2018

LEAD-TIME QUOTATION BY SYNERGISTICALLY MODELING REAL AND SIMULATION DATA

Hoda Sabeti

LEAD-TIME QUOTATION BY SYNERGISTICALLY MODELING REAL AND SIMULATION DATA

Hoda Sabeti

Dissertation submitted to the
Benjamin M. Statler College of Engineering and Mineral Resources
at West Virginia University

in partial fullfillment of the requirements for the degree of

Doctor of Philosophy
in
Industrial Engineering

Feng Yang, Chair Ph.D.
Majid Jaridi, Ph.D.
Kenneth Currie, Ph.D.
Robert Mnatsakanov, Ph.D.
Xi Chen, Ph.D.

Department of Industrial and Management Systems Engineering

Morgantown, West Virginia
2018

Keywords: Lead Time Quotation, Metamodeling, Stochastic Kriging, Gaussian Process Regression, Information Infusing

ABSTRACT

Lead-Time Quotation by Synergistically Modeling Real and Simulation Data

Hoda Sabeti

The ability to quote a competitive and reliable lead time for a new order is a key competitive advantage for manufacturers and plays a significant role in customer acquisition and satisfaction. Upon the arrival of a customer's order, it is critical to accurately predict the flow time (the time needed to complete that job) and quote its lead time accordingly. Quoting a precise and reliable lead time requires a good prediction for the flow time of a new order. A new job's flow time through the system depends on the complex shop-floor status upon its arrival and is also subject to uncertainties in manufacturing processes such as stochastic processing times and random machine failures. Hence, it is challenging to provide high-quality flow time estimation for a new order at its arrival time.

This research focuses on quantifying the dependence of the flow time upon observed job shop status variables, the size of a new order, and the arrival rate of future orders. An iterative fitting procedure based on stochastic kriging with qualitative factors, is developed to synergistically model simulation and real manufacturing data, for the prediction of a new order's flow time. The fitting procedure aims at exploiting the strengths of both simulation data, which can be well designed, and real data, which are observed from manufacturing, to achieve a high-quality prediction model of flow time.

Dedicated to my inspiring mom and dad

Thank you for always believing in me.

Acknowledgments

I would like to thank my advisor, Dr. Yang, for giving me the opportunity to work with her, and for her support throughout my study. It is my great honor to work under her guidance. I am also thankful to Dr. Jaridi, Dr. Currie, Dr. Mnatsakanov, and Dr. Chen for being my committee members and for their assistance in preparing this dissertation.

Contents

List of Figures

List of Tables

Acronyms

ANN	Artificial neural network
BPN	Back-propagation network
CT	Cycle time
DES	Discrete Event Simulation
DOE	Design of Experiment
EDD	Earliest due date
ERMSE	Estimated root mean squared error
FCFS	First-come- first-served
GA	Genetic algorithm
GPR	Gaussian Process Regression
HL	High level
IB	Important buffers
IW	Important Work Station
LL	Lower level
ML	Medium level
MRP	Material requirements planning
MTO	Make to order
MTS	Make to stock
PM	Product mix
SK	Stochastic Kriging
SKQ	Stochastic Kriging with Qualitative Factors
DKQ	Deterministic Kriging with Qualitative Factors
SV	Status Variables
TH	Throughput
TTF	Time to failure
TTR	Time to repair
TWK	Total work content
WIP	Work In Process

Nomenclature

Symbol	Units	Description
$\alpha(\mathbf{w})$		Shape parameter of gamma distribution
β		Vector of related unknown parameters for functions
$\beta(\mathbf{w})$		Scale parameter
β_0		Constant term representing the overall surface mean
θ		Vector of parameters for quantitative factors x
$\mathbf{v}(\mathbf{w}_0, \theta, \Phi)$		Correlation vector
δ^2		Variance of the Gaussian process
$\overline{\varepsilon}$		Vector of sample average errors
$\mathbf{R}(\theta, \Phi)$		Correlation matrix
$\mathbf{r}(\mathbf{w}_i)$		Logarithm of the samples variance including the bias correction
Σ_ε		Intrinsic variance-covariance matrix
Σ_M		Variance-covariance matrix
$\mathsf{M}(\mathbf{w})$		Mean-zero stationary Gaussian process
τ		Set of unknown parameters for qualitative
ε_{obs}		The vector of observed errors
$\varepsilon_j(\mathbf{w})$		Intrinsic variability
$\varepsilon_j(\mathbf{w})$		Random error variability
$\widehat{e}^2(\mathbf{w}_i)$		Estimated squared residual at point $(\mathbf{w}_i)$
$f(\mathbf{w})$		Vector of known functions of w
I		Total number of design points
K		Number of simulated data points where replications are available
l_j		The quoted lead time
$\mathbf{r}(\mathbf{w}_i)$		The logarithm of estimated variances in re-sampling method
$M(\mathbf{w})$		Extrinsic variability
$n(\mathbf{w}_i)$		Number of replications at $(\mathbf{w}_i)$
$\mathbf{p}(\mathbf{w}_i)$		The logarithm of estimated variances in differenced based method

Nomenclature

Symbol	Units	Description
Q		Levels of the total WIP
QL		Lower bound of WIP
QU		Upper bound of WIP
R		Number of real data points where replications are not available
S_p		Number of resampling
SV		Job shop status variables
$SVs.A$		The number of jobs at each buffer
$SVs.B$		The status (busy or idle) of each server.
$SVs.C$		The elapsed processing time at each busy server
$SVs.D$		The status (up or down) of each server that is subject to random failures
$SVs.E$		The elapsed down time for a currently down server.
$SVs.F$		The elapsed up time for a currently up server.
$SVs.G$		The batch size currently being processed at a batch processing server
S^2		The vector of all estimated variances in resampling method
T^2		The vector of all estimated variances in differenced based method
X		Vector of quantitative factors
X_{ORG}		Vector of all potential factors of this problem
x_{WIP}		The number of jobs in the sth stage
$\mathbf{x}_C$		The elapsed processing times at important busy servers
$\mathbf{x}_E$		The elapsed down times for important down servers
$\mathbf{x}_F$		The elapsed up time for each important up server
$\mathbf{x}_G$		The batch sizes being handled by important busy servers involving batch processing
$\mathbf{x}_O$		The size of a newly arrived order
$\mathbf{x}_R$		The forecasted arrival rate of future orders
$\mathbf{z}_B$		The busy or idle status of important servers

Nomenclature

Symbol	Units	Description
$\mathbf{z}_D$		The status (up or down) of important servers subject to random failures
$\mathbf{z}_S$		The source of the data
$\mathbf{w}_i$		ith design point (variable setting of the manufacturing plant)
$\mathscr{Y}_j^*$		Flow time in the VD data set
$\mathscr{Y}_{obs}$		The vector of observed responses at $\mathbf{w}_i$
$\mathscr{Y}_{obs}(\mathbf{w}_i)$		Observed responses at $\mathbf{w}_i$
$\mathscr{Y}_j(\mathbf{w}_i)$		Response j at design point $\mathbf{w}_i$
$\overline{\mathscr{Y}}$		The vector of sample average at replicated points
$\overline{\mathscr{Y}}(\mathbf{w}_i)$		In K replicated data points, the sample average of the responses
Z		Vector of qualitative factors

Chapter 1

Introduction

1.1 Background

The ability to quote a competitive and reliable lead time for a new order is a key competitive advantage for manufacturers and plays a significant role in customer acquisition and satisfaction. In this study, lead time is defined as the difference between the promised due date of an order (or job) and its arrival time [2]. Quoting a precise and reliable lead time requires a good prediction for the flow time of a new order, the time it takes for a job to traverse through a manufacturing process [3].

A new job's flow time through the system depends on the complex shop-floor status upon its arrival and is also subject to uncertainties in manufacturing processes. Various sources of uncertainty affect the make to order (MTO) manufacturing environments and distinguish it from make to stock (MTS) firms [4]. Upon the arrival of a new order, some previous orders are still being processed in the manufacturing system, and some unfinished jobs are still queueing in front of stations at some stage of their processing sequence. Besides, some job shop features (e.g. random machine failure, random processing times, etc.) affect the flow time of a new order. Hence, the flow time of a new job depends on inherent uncertainties of the manufacturing system and the current status of the shop floor and cannot be predicted with exactness and it is challenging to provide high-quality flow time estimation for a new order at its arrival time.

In the literature, two types of approaches have been used for flow time estimation: analytical and numerical approaches. On the analytical side, a range of queueing models have been developed ([5, 6, 7, 8, 9]). Analytical models rely on restrictive assumptions such as the Markovian property,

and fall short in capturing the realistic features of manufacturing processes.

The majority of numerical approaches employ either real or simulation data to develop a surrogate model approximating the functional relationship between the expected flow time and the various shop-status factors ([10, 11, 12, 13]). These surrogate models include classic linear regression ([14, 15, 16, 17, 18]) as well as powerful models such as neural network ([16, 19, 20]).

In this stream of numerical work, [1] is the first paper that takes an experimental design effort based on discrete-event simulation of manufacturing: Simulation experiments are designed to provide a good coverage of the input space spanned by the typically large number of quantitative and qualitative factors depicting shop floor status. Good design of experiments is critical to the quality of the fitted prediction model for flow time, especially when the input space is large and complex. However, for the planning and control of manufacturing, experimental design can only be performed on simulation models, which is high-fidelity but nevertheless deviates somewhat from the real-world system. With increased capability to track and monitor manufacturing processes, more and more real data will be available for decision making. In contrast to simulation data, real data unquestionably reflects the actual behavior of the manufacturing system being investigated, but are not subject to experimental design.

To take advantage of both simulation and real data, this work adapts the stochastic kriging with qualitative factors (SKQ)[21] and develops an SKQ-based iterative procedure to synergistically model simulation and real data, aiming at exploiting the strengths of both types of data to achieve a prediction model of the high quality. Stochastic kriging with qualitative factors (SKQ) is highly flexible and able to provide an accurate approximation of practically any continuous response surfaces [21, 22, 23, 24] without requiring a presumed functional form as traditional nonlinear regression does [21, 25].

1.2 Research Objective

In this dissertation, we studied the problem of quoting lead time for randomly arriving customer orders to a manufacturing system and developed a method to synergistically model both simulation based data and real-time manufacturing actual data to obtain a high-quality prediction of flow time and lead time quotation. This research focuses on quantifying the dependence of the flow time

upon observed job shop status variables, the size of a new order, and arrival rate of future orders. The problem includes a variety of manufacturing systems with different features and conditions like machine failure, batch processing, multiple workstations, re-entrant flows, and multiple type job flows, etc.

A SKQ-based iterative procedure is developed for estimating the flow time and quoting lead time. To provide high-quality lead time quotation, we characterize the flow time of a job by modeling not only the first but also the second moment characteristics of flow time as a function of shop status variables. Based on flow time's first two moments, we estimate its percentiles, which enables real-time due date quotation with a desired service level.

The proposed method has been applied to a scaled-down manufacturing system. The quality of the models has been evaluated regarding commonly used performance criteria, based on well-designed validation data set.

1.3 Contribution of the Research

As will be discussed in Chapter 2 and 3, the existing SKQ ([21]) is able to model the variability arising from quantitative as well as qualitative factors, and the heterogeneous variability of random errors. However, the SKQ estimation requires the target data to have multiple replications at each factor setting, which is needed for the estimation of heterogeneous error variances. In this study, we adapted the intrinsic (random error) variance structure and developed an iterative procedure to enable the fitting of SKQ to a non-replicated or partially non-replicated data set. The iterative SKQ is adapted to estimate both the mean and variance of flow time.

Moreover, real data availability is limited to settings observed from a manufacturing process. There is no control on the design of observed settings, and typically there are no replications available. On the other hand simulation data can be generated in designed settings with replications. Iterative SKQ can pool information from both read and simulation data for the improved estimation of flow time.

Appendix 5.1 provides more detailed comparison between this work and the most related literature.

1.4 Statement of the Research problem

To assist the lead time quotation upon the arrival of a new order, prediction models that quantify the dependence of flow time characteristics (i.e., mean and variance) upon the various shop-status factors, will be estimated from the ensemble of two types of data:

- Discrete-event simulation data, which can be designed to provide a good coverage in the design/input region and to include adequate replications at each design point.
- Real data from a manufacturing system, which cannot be controlled at the level of experimental design and are typically non-replicated.

Table 1.1: Original variables.

Type	Variables in X_{ORG}	Number of variables in the example system
Status Variables(SVs) Job Shop	SVs.A: The number of jobs at each buffer including those being processed and those waiting to be processed by the station	22
	SVs.B: The status (busy or idle) of each server.	11
	SVs.C: The elapsed processing time at each busy server.	11
	SVs.D: The status (up or down) of each server that is subject to random failures.	2
	SVs.E: The elapsed down time for a currently down server.	2
	SVs.F: The elapsed up time for a currently up server.	2
	SVs.G: The batch size currently being processed at a batch processing server, if that server is busy at the moment.	2
Order Size	The size of a newly arrived order	1
Future Orders	The forecasted arrival rate of future orders	1

The target manufacturing system may involve features such as random processing times, machine failures, batch processing, re-entrant flows, etc. As detailed in [1], the original predictive factors can be divided into three categories: (a) the shop status variables (SVs), (b) the size of a new order, and (c) the arrival rate of future orders, which can be obtained from forecasting models. A scaled-down manufacturing system is considered as the example system in this paper, with the detailed configurations given in Appendix 5.2 for readers' convenience. For this example system,

the original factors and factor numbers are provided in Table 1.1. The concept of buffers by [26] is used to define SVs. That is, for all the jobs that are in the same step of their production sequence, a virtual location called buffer is considered.

