Montgomery, D.C. (2001) *Introduction to Statistical Quality Control, 4th ed*., Wiley, New York, NY.

Morgan, J. N. and Sonquist, J. A. (1963) Problems in the analysis of survey data, and a proposal, *Journal of American Statistical Association*, **58**, 415-434.

Muller, P., Sanso, B. and de Iorio, M. (2004) Optimal Bayesian design by inhomogeneous Markov Chain simulation, *Journal of the American Statistical Association*, **99**, 788-798.

Ozcelik, B. and Bayramoglu, M. (2006) The statistical modeling of surface roughness in high-speed flat end milling, *International Journal of Machine Tools and Manufacture*, **46**, 1395-1402.

Park, S., Fowler, J. W., Mackulak G. T., Keats, J. B. Carlyle, W. M. M. (2002) D-optimal sequential experiments for generating a simulation-based cycle time-throughput curve, *Operations Research*, **50**, 981-990.

Pei, Z. J., Xin, X. J. and Liu W. (2003) Finite element analysis for grinding of wire-sawn silicon wafers: a designed experiment, *International Journal of Machine Tools & Manufacture*, **43**, 7-16.

Pei, Z. J., Kassir, S., Bhagavat M. and Fisher, G. R. (2004) An experimental investigation into soft-pad grinding of wire-sawn silicon wafers, *International Journal of Machine Tools & Manufacture*, **44**, 299-306.

Pierre, D. A. (1986) *Optimization Theory with Applications*, Dover, New York, NY.

Rao, S., Strojwas, A.J., Lehoczky, J.P. and Schervish, M.J. (1996) Monitoring multistage integrated circuit fabrication processes, *IEEE Transactions on Semiconductor Manufacturing*, **9**, 495-505.

Santner, T. J., Williams, B. J. and Notz, W. I. (2003) *The design and analysis of computer experiments, Springer*, New York.

Schroder, P. (1996) Wavelets in computer graphics, in *Proceedings of the IEEE*, **84**, 615-625.

Schonlau, M., Welch, W. J. and Jones, D.R. (1998) Global versus local search in constrained optimization of computer models, *New Developments and Applications in Experimental Design*, **34**, 11-25.

Sederberg, T. W., Cardon, D. L., Finnigan, G. T., North, N. S., Zheng, J. and Lyche, T. (2004) T-spline simplification and local refinement, *ACM Transactions on Graphics*, **3**, 276-283.

Segal, M. R. (1992) Tree-structured methods for longitudinal data, *Journal of the American Statistical Association*, **87**, 407-418.

Seo, S., Wallat, M., Graepel, T. and Obermayer, K. (2000) Gaussian process regression: active data selection and test point rejection, in *Proceedings of the International Joint Conference on Neural Networks*, 241-246.

Shi, J. (2006) *Stream of Variation Modeling and Analysis for Multistage Manufacturing Processes*, CRC Press, New York, NY.

Shi, J. and Apley, D. (1998) A suboptimal N-Step-Ahead cautious controller for adaptive control applications, *ASME Transactions*, *Journal of Dynamic Systems, Measurement and Control*, **120**, 419-423.

Shi, J., Wu, C. F., Yang, X. and Zheng, H. (2005) Design of DOE-based automatic process controller for complex manufacturing processes, in *Proceedings of NSF DMII Grantees Conference*, Scottsdale, Arizona.

Shi, J. and Zhou, S. (2009) Quality control and improvement for multistage systems: a survey, *IIE Transactions*, **41**, 744-753.

Shu, L. J., Apley, D. W. and Tsung, F. (2003) Autocorrelated process monitoring using triggered cuscore charts, *Quality and Reliability Engineering International*, **18**, 411-421.

Shu, L. J. and Tsung, F. (2003) On multistage statistical process control, *Journal of the Chinese Institute of Industrial Engineers*, **20**, 1-8.

Shu, L. J., Tsung, F. and Tsui, K. L. (2004a) Run-length performance of regression control charts with estimated parameters, *Journal of Quality Technology*, **36**, 280-292.