Due to the typically large number of factors included in Table 1.1, a preliminary analytical analysis is used to (Appendix 5.4) to find a smaller set of important variables, which can be classified as WIP and non-WIP variables. Table 1.2 provides for the example system the WIP and non-WIP variables as well as the additional qualitative factor z_S introduced in this work: data source, which could be simulation or real data. The variables in Table 1.2 constitute the vector $\mathbf{w}$, which serves as the input of the SKQ model. Based on both simulation and real data, SKQ is to be fitted quantifying the expected flow time as a function of $\mathbf{w}$.

Table 1.2: List of input variables in $\mathbf{w}$.

Type	Variables in $\mathbf{w}$	Number of variables in the example system
WIP Variable	$\mathbf{x}_{WIP}$ stage WIP variables a subset of SVs.A	8
Non-WIP Variables	$\mathbf{z}_B$ the busy or idle status of important servers, which constitute a subset of SVs.B	4
	$\mathbf{x}_C$ the elapsed processing times at important busy servers, which constitute a subset of SVs.C	4
	$\mathbf{z}_D$ the status (up or down) of important servers subject to random failures, which constitute a subset of SVs.D	1
	$\mathbf{x}_E$ the elapsed down times for important down servers, which constitute a subset of SVs.E	1
	$\mathbf{x}_F$ the elapsed up time for each important up server, which constitute a subset of SVs.F	1
	$\mathbf{x}_G$ the batch sizes being handled by important busy servers involving batch processing, which constitute a subset of SVs.G	0
	$\mathbf{x}_O$ the size of a newly arrived order	1
	$\mathbf{x}_R$ the forecasted arrival rate of future orders	1
	$\mathbf{z}_S$ the source of data	1

In our first step, we develop a model to obtain a high-quality prediction of mean flow time. To provide high-quality lead time quotation, we also model the variance of flow time as a function of shop status variables. Finally, a distribution is fitted based on the mean and variance estimates. This fitted distribution is used to quote the lead time.

1.5 Organization of the Dissertation

The remainder of this dissertation is organized as follows. In Chapter 2, we discuss the iterative frame work based on SKQ models to estimate mean flow time. The dual modeling to estimate the variance where the replications are not available is detailed in Chapter 3. In Chapter 4, the flow time distribution is quantified and estimated. Finally, Chapter 5 includes the summary and conclusions.

Chapter 2

Mean Estimation

For the estimation of mean flow time, an iterative method is developed based on stochastic kriging with qualitative factors(SKQ) to model both simulation and real manufacturing data. In Section 2.1, a review of related work is provided. Methods for the modeling of mean flow time are detailed in Section 2.2. In Section 2.3, the estimation results are evaluated.

2.1 Literature Review

Kriging (also known as Gaussian process regression)is highly flexible and able to provide an accurate approximation of practically any continuous response surfaces [21, 22, 23, 24] without requiring a pre-assumed functional form as traditional nonlinear regression does [21, 25].

In this stream of work, Ankenman et al. [22] developed stochastic kriging, which models both intrinsic uncertainty and the extrinsic uncertainty. Qian et al. [27] developed deterministic kriging with both qualitative and quantitative factors. Wang et al. [21] developed stochastic kriging with qualitative factors (SKQ), which models the variability arising from quantitative factors, qualitative factors, and heterogeneous random errors.

A range of research efforts have been devoted to developing kriging-based methods for modeling data of different fidelity levels [28][29] [30] [31] [32]

2.2 Methodology

2.2.1 Iterative Procedure for Stochastic Kriging with Qualitative Factors (SKQ)

The SKQ developed in [21] is able to model the variability arising from quantitative as well as qualitative factors, and the heterogeneous variability of random errors. However, the SKQ estimation requires the target data to have multiple replications at each factor setting, which is needed for the estimation of heterogeneous error variances. In this study, we adapted the intrinsic (random error) variance structure and developed an iterative procedure to enable the fitting of SKQ to a non-replicated or partially non-replicated data set.

The data are represented as

$$\{(\mathbf{w}_i, \mathscr{Y}_j(\mathbf{w}_i)); i = 1,2,\ldots,I; j = 1,2,\ldots,n(\mathbf{w}_i)\}, \tag{2.1}$$

with a total of I distinct factor settings. The i^{th} setting $\mathbf{w}_i = (\mathbf{x}_i^\top, \mathbf{z}_i^\top)^\top$ includes the specified levels for the quantitative factors $\mathbf{x}_i$ and the qualitative factors $\mathbf{z}_i$. At $\mathbf{w}_i$, the number of replications $n(\mathbf{w}_i)$ is greater than or equal to 1, and $n(\mathbf{w}_i) = 1$ corresponds to the factor settings with no replications.

Without loss of generality, the data (2.1) is arranged into two subsets

$$\{(\mathbf{w}_i, \mathscr{Y}_j(\mathbf{w}_i)); i = 1,2,\ldots,K; j = 1,2,\ldots,n(\mathbf{w}_i) > 1\} \bigcup \{(\mathbf{w}_i, \mathscr{Y}(\mathbf{w}_i)); i = K+1, K+2,\ldots,I\}. \tag{2.2}$$

The replicated subset includes K ($0 \leq K \leq I$) distinct factor settings with multiple replications $n(\mathbf{w}_i) > 1$, and at each of the rest $I-K$ factor settings, there is only a single replication.

The SKQ model is written as

$$\begin{aligned} \mathscr{Y}_j(\mathbf{w}) &= \mathrm{E}[\mathscr{Y}(\mathbf{w})] + \varepsilon_j(\mathbf{w}) = \mathsf{Y}(\mathbf{w}) + \varepsilon_j(\mathbf{w}) && (2.3) \\ &= \mathbf{f}(\mathbf{w})^\top \beta + \mathsf{M}(\mathbf{w}) + \varepsilon_j(\mathbf{w}), && (2.4) \end{aligned}$$

quantifying the dependence of the continuous response $\mathscr{Y}(\mathbf{w})$ upon the factors $\mathbf{w} = (\mathbf{x}^\top, \mathbf{z}^\top)^\top$ including the d quantitative factors $\mathbf{x} = (x_1, x_2, \ldots, x_d)^\top \in \mathbb{R}^d$ and the L qualitative factors $\mathbf{z} = (z_1, z_2, \ldots, z_L)^\top$. In (2.4), $f(\mathbf{w})$ is a vector of known functions of $\mathbf{w}$, and β is a vector of unknown

coefficients. In this work, we set $f(\mathbf{w})^T\beta = \beta_0$, which is usually adequate for kriging-based modeling. $\mathsf{M}(\mathbf{w})$ represents a mean-zero stationary Gaussian process, which seeks to describe the extrinsic variability [22]. $\varepsilon_j(\mathbf{w})$ denotes the random error variability, and is referred to as the intrinsic variability [22]. The random errors $\varepsilon_1(\mathbf{w})$, $\varepsilon_2(\mathbf{w}),\ldots,$ are assumed to be independent and identically distributed with mean zero.

2.2.2 Extrinsic Variance Structure

The extrinsic variability model is inherited from [27], and reviewed as follows. The covariance of $\mathsf{M}(\mathbf{w})$ can be written as

$$\mathrm{Cov}[\mathsf{M}(\mathbf{w}),\mathsf{M}(\mathbf{w}')] = \delta^2 \cdot \mathrm{Corr}[\mathsf{M}(\mathbf{w}),\mathsf{M}(\mathbf{w}')] = \delta^2 \cdot \left[\prod_{\ell=1}^{L} \tau^{(\ell)}_{z_\ell,z'_\ell}\right] \cdot K(\mathbf{x},\mathbf{x}'), \tag{2.5}$$

where δ^2 is the variance of the Gaussian process. The correlation $\mathrm{Corr}[\mathsf{M}(\mathbf{w}),\mathsf{M}(\mathbf{w}')]$ is decomposed into two parts: $\prod_{\ell=1}^{L} \tau^{(\ell)}_{z_\ell,z'_\ell}$ and $K(\mathbf{x},\mathbf{x}')$. For SKQ estimation, functional forms need to be specified for both parts. The correlation across different settings of $\mathbf{x}$ is represented by $K(\mathbf{x},\mathbf{x}')$, which can take a range of functional forms in the literature [33, 27]. A popular function is the exponential correlation function

$$K(\mathbf{x},\mathbf{x}') = \exp\left\{\sum_{h=1}^{d} -\theta_h |x_h - x'_h|^p\right\} \tag{2.6}$$

with $\theta = (\theta_1,\theta_2,\ldots,\theta_d)$ being unknown parameters.

In (2.5), the term $\prod_{\ell=1}^{L} \tau^{(\ell)}_{z_\ell,z'_\ell}$ models the correlations across different categories of qualitative factors, and the vector $\Phi = (\theta_1,\theta_2,\ldots,\theta_d)$ denotes the unknown parameters involved in the cross-category correlation model. Potential functional forms for $\tau^{(\ell)}_{z_\ell,z'_\ell}$ are given in [21]. Isotropic (or exchangeable) correlation functions (EC) is one of the common correlation functions:

$$\tau^{(l)}_{z_l,z'_l} = \exp\{-\phi^{(l)} I(z_l \neq z'_l))\}; l = 1,2,\ldots,L \tag{2.7}$$

In (2.7), $\phi = \phi^{(l)}; l = 1,2,\ldots,L$ represents the set of unknown parameters to be estimated; and $I[A]$ is an indicator function that takes the value of 1 if event A is true and 0 otherwise.

Given the data (2.1) collected at I distinct settings, the $I \times I$ variance-covariance matrix Σ_{M} is

defined as

$$\Sigma_{\mathsf{M}} = \delta^2 \cdot \mathbf{R}(\theta, \Phi) \tag{2.8}$$

$$= \delta^2 \cdot \begin{pmatrix} 1 & \mathrm{Corr}[\mathsf{M}(\mathbf{w}_1), \mathsf{M}(\mathbf{w}_2)] & \cdots & \mathrm{Corr}[\mathsf{M}(\mathbf{w}_1), \mathsf{M}(\mathbf{w}_I)] \\ \mathrm{Corr}[\mathsf{M}(\mathbf{w}_2), \mathsf{M}(\mathbf{w}_1)] & 1 & \cdots & \mathrm{Corr}[\mathsf{M}(\mathbf{w}_2), \mathsf{M}(\mathbf{w}_I)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Corr}[\mathsf{M}(\mathbf{w}_I), \mathsf{M}(\mathbf{w}_1)] & \mathrm{Corr}[\mathsf{M}(\mathbf{w}_I), \mathsf{M}(\mathbf{w}_2)] & \cdots & 1 \end{pmatrix}, \tag{2.9}$$

where $\mathbf{R}(\theta, \Phi)$ denotes the correlation matrix with each element being a correlation. Each element correlation can be decomposed into two parts as explained above, and involves the unknown parameters θ and Φ. For an arbitrary $\mathbf{w}_0$, the $I \times 1$ vector $\Sigma_{\mathsf{M}}(\mathbf{w}_0, \cdot)$ is defined as

$$\Sigma_{\mathsf{M}}(\mathbf{w}_0, \cdot) = \delta^2 \mathbf{v}(\mathbf{w}_0, \theta, \Phi) = \delta^2 \begin{pmatrix} \mathrm{Corr}[\mathsf{M}(\mathbf{w}_0), \mathsf{M}(\mathbf{w}_1)] \\ \mathrm{Corr}[\mathsf{M}(\mathbf{w}_0), \mathsf{M}(\mathbf{w}_2)] \\ \vdots \\ \mathrm{Corr}[\mathsf{M}(\mathbf{w}_0), \mathsf{M}(\mathbf{w}_i)] \end{pmatrix}, \tag{2.10}$$

where $\mathbf{v}(\mathbf{w}_0, \theta, \Phi)$ is a correlation vector involving $\mathbf{w}_0$, θ and Φ.

2.2.3 Intrinsic Variance Structure

The variance of the random error at $\mathbf{w}$ is denoted as $\mathrm{Var}[\varepsilon(\mathbf{w})]$. Let Σ_ε be the $I \times I$ intrinsic variance matrix. Under the i.i.d. assumption for random errors, Σ_ε for the data (2.2) is a diagonal matrix

$$\Sigma_\varepsilon = \mathrm{diag}\{\frac{\mathrm{Var}[\varepsilon(\mathbf{w}_1)]}{n(\mathbf{w}_1)}, \ldots, \frac{\mathrm{Var}[\varepsilon(\mathbf{w}_K)]}{n(\mathbf{w}_K)}, \mathrm{Var}[\varepsilon(\mathbf{w}_{K+1})], \ldots, \mathrm{Var}[\varepsilon(\mathbf{w}_i)]\}. \tag{2.11}$$

A similar model without an intrinsic variance structure can be considered as an deterministic kriging model (DKQ) with quantitative factors. This family of models were introduced by [27].

2.2.4 Integrative Estimation of Replicated and Non-replicated Data

Recall that the stochastic response on replication j at design point $\mathbf{w}$ was modeled as follows:

$$\mathscr{Y}_j(\mathbf{w}_i) = \beta_0 + \mathsf{M}(\mathbf{w}_i) + \varepsilon_j(\mathbf{w}_i) \tag{2.12}$$

Here the objective is to build a model to predict the response $\mathsf{Y}(\mathbf{w}_0) = \beta_0 + \mathsf{M}(\mathbf{w}_0)$ at a desired

point $\mathbf{w}_0$ using both replicated data and non-replicated data. The Gaussian process regression estimation and inference require the following assumptions:

Assumption 1:

The random field M is a stationary Gaussian random field, and $\varepsilon_1(\mathbf{w}), \varepsilon_2(\mathbf{w}), \ldots,$ are i.i.d. $N(0, \mathrm{Var}[\varepsilon(\mathbf{w}_i)])$, and independent of M.

Stationary Gaussian random field assumption for M is a standard assumption based on [34] and [22]. Since the response is the summation of individual product process times in the manufacturing example, the normality of $\varepsilon_j(\mathbf{w}_i)$ is anticipated [22]. In K replicated data points, the sample average of the responses at $\mathbf{w}_i$, $\mathscr{Y}_j(\mathbf{w}_i)$, across the $n(\mathbf{w}_i)$ replications follows as:

$$\overline{\mathscr{Y}}(\mathbf{w}_i) = \frac{1}{n(\mathbf{w}_i)} \sum_{j=1}^{n(\mathbf{w}_i)} \mathscr{Y}_j(\mathbf{w}_i) \qquad \text{for} \quad i = 1, \ldots, K. \tag{2.13}$$

And $\overline{\mathscr{Y}} = (\overline{\mathscr{Y}}(\mathbf{w}_1), \overline{\mathscr{Y}}(\mathbf{w}_2), \ldots, \overline{\mathscr{Y}}(\mathbf{w}_K))^\top$ denotes the vector of sample average at replicated points. Similarly, the sample average error follows as:

$$\overline{\varepsilon}(\mathbf{w}_i) = \frac{1}{n(\mathbf{w}_i)} \sum_{j=1}^{n(\mathbf{w}_i)} \varepsilon_j(\mathbf{w}_i) \qquad \text{for} \quad i = 1, \ldots, K \tag{2.14}$$

And $\overline{\varepsilon} = ((\overline{\varepsilon}(\mathbf{w}_1)), (\overline{\varepsilon}(\mathbf{w}_2), \ldots, (\overline{\varepsilon}(\mathbf{w}_K))^\top$ denotes the vector of sample average errors. In real data points, there is only one observation at each point. The vector of observed responses at $\mathbf{w}_i$ follows as:

$$\mathscr{Y}_{obs} = (\mathscr{Y}(\mathbf{w}_{K+1}), \mathscr{Y}(\mathbf{w}_{K+2}), \ldots, \mathscr{Y}(\mathbf{w}_i))^\top \qquad \text{for} \quad i = K+1, \ldots, I \tag{2.15}$$

And let the vector of observed errors denoted by

$$\varepsilon_{obs} = ((\varepsilon(\mathbf{w}_{K+1})), (\varepsilon(\mathbf{w}_{K+2}), \ldots, (\varepsilon(\mathbf{w}_I))^\top \qquad \text{for} \quad i = K+1, \ldots, I \tag{2.16}$$

Theorem 1:

$(\mathsf{Y}(\mathbf{w}_0), \overline{\mathscr{Y}}, \mathscr{Y}_{obs})^\top$ is multivariate normal with constant mean vector $\beta_0 \mathbf{1}_{I+1}$, and following variance-

covariance matrix:

$$\begin{pmatrix} \delta^2 & R(\mathbf{w}_0,.;\theta,\phi) \\ R(\mathbf{w}_0,.;\theta,\phi) & \Sigma_\varepsilon + \Sigma_{\mathsf{M}} \end{pmatrix} \tag{2.17}$$

Proof: Under Assumption 1 and model (2.12), $(\mathsf{Y}(\mathbf{w}_0), \overline{\mathscr{Y}}, \mathscr{Y}_{obs})^\top$ is multivariate normal. Within replicated data:

$$\begin{aligned} \mathrm{Cov}[\mathscr{Y}_j(\mathbf{w}_i), \mathscr{Y}_l(\mathbf{w}_h)] &= \mathrm{Cov}[\mathsf{M}(\mathbf{w}_i) + \varepsilon_j(\mathbf{w}_i), \mathsf{M}(\mathbf{w}_h) + \varepsilon_l(\mathbf{w}_h)] \\ &= \begin{cases} \delta^2 + \mathrm{Var}[\varepsilon(\mathbf{w}_i)] & i = h, j = l \\ \delta^2 & i = h, j \neq l \\ \delta^2 R(\mathbf{w}_i, \mathbf{w}_h; \theta, \phi) & i \neq h \end{cases} \end{aligned} \tag{2.18}$$

Within non-replicated data:

$$\begin{aligned} \mathrm{Cov}[\mathscr{Y}(\mathbf{w}_i), \mathscr{Y}(\mathbf{w}_h)] &= \mathrm{Cov}[\mathsf{M}(\mathbf{w}_i) + \varepsilon(\mathbf{w}_i), \mathsf{M}(\mathbf{w}_h) + \varepsilon(\mathbf{w}_h)] \\ &= \begin{cases} \delta^2 + \mathrm{Var}[\varepsilon(\mathbf{w}_i)] & i = h \\ \delta^2 R(\mathbf{w}_i, \mathbf{w}_h; \theta, \phi) & i \neq h \end{cases} \end{aligned} \tag{2.19}$$

Between predicted response and replicated data:

$$\mathrm{Cov}[\mathsf{Y}(\mathbf{w}_0), \mathscr{Y}_j(\mathbf{w}_i)] = \mathrm{Cov}[\mathsf{M}(\mathbf{w}_0), \mathsf{M}(\mathbf{w}_i) + \varepsilon_j(\mathbf{w}_i)] = \delta^2 R(\mathbf{w}_0, \mathbf{w}_i; \theta, \phi) \tag{2.20}$$

Between predicted response and non-replicated data:

$$\mathrm{Cov}[\mathsf{Y}(\mathbf{w}_0), \mathscr{Y}_{obs}(\mathbf{w}_i)] = \mathrm{Cov}[\mathsf{M}(\mathbf{w}_0), \mathsf{M}(\mathbf{w}_i) + \varepsilon(\mathbf{w}_i)] = \delta^2 R(\mathbf{w}_0, \mathbf{w}_i; \theta, \phi) \tag{2.21}$$

Since the $\varepsilon_j(\mathbf{w}_i)$ are independent across replications and design points, averaging the $n(\mathbf{w}_i)$ replications at design point $\mathbf{w}_i$ only affects

$$\mathrm{Cov}[\overline{\mathscr{Y}}(\mathbf{w}_i), \overline{\mathscr{Y}}(\mathbf{w}_h)] = \mathrm{Var}[\overline{\mathscr{Y}}(\mathbf{w}_i)] = \delta^2 + \frac{\mathrm{Var}[\varepsilon(\mathbf{w}_i)]}{n(\mathbf{w}_i)} \tag{2.22}$$

Therefore, following holds:

$$\begin{pmatrix} \mathsf{Y}(\mathbf{w}_0) \\ \overline{\mathscr{Y}} \\ \mathscr{Y}_{obs} \end{pmatrix} = \text{MVN}[\beta_0 \mathbf{1}_{I+1}, \begin{pmatrix} \delta^2 & R(\mathbf{w}_0,.;\theta,\phi) \\ R(\mathbf{w}_0,.;\theta,\phi) & \delta^2 R(\mathbf{w}_i,\mathbf{w}_j;\theta,\phi) + Diag\{\frac{\text{Var}[\varepsilon(\mathbf{w}_i)]}{n(\mathbf{w}_i)}, \text{Var}[\varepsilon(\mathbf{w}_i)]\} \end{pmatrix}]. \tag{2.23}$$

where 1_{I+1} is a $A1$ vector of ones. Considering the definition of Σ_ε and Σ_M in (2.8) and (2.11) respectively, the variance-covariance matrix can be shown as:

$$\begin{pmatrix} \mathsf{Y}(\mathbf{w}_0) \\ \overline{\mathscr{Y}} \\ \mathscr{Y}_{obs} \end{pmatrix} = \text{MVN}[\beta_0 \mathbf{1}_{I+1}, \begin{pmatrix} \delta^2 & R(\mathbf{w}_0,.;\theta,\phi) \\ R(\mathbf{w}_0,.;\theta,\phi) & \Sigma_\mathsf{M} + \Sigma_\varepsilon \end{pmatrix}]. \tag{2.24}$$

Since $(\mathsf{Y}(\mathbf{w}_0), \overline{\mathscr{Y}}, \mathscr{Y}_{obs})^\top$ is multivariate normal , the following stochastic predictor can be used to estimate the expected response at point $\mathbf{w}_0$ [21, 22, 34].

$$\widehat{\mathsf{Y}}(\mathbf{w}_0) = \beta_0 + \Sigma_\mathsf{M}(\mathbf{w}_0,.)^\top [\Sigma_\mathsf{M} + \Sigma_\varepsilon]^{-1} (\mathscr{Y} - \beta_0 \mathbf{1}_I)) \tag{2.25}$$

According to [21] and [22] it can be shown that (2.25) is the best linear unbiased estimator for $\mathsf{Y}(\mathbf{w}_0)$.

2.2.5 Iterative Procedure for Model Estimation

In the data set (2.2), the vector of sample averages for replicated data is denoted as

$$\overline{\mathscr{Y}} = (\overline{\mathscr{Y}}(\mathbf{w}_1), \overline{\mathscr{Y}}(\mathbf{w}_2), \ldots, \overline{\mathscr{Y}}(\mathbf{w}_K))^\top \tag{2.26}$$

with

$$\overline{\mathscr{Y}}(\mathbf{w}_i) = \frac{1}{n(\mathbf{w}_i)} \sum_{j=1}^{n(\mathbf{w}_i)} \mathscr{Y}_j(\mathbf{w}_i) \qquad\qquad i = 1, \ldots, K. \tag{2.27}$$

Where $\mathscr{Y}_j(\mathbf{w}_i)$ is the observed response from the jth replication at $\mathbf{w}_i$.

The vector of single observations for non-replicated data is written as

$$\mathscr{Y}_{obs} = (\mathscr{Y}(\mathbf{w}_{K+1}), \mathscr{Y}(\mathbf{w}_{K+2}), \ldots, \mathscr{Y}(\mathbf{w}_i))^\top \tag{2.28}$$

The random vector $\mathscr{Y} = (\overline{\mathscr{Y}}^\top, \mathscr{Y}_{obs}^\top)^\top$ follows multivariate normal (MVN) distribution

$$\mathscr{Y} \sim \text{MVN}[\beta_0 \mathbf{1}_I, \Sigma_{\text{M}} + \Sigma_\varepsilon] \tag{2.29}$$

The log-likelihood function with respect to the unknown parameters $(\beta_0, \delta, \theta, \phi)$ is thus written as:

$$\ln L(\beta_0, \delta^2, \theta, \phi) = -\ln[(2\pi)^{\frac{I}{2}}] - \frac{1}{2}\ln[|\delta^2 R(\theta,\phi) + \Sigma_\varepsilon|] - \frac{1}{2}(\mathscr{Y} - \beta_0 \mathbf{1}_I)^\top [\delta^2 R(\theta,\phi) + \Sigma_\varepsilon]^{-1} (\mathscr{Y} - \beta_0 \mathbf{1}_I). \tag{2.30}$$

Since the data set (2.2) involves non-replicated data, some variance components in Σ_ε (corresponding to the non-replicated data) cannot be straightforwardly estimated and replaced by their sample variances. To circumvent that, we adapted the SKQ estimation/inference in [21] into the following iterative procedure for SKQ fitting of both replicated and non-replicated data.

Stage 1: Obtain an estimate of the intrinsic variance matrix Σ_ε.

- For replicated data, estimate $\text{Var}[\varepsilon(\mathbf{w}_i)]$ $(i = 1, 2, \ldots, K)$ by

$$\widehat{\text{Var}}[\varepsilon(\mathbf{w}_i)] = \frac{1}{n(\mathbf{w}_i) - 1} \sum_{j=1}^{n(\mathbf{w}_i)} (\mathscr{Y}_j(\mathbf{w}_i) - (\overline{\mathscr{Y}}(\mathbf{w}_i))^2. \tag{2.31}$$

- For non-replicated data, set $\widehat{\text{Var}}[\varepsilon(\mathbf{w}_i)] = v_0$ for $i = K+1, \ldots, I$. The initial variance estimate v_0 can be set as the median of the sample variances $\{\widehat{\text{Var}}[\varepsilon(\mathbf{w}_i)]; i = 1, 2, \ldots, K\}$.

- Assemble $\{\frac{\widehat{\text{Var}}[\varepsilon(\mathbf{w}_i)]}{n(\mathbf{w}_i)}; i = 1, 2, \ldots, K\}$ and v_0 to obtain the initial estimate $\widehat{\Sigma}_\varepsilon$.

Stage 2: Estimate the hyperparameters by solving the maximum likelihood problem. Replace Σ_ε by $\widehat{\Sigma}_\varepsilon$ in (2.30), and maximize the log-likelihood function with respect to $(\beta_0, \delta, \theta, \phi)$, which can be achieved in two steps.

- Given δ,θ and ϕ, the maximum likelihood estimate (MLE) of β_0 is derived from

$$\frac{\partial \ln L(\beta_0, \delta^2, \theta, \phi)}{\partial \beta_0} = 0, \tag{2.32}$$

and expressed as

$$\widehat{\beta}_0(\delta^2,\theta,\phi) = (1_I^T[\delta^2 R(\theta,\phi)+\widehat{\Sigma}_\varepsilon]^{-1}1_I)^{-1}(1_I^T[\delta^2 R(\theta,\phi)+\widehat{\Sigma}_\varepsilon]^{-1}\mathscr{Y}) \tag{2.33}$$

- Substitute $\widehat{\beta}_0(\delta^2,\theta,\phi)$ into (2.30) and maximize

$$\begin{aligned}\ln L(\delta^2,\theta,\phi) = &-\ln[(2\pi)^{\frac{I}{2}}] - \frac{1}{2}\ln[|\delta^2 R(\theta,\phi)+\widehat{\Sigma}_\varepsilon|] \\ &-\frac{1}{2}(\mathscr{Y}-\widehat{\beta}_0(\delta^2,\theta,\phi)1_I)^T[\delta^2 R(\theta,\phi)+\widehat{\Sigma}_\varepsilon]^{-1}(\mathscr{Y}-\widehat{\beta}_0(\delta^2,\theta,\phi)1_I)\end{aligned} \tag{2.34}$$

  with respect to (δ^2,θ,ϕ).