Shu, L. J., Tsung, F. and Kapur, K. C. (2004b) Design of multiple cause selecting charts for multistage processes with model uncertainty, *Quality Engineering*, **16**, 437 - 450.

Stoumbos, Z. G., Reynolds, M. R., Ryan, T. P. and Woodall, W. H. (2000) The state of statistical process control as we proceed into the 21st century, *Journal of the American Statistical Association*, **95**, 992-998.

Stutzle, T. (1998) *Local Search Algorithms for Combinatorial Problems-Analysis, Improvements and New Applications*, Ph.D thesis, FB Informatik, TU Darmstadt.

Sukchotrat, T., Kim S. B. and Tsung, F. (2010) One-class classification-based control charts for multivariate process monitoring, *IIE Transactions*, **42**, 107-120.

Taguchi, G., Wlsayed, E. and Hsiang, T. (1989) *Quality Engineering in Production Systems*, McGraw-Hill, New York, NY.

Valette, S. and Prost, R. (2004) Wavelet-based multiresolution analysis of irregular surface meshes, *IEEE Transaction on Visualization and Computer Graphics*, **10**, 113-122.

Wade, M. R. and Woodall, W. H. (1993) A review and analysis of cause selecting control charts, *Journal of Quality and Technology*, **25**, 161-169.

Wang, G., Gertner, Z. G. and Anderson, A. B. (2005) Sampling design and uncertainty based on spatial variability of spectral reflectance for mapping vegetation cover, *International Journal of Remote Sensing*, **26**, 3255-3274.

Williams, B. J., Santner T. J. and Notz, W. I. (2000) Sequential design of computer experiments to minimize integrated response functions, *Statistica Sinica*, **10**, 1133-1152.

Woodall, W. H. and Montgomery, D. C. (1999) Research issues and ideas in statistical process control, *Journal of Quality Technology*, **31**, 376-386.

Wu, Z., Lam, Y.C., Zhang, S. and Shamsuzzaman, M. (2004) Optimization design of control chart systems, *IIE Transactions*, **36**, 447-455.

Xiang, L. and Tsung, F. (2008) Statistical monitoring of multi-stage processes based on engineering models, *IIE Transactions*, **40**, 957-970.

Xiao, X., Gertner, G. Z., Wang, G. and Anderson, A. B. (2005) Optimal sampling scheme for estimation and landscape mapping of vegetation cover, *Landscape Ecology*, **20**, 375-387.

Zantek, P. F., Wright, G. P. and Plante, R. D. (2002) Process and product improvement in manufacturing systems with correlated stages, *Management Science*, **48**, 591-606.

Zhang, G. M. and Kapoor, S. G. (1990) Dynamic generation of machined surfaces part 1: description of a random excitation system, Technical Report.

Zhang, G. X. (1984) A new type of control charts and a theory of diagnosis with control charts, *World Quality Congress Transactions*, *American Society for Quality Control*, London, 175-185.

Zhang, G. X. (1985) Cause-selecting control charts - a new type of quality control charts, *Operation Research*, **12**, 221-225.

Zhang, G. X. (1989) A new diagnosis theory with two kinds of quality, *43rd ASQC Annual Quality Congress Transactions*, Toronto, **43**, 594-599.

Zhang, G. X. (1992) Cause-selecting control chart and diagnosis, theory and practice, Aarhus School of Business, Department of Total Quality Management, Aarhus, Denmark.

Zhang, H. and Albin, S. L. (2007) Detecting the number of operational modes in baseline multivariate SPC data, *IIE Transactions*, **39**, 1103-1110.

Zhao, H., Jin, R. and Shi, J. (2011) PDE-constrained Gaussian process model on material removal rate of wiresaw slicing process modeling, *ASME transactions*, *Journal of Manufacturing Science and Engineering*, in-press.

Zhong, J., Shi, J. and Wu, C. F. (2010) Design of DOE-based automatic process controller with consideration of model and observation. *IEEE Transactions on Automation Science and Engineering*, **7**, 266-273.

Zhu, L. and Kao, I. (2005) Galerkin-based modal analysis on the vibration of wire–slurry system in wafer slicing using a wire saw, *Journal of Sound and Vibration*, **283**, 589-620.