Stage 3: With the MLE $(\widehat{\beta}_0,\widehat{\delta^2},\widehat{\theta},\widehat{\phi})$, estimate the expected responses at the non-replicated factor settings as

$$\widehat{\Upsilon}(\mathbf{w}_i) = \widehat{\beta}_0 + \widehat{\delta}^2 \nu(\mathbf{w}_i,\widehat{\theta},\widehat{\phi})^T[\widehat{\Sigma}_M+\widehat{\Sigma}_\varepsilon]^{-1}(\mathscr{Y}-\widehat{\beta}_0\mathbf{1}_I)). \tag{2.35}$$

Stage 4: Update the variance estimates for non-replicated data.

- Based on the estimates obtained from (2.35), calculate the squared residuals:

$$\widehat{e}^2(\mathbf{w}_i) = (\mathscr{Y}(\mathbf{w}_i)-\widehat{Y}(\mathbf{w}_i))^2 \quad i = K+1,K+2,\ldots,I \tag{2.36}$$

- Update the estimate $\widehat{\Sigma}_\varepsilon$ by replacing its non-replicated components by $\widehat{e}^2(\mathbf{w}_i)$ for $i = K+1,\ldots,I$.

- Repeat Stages 2-4 until there is no significant changes in the parameter estimates $(\widehat{\beta}_0,\widehat{\delta^2},\widehat{\theta},\widehat{\phi})$.

2.3 Empirical Results

To demonstrate the information-pooling effects of SKQ, a simulation-based case study is designed as follows. Two DES models were coded in Microsoft Visual $C++$. They share the same configuration of a scaled-down manufacturing system, and only differ in some processing-time parameters at certain workstations. The DES model specified in Appendix 5.3 will be referred to as *DES_Real*

representing the target real system, and *DES_Real* is used to generate data mimicking system observations that cannot be designed with control but are real. The DES model detailed in Appendix 5.2 will be referred to as *DES_Sim*, which serves as the high-fidelity simulation model of the "real system" *DES_Real* while slightly deviating from the reality . Experimental design strategies are applied to *DES_Sim* for the collection of well-designed data.

The preliminary analytical analysis by [35] is first performed to identify **w**, a set of relatively important variables. In this case, **w** includes 16 quantitative and 6 qualitative variables.

2.3.1 Estimation Data (ED)

The estimation data set includes two subsets: *ED_Real* and *ED_Sim*, which are described as follows.

ED_Real: a data set which is typically obtained from observing or tracking a real system. 32 distinct points were generated in the space of **w** following some random scheme. At each point, a single simulation run was carried out to obtain a CT observation.

ED_Sim: a data set resulting from simulation experiments. The simulation design points are two folds. First, the well-designed 120 distinct points generated in the space of **w** following the experimental design method developed by [1], which seeks to optimize the D-criterion while providing a decent coverage of the design space. This part of the simulation data gathering can be performed off-line before receiving the orders. The second part of the simulation design points are real design points observed in *ED_Real*. At each design point, multiple replications were performed to enable the estimation of heterogeneous variance. The number of replications at a design point was determined by the two-stage process following [1] method, and ranges from 50 to 400 among the 120 design points.

For the estimation data sets the WIP levels are 15,30,45, and 60. The future orders arrival rate $\mathbf{x}_R$ takes three levels, which correspond to a steady-state real system utilization of 75%, 80%, and 90%, respectively. With selected utilization rates, the arrival rates are 0.129, 0.136, and 0.150 orders per hour. Table 5.4 shows the level or other Non-WIP variables.

2.3.2 Validation Data (VD)

The goal is to obtain a prediction model relating the mean CT to **w** for the target system, *DES_Real*. Thus, *DES_Real* was used to generate the VD, for the evaluation of fitted models. A total of 2400 check points were generated in the **w** space providing a dense and fairly even coverage of the design space. The WIP levels are 20, 25, 35, 40,50, and 55 which creates checking points all different from the points in *ED_Real* or *ED_Sim*. The arrival rates of future orders are 0.129, 0.136, and 0.150 orders per hour which is equivalent to 75%, 80%, and 90% utilization rate, respectively. At a check point **w**, 1000 replications were carried out, from which a highly accurate estimate of the mean FT can be obtained and denoted as $\mathsf{Y}_T(\mathbf{w})$. $\mathsf{Y}_T(\mathbf{w})$ is considered as nearly free of errors and serves as the "true" expected FT for the assessment of prediction models.

2.3.3 Model Evaluation Criteria

The quality of a prediction model is evaluated by the deviations of its estimated responses from their true values. With the "true" expected FT $\mathsf{Y}_T(\mathbf{w})$ obtained from the VD, the following two criteria are employed here for model assessment.

The mean absolute percentage error (MAPE):

$$\text{MAPE} = \frac{100\%}{2400}\sum_{i=1}^{2400}\left|\frac{\widehat{\mathsf{Y}}(\mathbf{w}_i)-\mathsf{Y}_T(\mathbf{w}_i)}{\mathsf{Y}_T(\mathbf{w}_i)}\right| \tag{2.37}$$

The estimated root mean squared error(ERMSE):

$$\text{ERMSE} = \sqrt{\frac{1}{2400}\sum_{i=1}^{2400}(\widehat{\mathsf{Y}}(\mathbf{w}_i)-\mathsf{Y}_T(\mathbf{w}_i))^2} \tag{2.38}$$

In (2.37) and (2.38), $\widehat{\mathsf{Y}}(\mathbf{w}_i)$ is the estimated mean FT at a check point $\mathbf{w}_i$.

2.3.4 Comparison of Modeling Methods

Three prediction models are respectively obtained through three different venues.

- Iterative SKQ on *ED_Real* and *ED_Sim*: The iterative SKQ procedure (Section 2.2.5) was applied to model the data ensemble of *ED_Real* and *ED_Sim* with the predictors being **w**

including the qualitative variable that has two categorical levels, Real or Simulation.

- Iterative SKQ on *ED_Real*: On *ED_Real* alone, the iterative SKQ procedure was applied with the predictors being **w** excluding the qualitative variable for real or simulation data.

- Regression on *ED_Real*: On *ED_Real* alone, the linear regression by [1] was applied with the predictors being **w** excluding the qualitative variable for real or simulation data.

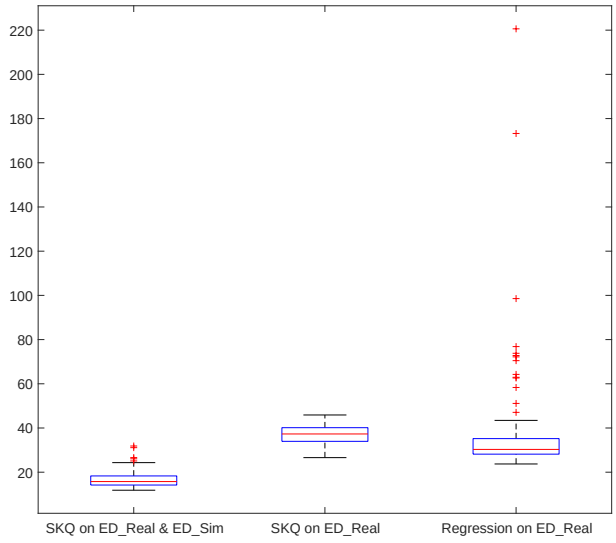
To statistically compare these three approaches, 100 macro-replications were performed. For each macro-replication, design points for *ED_Sim* and observation points for *ED_Real* were regenerated following the schemes as briefed in 2.3.1, and simulation runs were carried for data collection using a different random stream; with the obtained *ED_Sim* and *ED_Real*, all three approaches were applied respectively. Thus, each of the three approaches leads to 100 fitted models (e.g., regression models), and 100 MAPEs and ERMSEs (2.3.3).

Figure 2.1(a) and (b) display the MAPE and ERMSE box plots respectively for the three approaches. Each box is plotted from the 100 MAPEs or ERMSEs for the corresponding approach. The medians of the boxes are also given in Table 2.1. Clearly, by borrowing information from the well-designed simulation data *ED_Sim*, the iterative SKQ achieves the fitted models of the smallest deviations and most consistent performance, which are evident from the lowest and narrowest boxes for "Iterative SKQ on *ED_Real* and *ED_Sim*" in Figure 2.1. From the same scarce "real" data *ED_Real*, the iterative SKQ leads to better fitted models than the regression method with boxes of close heights (medians) and substantially narrower boxes and whiskers. The medians of the MAPEs and ERMSEs obtained from these three approaches are given in Table 2.1

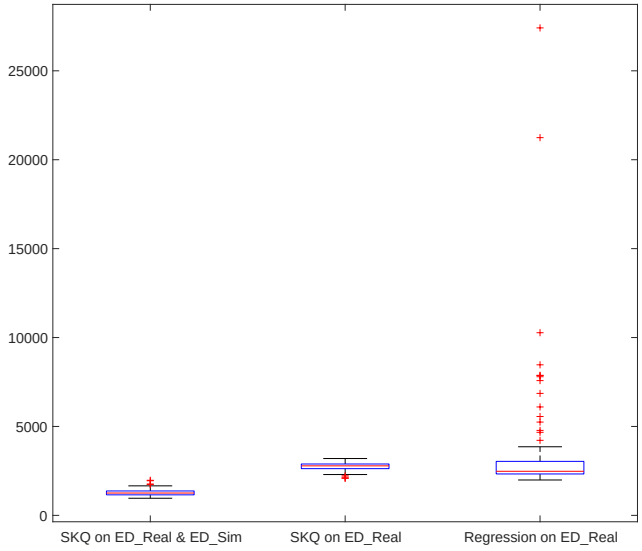
Table 2.1: Medians of MAPEs and ERMSEs from macro-replications for mean models, WIP level [15 60]

Method	MAPE	ERMSE
Iterative SKQ on *ED_Real* and *ED_Sim*	15.76%	1255.8
Iterative SKQ on *ED_Real*	37.30%	2784.4
Regression on *ED_Real*	30.31%	2479.2

As we will discuss in next chapter, the variance model works better under lower levels of WIP, So we decreased the WIP level into a range of [15 42]. Table 2.2 shows the MAPEs and ERMSEs for the new WIP levels. All methods are more stable in the narrower WIP levels.

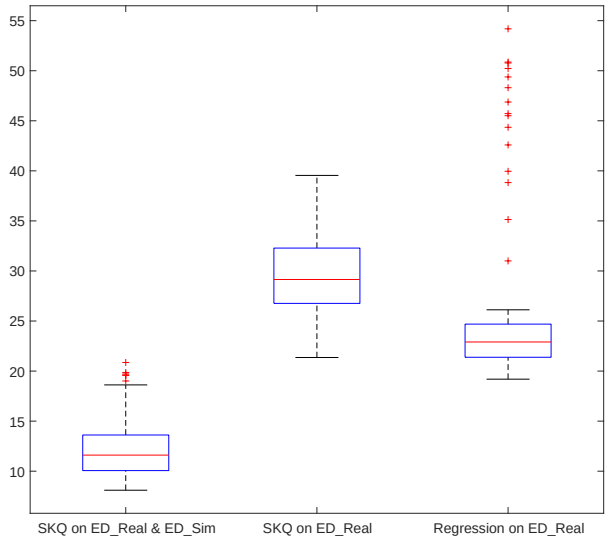


(a) MAPE box plots.

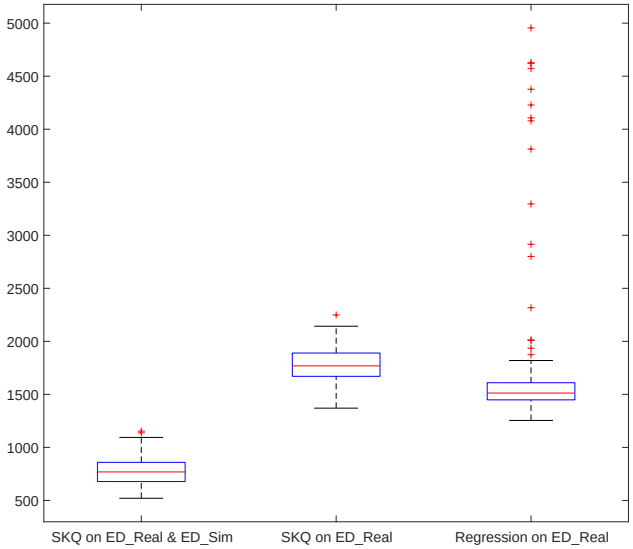


(b) ERMSE box plots.

Figure 2.1: Comparison of mean models quality for WIP level [15 60].



(a) MAPE box plots.



(b) ERMSE box plots.

Figure 2.2: Comparison of mean models quality for WIP level [15 42].

Table 2.2: Medians of MAPEs and ERMSEs from macro-replications for mean models, WIP level [15 42].

Method	MAPE	ERMSE
Iterative SKQ on *ED_Real* and *ED_Sim*	11.60%	768.8
Iterative SKQ on *ED_Real*	29.15%	1768.9
Regression on *ED_Real*	22.90%	1512.5

Chapter 3

Variance Estimation

In this research, not only the first but also the second moment of the flow time is of interest. The flow time variance is particularly important to ensure the reliability of a quoted lead time. To provide high-quality lead time quotation, we model both the mean and variance of flow time as a function of the shop status variables. In this chapter we use a dual modeling frame work based on kriging method to estimate the mean and variance simultaneously. Both simulation and real data are integrated to improve the fitting of the target response surface for the real system.

The remainder of this chapter is organized as follows. Section 3.1 provides a review of the most related work. An iterative procedure to model the heteroscedastic variance is given in Section 3.2. In Section 3.3, the methods are evaluated via empirical results.

3.1 Literature Review

To model the response variance as a function of independent variables, usually multiple replications are required to provide variance estimates at various factor settings [22]. In the absence of replications, two main approaches have been developed in the literature for variance modeling: difference-based methods and resampling methods.

3.1.1 Difference-based Methods

Difference-based method estimates variances based on squared residuals from an initial fit of the mean model: a mean model is fitted first, and then the variance model is built on the residuals

obainted from the estimated means and observations.

In this stream of work, Carroll [36] considered a dual model with a parametric mean function and a kernel regression variance model. The author studied a linear model for the mean with one regressor and suggested using squared residuals from the means to fit the variance model.

Gasser et al.[37] considered the accuracy of the mean model and its effect on the variance model in the dual modeling structure. They used nonparametric models for both mean and variance, and proposed a variance estimate which is independent of the fitted mean model. The variance estimates are pseudo residuals calculated from neighbor design points.

Muller and Stadtmuller[38] extended the approach presented by Gasser et al. [37]. They suggested a new difference-based scheme and estimated the local variances using several neighborhood design points. They considered a set of weights for each neighborhood design point. Later, Muller[39] expanded their previous work and considered quadratic forms for variance models.

Hall and Carroll[40] also studied the problem of estimating the variance function in regression problems. They declared that "such estimation requires simultaneous estimation of the mean and variance functions". They also considered nonparametric models for both mean and variance and discussed the effect of not knowing the mean function on the estimation of variance. They used squared residuals from the mean function to fit the variance function.

Herrmann[41] provided a dual nonparametric modeling for mean and variance. A new bandwidth selector with local variable bandwidth kernel estimators are proposed to include heteroscedasticity. They used a simulation study to evaluate the proposed method. Fan and Yao[42] also modeled the dual problem with nonparametric model for both mean and variance function. They also used residuals from the mean model and estimated the variance model by using local polynomial smoothing of the squared residuals. They evaluated the proposed method with financial time series data.

Opsomer et al. [43] presented an iterative procedure for dual modeling. They assumed a linear model for the mean and kriging variance model. Brown et al. [44] considered a dual modeling problem with Gaussian nonparametric regression for both mean and variance models. They proposed a class of difference-based kernel estimators for the variance model.

Wang et al. [45] also considered the mean and variance modeling in nonparametric regression. They studied the effect of mean model on the estimation of the variance model. Nonparametric models were considered for both mean and variance, and the minimax rate of convergence was

derived. They showed that the residual-based estimator performed better than the difference estimator when the mean function is very smooth. However, when the mean function is not smooth, the difference-based estimator is significantly better.

Cai et al. [46] expanded Wang et al.'s [45] work and proposed a wavelet thresholding approach to adaptive variance modeling in the heteroscedastic nonparametric regression model. Robinson et al. [47] proposed a semi-parametric dual modeling approach to simultaneously model the mean and variance when no replication is available. Their semi-parametric dual modeling approach combines a nonparametric fit for the mean component and a parametric fit for the squared residuals to model the variance.

Marrel et al. [48] applied a joint mean and variance modeling framework for heteroscedastic data. They started with a homoscedastic model for the mean followed by the iterative method of Robinson et al. [47]. Gaussian process regression was used for modeling both mean and variance. In their study, variable settings are controllable for sampling data. The authors used all the possible experiments to cover more space instead of having replications.

Navaee et al. [49] presented a dual semi-parametric modeling approach. The proposed dual model robust regression (DMRR), is robust against user misspecification of the mean variance models. They started with fitting a nonparametric model to variance using replicated data. Next, an expected weighted least square technique is applied to model. Finally the residuals of the fitted mean model are used to build the robust mean model.

3.1.2 Resampling Methods

The second stream of approaches employs resampling methods to estimate variance in dual modeling. Goldberg et al. [50] used a GP regression to model the mean and another independent GP to model the logarithms of noise levels. They applied a Markov chain Monte Carlo method and simulated a sample based on the predictive distribution of the mean model. The sample is used to estimate the noise level at each point and fit the GP model for the noise level.

Le et al. [51] followed Goldberg et al. [50] and presented an algorithm for dual modeling. They applied nonparametric regression model and suggested a method to maximize the posteriori estimation of the exponential parameters.

Kersting et al. [52] proposed a framework similar to Goldberg et al. [50]. They introduced

a dual model framework with Gaussian process regression models for both mean and variance. In their method, first a homoscedastic Gaussian Process regression was fitted for the mean model (GP1) and sample data were simulated based on the predictive distribution. However, to avoid the significant computational cost of Markov chain Monte Carlo, they used a most likely value of the variance at each point. The simulated data set was used to estimate an empirical noise level at each point and fit a second GP2 model for it. Finally a combined GP (GP3) was fitted using the GP2 to predict the logarithmic noise levels. This process was repeated until convergence.

Boukouvalas and Cornford[53] developed a method based on Kersting et al. [52] to perform the dual Gaussian Process regression on computer data with replicated observations at some selected points. They made some corrections to remove the bias due to the log transformation in Kersting et al. [52].

Titsias and Lazaro[54] improved Kersting et al. [52] and presented a non-standard variation estimation to enable inferences in heteroscedastic GPs. Their framework applies Bayesian approach and maximizes an analytically tractable lower bound on the exact marginal likelihood.

3.2 Methodology

As discussed in Chapter 2.2.5, the SKQ developed in Wang et al. [21] is able to model the variability arising from quantitative as well as qualitative factors, and the heterogeneous variability of random errors. However, the SKQ estimation requires the target data to have multiple replications at each factor setting. In this Chapter we use resampling and difference based methods and extend the framework developed in Chapter 2.2.5 for both mean and variance modeling.

3.2.1 Resampling-based Variance Estimation

Employing resampling, an iterative procedure is adapated to model the mean and variance of flow time. In this procedure, an initial SKQ is trained based on the available data. Next, the predictive distribution of the current Gaussian Process is used for resampling and subsequent variance estimation. Herein, we extend Kersting et al. [52] work to a multi data source environment. First, the developed method in Chapter 2 is applied to infuse two sources of data and estimate the mean model. Then a resampling process is applied to generate a resample for the real system. Finally,

a new kriging model is fitted for the variance based on the new sample generated. We also follow] Boukouvalas and Cornford[53] corrections to estimate the log transformation bias. The iterative procedure developed in Section 2.2.5 is adapted as follows.

Stage 1: Obtain an estimate of the intrinsic variance matrix Σ_ε as in Section 2.2.5

Stage 2: Estimate the hyperparameters by solving the maximum likelihood problem. Replace Σ_ε by $\widehat{\Sigma}_\varepsilon$ in (2.30), and maximize the log-likelihood function with respect to $(\beta_0, \delta, \theta, \phi)$.

Stage 3: With the MLE $(\widehat{\beta_0}, \widehat{\delta^2}, \widehat{\theta}, \widehat{\phi})$, estimate the expected responses at the non-replicated factor settings as

$$\widehat{\mathrm{Y}}(\mathbf{w}_i) = \widehat{\beta}_0 + \widehat{\delta}^2 \nu(\mathbf{w}_i, \widehat{\theta}, \widehat{\phi})^T [\widehat{\Sigma}_M + \widehat{\Sigma}_\varepsilon]^{-1} (\mathscr{Y} - \widehat{\beta}_0 \mathbf{1}_I)). \tag{3.1}$$

The mean squared error (MSE) also can be obtained as in [21]:

$$\widehat{MSE}[\widehat{\mathrm{Y}}(\mathbf{w}_i)] = \widehat{\delta}^2 \nu(\mathbf{w}_i, \widehat{\theta}, \widehat{\phi})^T [\widehat{\Sigma}_M + \widehat{\Sigma}_\varepsilon]^{-1} \nu(\mathbf{w}_i, \widehat{\theta}, \widehat{\phi}) + \eta^2 (1_I^T [\widehat{\Sigma}_M + \widehat{\Sigma}_\varepsilon]^{-1} 1_I)^{-1} \tag{3.2}$$

where

$$\eta = 1 - 1_I^T [\widehat{\Sigma}_M + \widehat{\Sigma}_\varepsilon]^{-1} \nu(\mathbf{w}_i, \widehat{\theta}, \widehat{\phi}) \widehat{\delta}^2 \tag{3.3}$$

Stage 4: For each non-replicated data point,

- Randomly sample new data points from the normal distribution

$$N(\widehat{\mathrm{Y}}(\mathbf{w}_i), \widehat{MSE}[\widehat{\mathrm{Y}}(\mathbf{w}_i)] + \widehat{\Sigma}_\varepsilon) \tag{3.4}$$

The new resampled data are represented as:

$$\{(\mathbf{w}_i, \mathscr{Y}_s(\mathbf{w}_i)); i = K+1, K+2, \ldots, I; s = 1, 2, \ldots, S_p\} \tag{3.5}$$

- Update the variance estimates for non-replicated data based on the estimator suggested by

[52]

$$\widehat{e^2}(\mathbf{w}_i) = \frac{1}{S_p}\sum_{s=1}^{S_p}\frac{(\mathscr{Y}_s(\mathbf{w}_i) - \mathscr{Y}_{obs})^2}{2}. \quad i = K+1, K+2, \ldots, I \tag{3.6}$$

- Take the logarithm of the sample variances (for both simulation and real systems) and add the bias corrections suggested by [53]

$$\mathbf{r}(\mathbf{w}_i) = log(S_i^2) + (d + dlog(2) - \Psi(d/2))^{-1} \tag{3.7}$$

where S_i^2 is the sample variance. For replicated data, the sample variance ,$\widehat{\text{Var}}[\varepsilon(\mathbf{w}_i)]$, is calculated as in (2.31). For non-replicated data ,S_i^2 , is $\widehat{e^2}(\mathbf{w}_i)$ as in (3.6). In (3.7), d is the number of samples, and the Ψ digamma function.

Stage 5:

- Fit a deterministic kriging model (DKQ) [27] with qualitative factors to the variance data set

$$\{(\mathbf{w}_i, \mathbf{r}(\mathbf{w}_i)); i = 1, 2, \ldots, I\} \tag{3.8}$$

- Update the estimate $\widehat{\Sigma}_\varepsilon$ by replacing its non-replicated components by $\widehat{e^2}(\mathbf{w}_i)$ for $i = K+1, \ldots, I$.

- Repeat Stages 2-5 until there is no significant changes in the DKQ parameter estimates.

3.2.2 Difference-based Variance Estimation

In this section, we employ difference-based methods [36, 37, 38, 55] for variance estimates in the iterative procedure for dual modeling of flow time. The iterative procedure developed in Section 2.2.5 is adapted as follows.

Stage 1: Obtain an estimate of the intrinsic variance matrix Σ_ε as in section 2.2.5.

Stage 2: Estimate the hyperparameters by solving the maximum likelihood problem as explained in 2.2.5.

Stage 3: With the MLE $(\widehat{\beta_0}, \widehat{\delta^2}, \widehat{\theta}, \widehat{\phi})$, estimate the expected responses at the non-replicated factor settings as in (2.35).

Stage 4: For each non-replicated data point,

- Based on the estimates obtained from (2.35), calculate the squared residuals:

$$\widehat{e}^2(\mathbf{w}_i) = (\mathscr{Y}(\mathbf{w}_i) - \widehat{Y}(\mathbf{w}_i))^2 \quad i = K+1, K+2, \ldots, I \tag{3.9}$$

- Take the logarithm of the estimated variance (both simulation and real data).

$$\mathbf{p}(\mathbf{w}_i) = log(T_i^2) \tag{3.10}$$

where T_i^2 is the estimated variance. For replicated data, the estimated variance ,$\widehat{\text{Var}}[\varepsilon(\mathbf{w}_i)]$, is calculated as in (2.31). For non-replicated data ,T_i^2 , is $\widehat{e}^2(\mathbf{w}_i)$ as in 3.9.

Stage 5:

- Fit a deterministic Kriging model with qualitative factors to the variance data set

$$\{(\mathbf{w}_i, \mathbf{p}(\mathbf{w}_i)); i = 1, 2, \ldots, I\} \tag{3.11}$$

- Update the estimate $\widehat{\Sigma}_\varepsilon$ by replacing its non-replicated components by $\widehat{e}^2(\mathbf{w}_i)$ for $i = K+1, \ldots, I$.

- Repeat Stages 2-5 until there is no significant changes in the DKQ parameter estimates.

3.3 Empirical Results

To assess the variance estimation procedures, the case in Chapter 2 was used: two DES models , *DES_Real* and *DES_Sim*, representing the real and simulation systems respectively. The estimation data (ED) set in 2.3.1 was used in this chapter, and the validation data (VD) set in 2.3.2 has been expanded to include the "true" variances at check points. Based on what is recommended in [21], the resampling size was set as S_p= 999.

3.3.1 Comparison of Variance Estimation Results

The variance estimation results obtained from the following scenarios are compared.

- Iterative RS on *ED_Real* and *ED_Sim*: The iterative resampling procedure for variance model (Section 3.2.1) was applied to model the data ensemble of *ED_Real* and *ED_Sim* with the predictors being **w** including the qualitative variable for real or simulation data.

- Iterative DB on *ED_Real* and *ED_Sim*: The iterative difference-based procedure for variance model (Section 3.2.2) was applied to model the data ensemble of *ED_Real* and *ED_Sim* with the predictors being **w** including the qualitative variable for real or simulation data.

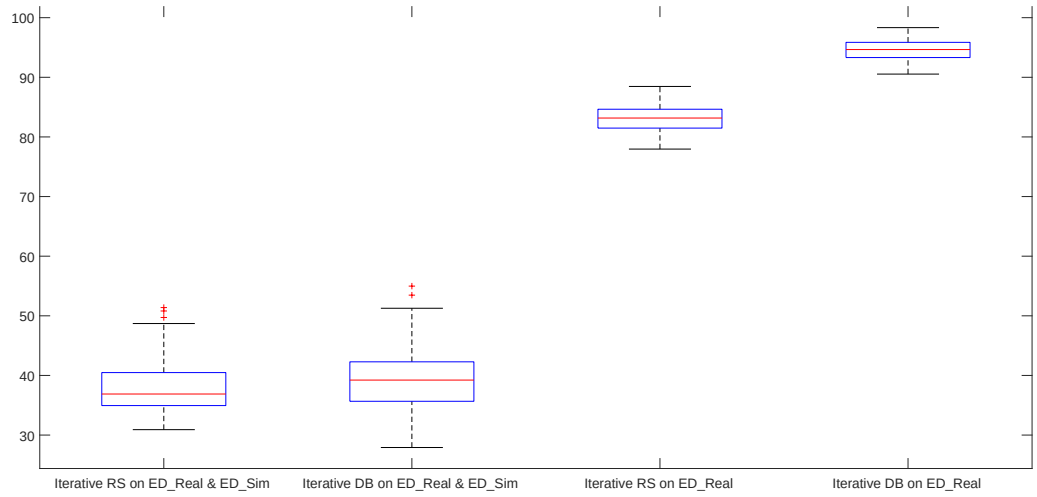
- Iterative RS on *ED_Real*: On *ED_Real* alone, the iterative resampling procedure was applied with the predictors being **w** excluding the qualitative variable for real or simulation data.

- Iterative DB on *ED_Real*: On *ED_Real* alone, the iterative difference-based procedure was applied with the predictors being **w** excluding the qualitative variable for real or simulation data.

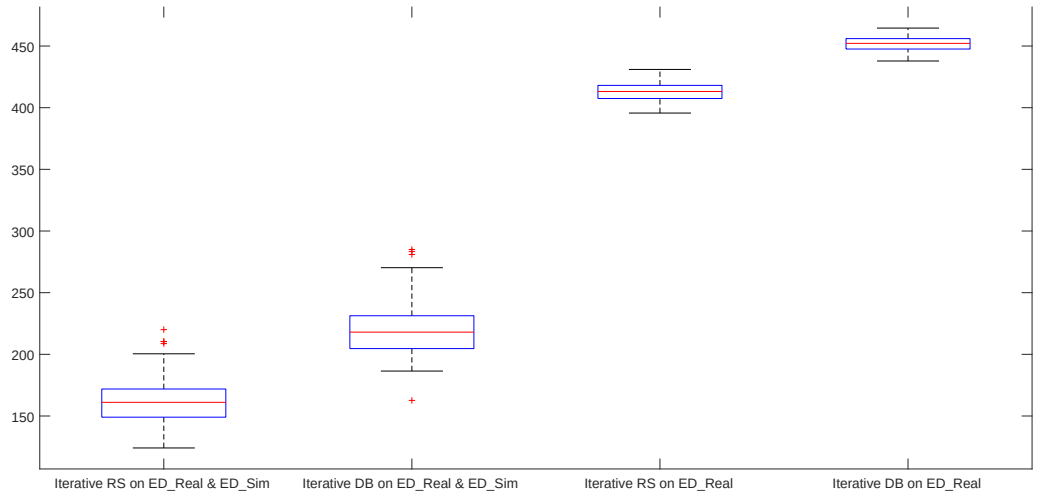
To compare the variance estimation results in a statistical manner, 100 macro-replications were performed. These macro-replications follow the schemes as briefed in 2.3.1. Simulation runs were carried for data collection using a different random stream. Each of the four scenarios above leads to 100 fitted models and 100 MAPEs and ERMSEs. In this chapter, the same equations (2.37 and 2.38) are used to calculate MAPE and ERMSE to measure the deviations of the estimated standard deviations from their true values.

Figure 3.1(a) and (b) display the MAPE and ERMSE box plots for the four scenarios. Each box is plotted from the 100 MAPEs or ERMSEs for the corresponding scenario. As can be seen from the box plots, including simulation data in the modeling of real system's variance modeling improves the quality of the estimation results substantially. The medians of the MAPEs and ERMSEs obtained from these scenarios are given in Table 3.1

Figure 3.2 plots the estimated standard deviations resulting from "Iterative RS on *ED_Real* and *ED_Sim*" scenario and their "true" values against the index of the check points (horizontal axis). The check points are roughly sorted based on their WIP levels of the validation set. From the "true" plots, it can be seen that the standard deviations vary widely over the check-point region, from about 200 to 1100. The standard deviation estimates are not able to capture the drastic changes throughout



(a) MAPE box plots.



(b) ERMSE box plots.

Figure 3.1: Comparison of variance estimation results using *ED* over different scenarios.

Table 3.1: Medians of MAPEs and ERMSEs from macro-replications for variance models using *ED*.

Method	MAPE	ERMSE
Iterative RS on *ED_Real* and *ED_Sim*	36.88%	161.1
Iterative DB on *ED_Real* and *ED_Sim*	39.22%	218.0
Iterative RS on *ED_Real*	83.17%	413.1
Iterative DB on *ED_Real*	94.65%	452.1

the check settings, and only show a slightly increasing trend along the point index (over the WIP range).

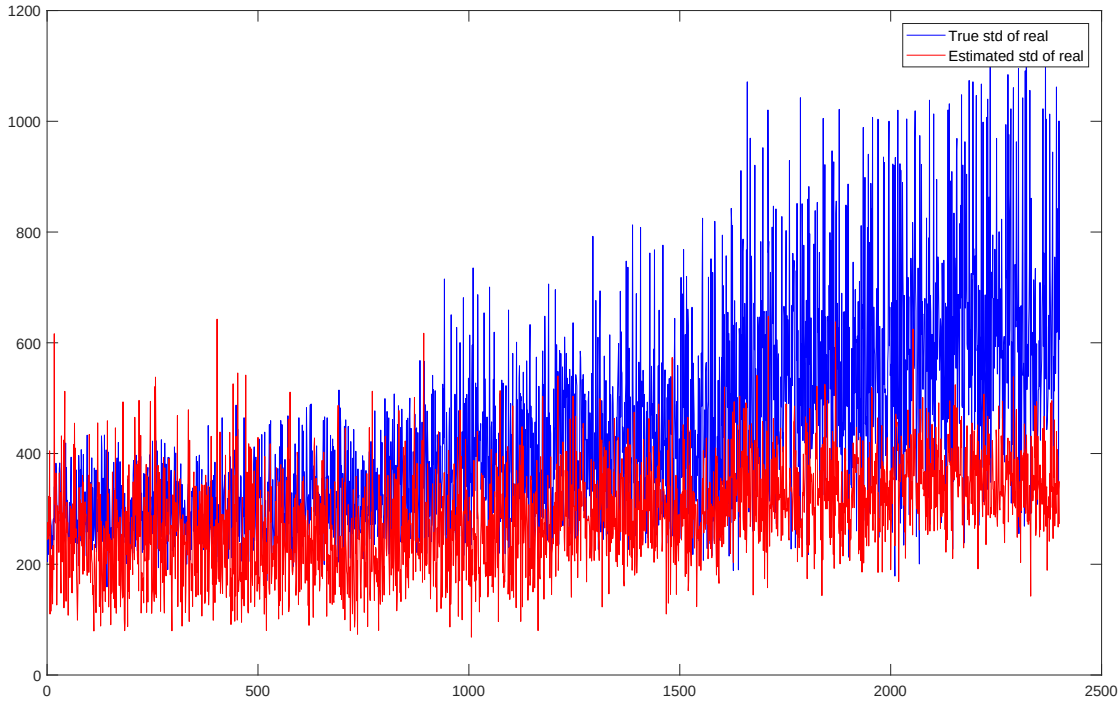


Figure 3.2: True and estimated standard deviations using *ED* over WIP

Considering the limitation of the model for the WIP levels, we generate a new Estimation Data set for both simulation and real system. In the new estimation data set *NewED_Sim* and *NewED_Real*, the same scheme was employed to generate design points with pre-specified WIP range being [15,42], instead of [15 60].

Boxplots (Figure 3.4) also show that the predictions are more accurate within new data sets, however the standard deviation is still underestimated. Table 3.2 shows the numerical results of the evaluation of different methods using fitting by *NewED_Sim* and *NewED_Real* and estimated with

the new validation data set.

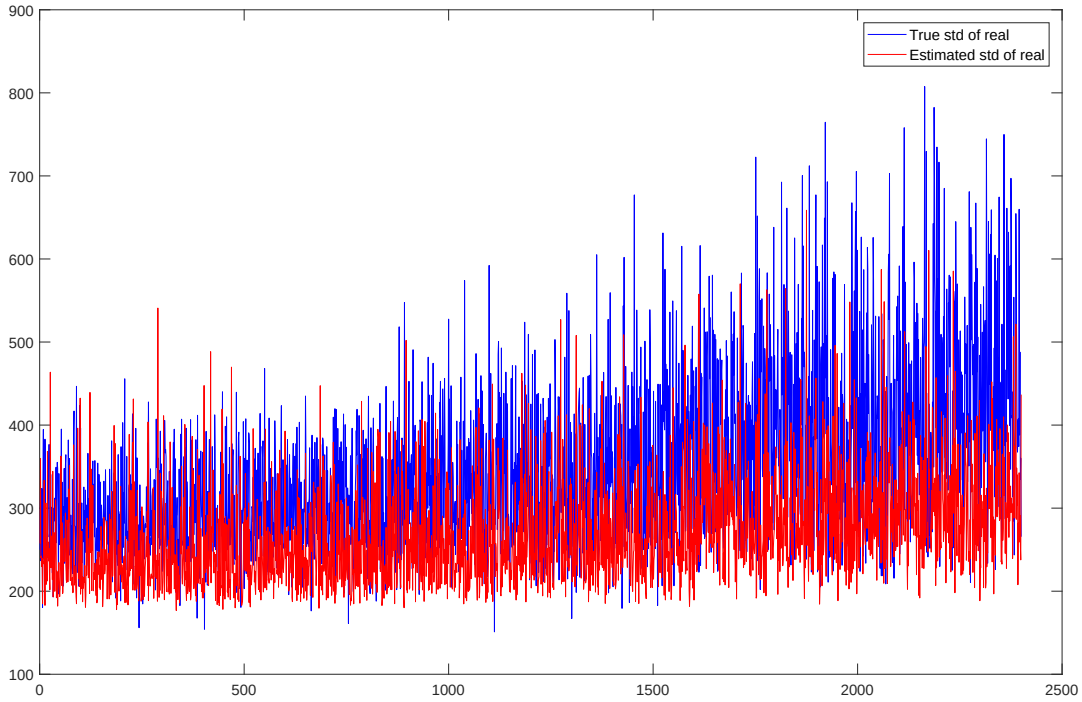
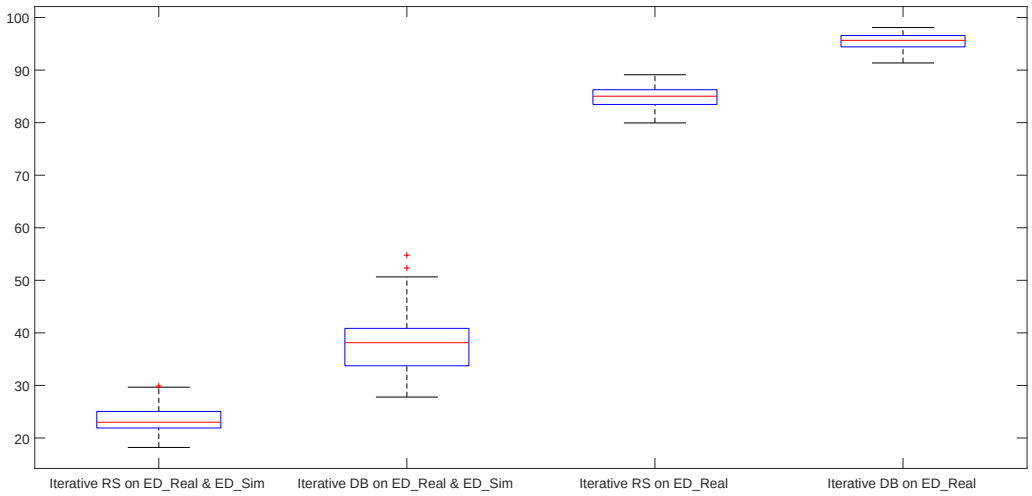


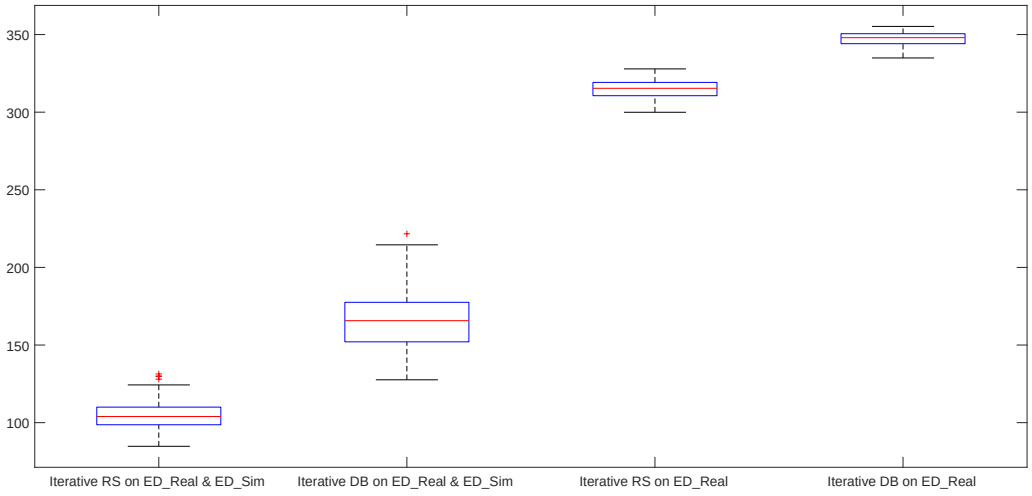
Figure 3.3: True and estimated standard deviations using *NewED* over WIP

Table 3.2: Medians of MAPEs and ERMSEs from macro-replications for variance models using *NewED* .

Method	MAPE	ERMSE
Iterative RS on *NewED_Real* and *NewED_Sim*	23.01%	103.9
Iterative DB on *NewED_Real* and *NewED_Sim*	38.13%	165.7
Iterative RS on *NewED_Real*	85.03%	315.4
Iterative DB on *NewED_Real*	95.66%	347.9



(a) MAPE box plots.



(b) ERMSE box plots.

Figure 3.4: Comparison of variance estimation results using *NewED* over different scenarios

Chapter 4

Quoting Lead Time

In the previous chapters, statistical procedures were developed to quantify the flow time characteristics (mean and variance) as a function of the predictor variables $\mathbf{w}$. To quote lead time with desired service level, percentile estimates of flow time are needed. Thus, in this chapter, a distribution is fitted based on the mean and variance models, and will be used to provide percentile estimates for lead time quotation.

In this work, both normal and gamma are considered as potential distribution families for flow time [22, 14]. Normal distribution is widely used to model continuous quantities with mean and standard deviation characteristics. Gamma is a highly flexible distribution suitable to model flow time in manufacturing [14], and has been adopted in a range of lead-time modeling work [56, 57, 58, 59, 6, 1].

4.1 Flow Time Distribution

Normal and Gamma distributions are used to model flow time. Gamma distribution's probability density function is:

$$g(y;\alpha(\mathbf{w}),\beta(\mathbf{w})) = \frac{1}{\Gamma(\alpha(\mathbf{w}))\beta(\mathbf{w})^{\alpha(\mathbf{w})}} y^{\alpha(\mathbf{w})-1} \exp(-\frac{y}{\beta(\mathbf{w})}) y > 0, \tag{4.1}$$

where α is the shape parameter, β the scale parameter. As shown in (4.1), the distribution parameters are dependent on the predictor variables $\mathbf{w}$. The relationship among the predictor variables $\mathbf{w}$

and the flow time distribution is quantified in two steps.

- First, the developed methods in the previous chapters are used to predict mean and variance of the flow time for a new item upon its arrival (at setting $\mathbf{w}$).
- Second, the predicted mean and variance are employed to estimate the gamma distribution parameters as follows [1]:

$$\alpha(\mathbf{w}) = \frac{[\widehat{E}[\mathsf{Y}(\mathbf{w})]]^2}{\widehat{Var}[\mathsf{Y}(\mathbf{w})]} \tag{4.2}$$

$$\beta(\mathbf{w}) = \frac{\widehat{Var}[\mathsf{Y}(\mathbf{w})]}{\widehat{E}[\mathsf{Y}(\mathbf{w})]}, \tag{4.3}$$

where $\widehat{E}[\mathsf{Y}(\mathbf{w})]$ and $\widehat{Var}[\mathsf{Y}(\mathbf{w})]$ denote the estimated mean and variance of the flow time, respectively. For the normal distribution the estimated mean and variance of the flow time , $\widehat{E}[\mathsf{Y}(\mathbf{w})]$ and $\widehat{Var}[\mathsf{Y}(\mathbf{w})]$, can be used to specify the distribution directly.

To quote the lead time for an order upon its arrival with a desired service level, say 95%, the status variables of the job shop is observed and fed to the mean and variance models of flow time. With the mean and variance esitmates, the gamma distribution parameters are calculated by (4.2) and (4.3). The fitted distribution renders percentile estimates of flow time, which serves as the lead time quoted for a desired service level.

4.2 Empirical Results for Lead Time Quotation

In this section, we estimate the distributions using the New Estimation Data Set (*NewED*) and New Validation Data Set (*NewVD*) (as in 3.3.1). The target service level is set as 95%. The lead time is quoted as the 95^{th} percentile estimates of the flow-time distribution.

4.2.1 Model Evaluation Criteria

At each check point, a realization of the new job's flow time is denoted as $\mathscr{Y}_j^*$ ($j = 1,2,\ldots,2400$), and the quoted lead time l_j ($j = 1,2,\ldots,2400$). The lead time quoted by the approach here is evaluated based on the following metrics.

- Achieved service level, which is calculated as

$$\frac{1}{2400} * \sum_{j=1}^{2400} I(\mathscr{Y}_j^* \leq l_j), \tag{4.4}$$

where I is the indicator function.

- Mean absolute percent error, which is calculated as

$$\frac{1}{2400} * \sum_{j=1}^{2400} \frac{|\mathscr{Y}_j^* - l_j|}{\mathscr{Y}_j^*}. \tag{4.5}$$

- Mean earliness, which is calculated as

$$\frac{1}{2400} * \sum_{j=1}^{2400} Max(0, l_j - \mathscr{Y}_j^*). \tag{4.6}$$

- Mean tardiness, which is calculated as

$$\frac{1}{2400} * \sum_{j=1}^{2400} Max(0, \mathscr{Y}_j^* - l_j). \tag{4.7}$$

- Mean missed due date, which is calculated as

$$\frac{1}{2400} * \sum_{j=1}^{2400} |\mathscr{Y}_j^* - l_j|. \tag{4.8}$$

- Mean of the lead time quoted, which is calculated as

$$\frac{1}{2400} * \sum_{j=1}^{2400} l_j. \tag{4.9}$$

4.2.2 Evaluation of Lead-Time Quotation

The two approaches below were applied and used to quote lead times.

Iterative DB on *NewED_Real* and *NewED_Sim*: The iterative difference-based procedure for variance model (Section 3.2.2) was applied to model the data ensemble of *NewED_Real* and

NewED_Sim with the predictors being **w** including the qualitative variable for real or simulation data.

Iterative RS on *NewED_Real* and *NewED_Sim*: The iterative resampling procedure for variance model (Section 3.2.1) was applied to model the data ensemble of *NewED_Real* and *NewED_Sim* with the predictors being **w** including the qualitative variable for real or simulation data.

The new estimation data sets (*NewED_Real* and *NewED_Sim* with WIP range of 15 to 42) and the new validation data set were to used in the evaluation.

Table 4.1: Evaluation of quoted lead times in terms of the performance metrics.

	Gamma Distribution		Normal Distribution	
	Iterative DB on *NewED_Real* and *NewED_Sim*	Iterative RS on *NewED_Real* and *NewED_Sim*	Iterative DB on *NewED_Real* and *NewED_Sim*	Iterative RS on *NewED_Real* and *NewED_Sim*
Achieved service level	0.93076	0.9614	0.92939	0.95958
Mean absolute percent error	0.21242	0.24128	0.2105	0.23842
Mean earliness(Minutes)	976.53	1115	966.58	1101.2
Mean tardiness(Minutes)	16.209	8.1413	16.524	8.5289
Mean missed due date(Minutes)	993.81	1124.1	986.54	1110.8
Mean of the lead time quoted(Minutes)	5619.6	5770.2	5610.2	5755.9

Table 4.1 compares the two approaches in terms of the six performance metrics with the target service level being 95%. Compared to the difference-based approach, the resampling-based approach tends to quote longer lead times (longer mean quoted lead time), resulting in higher service levels (slightly higher than the target), longer mean earliness, and shorter mean tardiness. The table does not show significant differences between Normal and Gamma distributions.

Chapter 5

Summary

In this work, statistical procedures were adapted to assist lead time quotation of a new customer's order upon its arrival. The kriging-based modeling procedure integrates well-designed simulation data and observed real data to achieve models of improved quality for the estimation of both mean and variance of flow time. Built on the mean and variance models, the flow time distribution is fitted rendering percentile estimates which can be used for lead time quotation with desired service levels.

Through simulation studies, it has been shown that well-designed simulation data, though deviates somewhat from the real-world system, help to substantially improve the modeling of the real system's behavior. The iterative procedures in this work are able to achieve accurate mean flow time models when synergistically modeling replicated simulation and non-replicated real data. However, the variance models generated by the procedures fall short in capturing the drastic changes in flow time variance for the real system, even when the non-replicated real data are supplemented with replicated simulation data.

The design of simulation experiments carried in this work was developed in Li et al. [1] for linear regression of simulation data alone. An immediate next step is to develop experimental design methods for simulation experiments particularly tailored to the modeling of both simulation and real data.

Bibliography

[1] Li, M., Yang, F., Wan, H., and Fowler, J. W., "Simulation-based Experimental Design and Statistical Modeling for Lead Time Quotation," *Journal of Manufacturing Systems*, Vol. 37, 2015, pp. 362–374.

[2] Adams, J., Balas, E., and Zawack, D., "The Shifting Bottleneck Procedure for Job Shop Scheduling," *Management Science*, Vol. 34, No. 3, 1988, pp. 391–401.

[3] Ioannou, G. and Dimitriou, S., "Lead Time Estimation in MRP/ERP for Make-to-order Manufacturing Systems," *International Journal of Production Economics*, Vol. 139, No. 2, 2012, pp. 551–563.

[4] Corti, D., Pozzetti, A., and Zorzini, M., "A Capacity-driven Approach to Establish Reliable Due Dates in a MTO Environment," *International Journal of Production Economics*, Vol. 104, No. 2, 2006, pp. 536–554.

[5] Wein, L. M., "Due-date Setting and Priority Sequencing in a Multiclass M/G/1 Queue," *Management Science*, Vol. 37, No. 7, 1991, pp. 834–850.

[6] Duenyas, I. and Hopp, W. J., "Quoting customer lead times," *Management Science*, Vol. 41, No. 1, 1995, pp. 43–57.

[7] Spearman, M. L. and Zhang, R. Q., "Optimal Lead Time Policies," *Management Science*, Vol. 45, No. 2, 1999, pp. 290–295.

[8] Savaşaneril, S., Griffin, P. M., and Keskinocak, P., "Dynamic Lead-time Quotation for an M/M/1 Base-stock Inventory Queue," *Operations Research*, Vol. 58, No. 2, 2010, pp. 383–395.

[9] Altendorfer, K. and Jodlbauer, H., "An Analytical Model for Service Level and Tardiness in a Single Machine MTO Production System," *International Journal of Production Research*, Vol. 49, No. 7, 2011, pp. 1827–1850.

[10] Lawrence, S. R., "Estimating Flowtimes and Setting Due-dates in Complex Production Systems," *IIE Transactions*, Vol. 27, No. 5, 1995, pp. 657–668.

[11] Öztürk, A., Kayalıgil, S., and Özdemirel, N. E., "Manufacturing Lead Time Estimation Using Data Mining," *European Journal of Operational Research*, Vol. 173, No. 2, 2006, pp. 683–700.

[12] Pearn, W., Chung, S., and Lai, C., "Due-date Assignment for Wafer Fabrication under Demand Variate Environment," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 20, No. 2, 2007, pp. 165–175.

[13] Baykasoğlu, A., Göçken, M., and Unutmaz, Z. D., "New Approaches to Due Date Assignment in Job Shops," *European Journal of Operational Research*, Vol. 187, No. 1, 2008, pp. 31–45.

[14] Hopp, W. J. and Sturgis, M. L. R., "Quoting Manufacturing Due Dates Subject to a Service Level Constraint," *IIE Transactions*, Vol. 32, No. 9, 2000, pp. 771–784.

[15] Vig, M. M. and Dooley, K. J., "Dynamic Rules for Due-date Assignment," *The International Journal of Production Research*, Vol. 29, No. 7, 1991, pp. 1361–1377.

[16] Hsu, S. and Sha, D., "Due Date Assignment Using Artificial Neural Networks Under Different Shop Floor Control Strategies," *International Journal of Production Research*, Vol. 42, No. 9, 2004, pp. 1727–1745.

[17] Sabuncuoglu, I. and Comlekci, A., "Operation-based Flowtime Estimation in a Dynamic Job Shop," *Omega*, Vol. 30, No. 6, 2002, pp. 423–442.

[18] Sha, D., Storch, R., and Liu, C.-H., "Development of a Regression-based Method with Case-based Tuning to Solve the Due Date Assignment Problem," *International Journal of Production Research*, Vol. 45, No. 1, 2007, pp. 65–82.

[19] Philipoom, P. R., Rees, L. P., and Wiegmann, L., "Using Neural Networks to Determine Internally-Set Due-Date Assignments for Shop Scheduling," *Decision Sciences*, Vol. 25, No. 5-6, 1994, pp. 825–851.

[20] Li, S., Li, Y., Liu, Y., and Xu, Y., "A GA-based NN Approach for Makespan Estimation," *Applied Mathematics and Computation*, Vol. 185, No. 2, 2007, pp. 1003–1014.

[21] Wang, K., Chen, X., Yang, F., Porter, D. W., and Wu, N., "A New Stochastic Kriging Method for Modeling Multi-Source Exposure–Response Data in Toxicology Studies," *ACS Sustainable Chemistry & Engineering*, Vol. 2, No. 7, 2014, pp. 1581–1591.

[22] Ankenman, B., Nelson, B. L., and Staum, J., "Stochastic Kriging for Simulation Metamodeling," *Operations Research*, Vol. 58, No. 2, 2010, pp. 371–382.

[23] Chen, X., Ankenman, B. E., and Nelson, B. L., "Enhancing Stochastic Kriging Metamodels with Gradient Estimators," *Operations Research*, Vol. 61, No. 2, 2013, pp. 512–528.

[24] Chen, X., Ankenman, B. E., and Nelson, B. L., "The Effects of Common Random Numbers on Stochastic Kriging Metamodels," *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, Vol. 22, No. 2, 2012, pp. 7.

[25] Seber, G. and Wild, C., "Computational Methods for Nonlinear Least Squares," *Nonlinear Regression*, 2005, pp. 619–660.

[26] Riano, G., *Transient Behavior of Stochastic Networks: Application to Production Planning with Load-dependent Lead Times*, Ph.D. thesis, School of Industrial and Systems Engineering, Georgia Institute of Technology, 2002.

[27] Qian, P. Z. G., Wu, H., and Wu, C. J., "Gaussian Process Models for Computer Experiments with Qualitative and Quantitative Factors," *Technometrics*, Vol. 50, No. 3, 2008, pp. 383–396.

[28] Xiong, Y., Chen, W., and Tsui, K.-L., "A New Variable-fidelity Optimization Framework Based on Model Fusion and Objective-oriented Sequential Sampling," *Journal of Mechanical Design*, Vol. 130, No. 11, 2008, pp. 111401.

[29] Mühlenstädt, T., Gösling, M., and Kuhnt, S., "How to Choose the SImulation Model for Computer Experiments: a Local Approach," *Applied Stochastic Models in Business and Industry*, Vol. 28, No. 4, 2012, pp. 354–361.

[30] Gratiet, L. L. and Cannamela, C., "Kriging-based Sequential Design Strategies Using Fast Cross-validation Techniques with Extensions to Multi-fidelity Computer Codes," *arXiv preprint arXiv:1210.6187*, 2012.

[31] Chen, R.-B., Hung, Y.-C., Wang, W., and Yen, S.-W., "Contour Estimation via Two Fidelity Computer Simulators Under Limited Resources," *Computational Statistics*, Vol. 28, No. 4, 2013, pp. 1813–1834.

[32] Goh, J., Bingham, D., Holloway, J. P., Grosskopf, M. J., Kuranz, C. C., and Rutter, E., "Prediction and Computer Model Calibration Using Outputs from Multifidelity Simulators," *Technometrics*, Vol. 55, No. 4, 2013, pp. 501–512.

[33] Santner, T. J., Williams, B. J., and Notz, W., *The Design and Analysis of Computer Experiments*, Springer Science & Business Media. New York., 2003.

[34] Santner, T., Williams, B., and Notz, W., "The Design and Analysis of Computer Experiments Springer-Verlag," *New York. 283pp*, 2003.

[35] Hopp, W. J., Spearman, M. L., Chayet, S., Donohue, K. L., and Gel, E. S., "Using an Optimized Queueing Network Model to Support Wafer Fab Design," *Iie Transactions*, Vol. 34, No. 2, 2002, pp. 119–130.

[36] Carroll, R. J., "Adapting for heteroscedasticity in linear models," *The Annals of Statistics*, 1982, pp. 1224–1233.

[37] Gasser, T., Sroka, L., and Jennen-Steinmetz, C., "Residual variance and residual pattern in nonlinear regression," *Biometrika*, Vol. 73, No. 3, 1986, pp. 625–633.

[38] Muller, H.-G. and Stadtmuller, U., "Variable bandwidth kernel estimators of regression curves," *The Annals of Statistics*, 1987, pp. 182–201.

[39] Müller, H.-G. and Stadtmüller, U., "On variance function estimation with quadratic forms," *Journal of Statistical Planning and Inference*, Vol. 35, No. 2, 1993, pp. 213–231.

[40] Hall, P. and Carroll, R., "Variance function estimation in regression: the effect of estimating the mean," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1989, pp. 3–14.

[41] Herrmann, E., "Local bandwidth choice in kernel regression estimation," *Journal of Computational and Graphical Statistics*, Vol. 6, No. 1, 1997, pp. 35–54.

[42] Fan, J. and Yao, Q., "Efficient estimation of conditional variance functions in stochastic regression," *Biometrika*, Vol. 85, No. 3, 1998, pp. 645–660.

[43] Opsomer, J. D., Ruppert, D., Wand, M. P., Holst, U., and Hössjer, O., "Kriging with nonparametric variance function estimation," *Biometrics*, Vol. 55, No. 3, 1999, pp. 704–710.

[44] Brown, L. D., Levine, M., et al., "Variance estimation in nonparametric regression via the difference sequence method," *The Annals of Statistics*, Vol. 35, No. 5, 2007, pp. 2219–2232.

[45] Wang, L., Brown, L. D., Cai, T. T., and Levine, M., "Effect of mean on variance function estimation in nonparametric regression," *The Annals of Statistics*, 2008, pp. 646–664.

[46] Cai, T. T., Wang, L., et al., "Adaptive variance function estimation in heteroscedastic nonparametric regression," *The Annals of Statistics*, Vol. 36, No. 5, 2008, pp. 2025–2054.

[47] Robinson, T. J., Birch, J. B., and Starnes, B. A., "A Semi-parametric Approach to Dual Modeling When no Replication Exists," *Journal of Statistical Planning and Inference*, Vol. 140, No. 10, 2010, pp. 2860–2869.

[48] Marrel, A., Iooss, B., Da Veiga, S., and Ribatet, M., "Global sensitivity analysis of stochastic computer models with joint metamodels," *Statistics and Computing*, Vol. 22, No. 3, 2012, pp. 833–847.

[49] Navaee, M., Mobin, M., Vardani, M., and Ahmadi, N., "A Semi parametric approach to dual modeling," *Management Science Letters*, Vol. 2, No. 2, 2002, pp. 665–672.

[50] Goldberg, P. W., Williams, C. K., and Bishop, C. M., "Regression with input-dependent noise: A Gaussian process treatment," *Advances in neural information processing systems*, 1998, pp. 493–499.

[51] Le, Q. V., Smola, A. J., and Canu, S., "Heteroscedastic Gaussian process regression," *Proceedings of the 22nd international conference on Machine learning*, ACM, 2005, pp. 489–496.

[52] Kersting, K., Plagemann, C., Pfaff, P., and Burgard, W., "Most likely heteroscedastic Gaussian process regression," *Proceedings of the 24th international conference on Machine learning*, ACM, 2007, pp. 393–400.

[53] Boukouvalas, A. and Cornford, D., "Learning heteroscedastic Gaussian processes for complex datasets," *Technical report*, 2009.

[54] Titsias, M. K. and Lázaro-Gredilla, M., "Variational heteroscedastic Gaussian process regression," *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 841–848.

[55] Ruppert, D., Wand, M. P., Holst, U., and HöSJER, O., "Local polynomial variance-function estimation," *Technometrics*, Vol. 39, No. 3, 1997, pp. 262–273.

[56] Karmarkar, U. S., "Lot sizes, lead times and in-process inventories," *Management science*, Vol. 33, No. 3, 1987, pp. 409–418.

[57] Yano, C. A., "Setting planned leadtimes in serial production systems with tardiness costs," *Management science*, Vol. 33, No. 1, 1987, pp. 95–106.

[58] Hendry, L. C. and Kingsman, B., "Production planning systems and their applicability to make-to-order companies," *European journal of operational research*, Vol. 40, No. 1, 1989, pp. 1–15.

[59] Hill, A. V. and Khosla, I. S., "Models for optimal lead time reduction," *Production and Operations Management*, Vol. 1, No. 2, 1992, pp. 185–197.

[60] Hopp, W. J. and Spearman, M. L., "Factory Physics: Foundations of Manufacturing Management, Richard D," *Irwin, Chicago, IL*, 1996.

[61] Backus, P., Janakiram, M., Mowzoon, S., Runger, C., and Bhargava, A., "Factory Cycle-time Prediction with a Data-mining Approach," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 19, No. 2, 2006, pp. 252–258.

[62] Fang, K.-T. and Yang, Z.-H., "On Uniform Design of Experiments with Restricted Mixtures and Generation of Uniform Distribution on Some Domains," *Statistics & Probability Letters*, Vol. 46, No. 2, 2000, pp. 113–120.

[63] Wu, C. J. and Hamada, M. S., *Experiments: Planning, Analysis, and Optimization*, Vol. 552, John Wiley & Sons, 2011.

[64] Banks, J., Carson, J., Nelson, B., and Nicol, D., "Discrete-Event System Simulation 5th ed. Upper Saddle River, NeW Jersey: Prentice Hall," 2010.

Appendix

5.1 Comparison with the most related literature

The following table shows the contribution of this research and compares it to most related work.

Table 5.1: Most related literature

Source	Method	Data Source	Data Type
Li et al. [1]	Regression methods with some Pre-assumed functional forms	Simulation data of a scaled-down semi-conductor manufacturing system	Replicated
Ankenman et al. [22]	Stochastic kriging with only quantitative factors	Simulation data of a scaled-down semi-conductor manufacturing system	Replicated
Qian et al. [27]	Kriging with both quantitative and qual-itative factors	Real Data for modeling the thermal dis-tribution of a data center	The model is not stochastic
Wang et al. [21]	Stochastic kriging with both quantitative and qualitative factors	Synthetic data based on known func-tions	Replicated
Marrel et al. [48]	Gaussian process regression for model-ing both mean and variance	Variable settings are controllable for sampling data from a single source	Non-replicated
This research	Stochastic kriging with both quantitative and qualitative factors	Simulation and Real data	Both non-replicated and replicated

5.2 Configuration of the Example System

In the example system investigated in this paper, customer orders arrive to the system is a homo-geneous compound Poisson process with a rate ranging within $[0.129, 0.150]$ per hour. Order size (i.e., the number of jobs requested by customer) distribution is discrete uniform with possible values being 1, 2, and 3. There are 22 processing steps for each job through 10 workstations. Figure 5.1 shows the sequence of required processing steps and the stations that a job has to visit.

As shown in Figure 5.1, Stations 1, 4, 5, 6 and 7 are revisited by jobs. Table 5.2 provides for each station the number of machines available, batch processing size, mean and standard deviation of the processing time (Mean PT and Stdev PT), and whether or not the machines are subject to random failures. Based on Figure 5.1 and Table 5.2, this manufacturing system involves major fea-

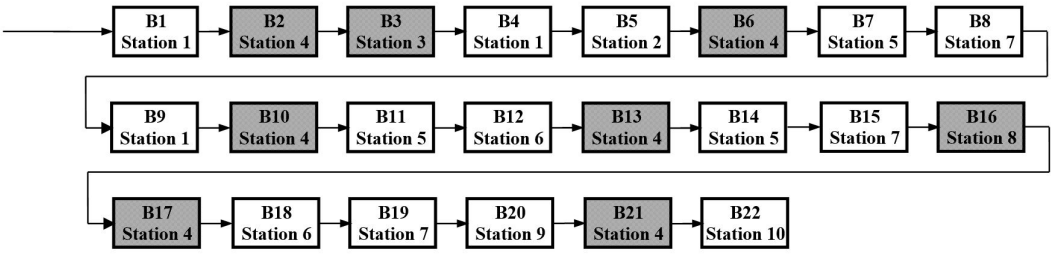


Figure 5.1: Job processing sequence and important workstations.

tures present in real semiconductor fabrication system: re-entrant flows (revisited stations), machine failures, and batch processing. The processing time at each machine follows a log-normal distribution. Machines at Stations 3 and 7 are subject to random failures. At Station 3, time to failure (TTF) follows a gamma distribution with parameters $(\alpha, \beta) = (3600, 1)$, and time to repair (TTR) has $(\alpha, \beta) = (600, 1.5)$ for a gamma distribution. At Station 7, both TTF and TTR follow gamma distribution with the distribution parameters $(\alpha, \beta) = (720, 1)$ and $(\alpha, \beta) = (120, 1.5)$, for TTF and TTR respectively. Besides, Stations 1 and 2 involve batch processing. The maximum and minimum batch sizes allowed for these two stations are 4 and 2.

Table 5.2: Configuration of the Example System.

Station index #	1	2	3	4	5	6	7	8	9	10
# of machines	1	1	1	2	1	1	1	1	1	1
Batch size (min/max)	2/4	2/4	1	1	1	1	1	1	1	1
Failure	No	No	Yes	No	No	No	Yes	No	No	No
Mean PT (min)	78	255	145	63	40	45	37	185	80	82
Stdev PT(min)	8	16	7	4	3	4	4	12	6	7

5.3 Configuration of the System for "Real" Data

The "real" system follows the same configuration as the one in Appendix 5.2, and its parameters are provided in Table 5.3.

Table 5.3: Configuration of each workstation for collecting real data (time units: min).

Station index #	1	2	3	4	5	6	7	8	9	10
# of machines	1	1	1	2	1	1	1	1	1	1
Batch size (min/max)	2/4	2/4	1	1	1	1	1	1	1	1
Failure	No	No	Yes	No	No	No	Yes	No	No	No
Mean PT (min)	70	235	135	56	30	25	35	160	54	54
Stdev PT(min)	7	16	7	4	2	2.4	4	12	4	5

5.4 Preliminary Analysis

In this section we review the preliminary analysis based on [1, 35].Queueing theory and empirical experience assert that a manufacturing systems performance is mainly affected by a small number of important workstations (IWs) which most limit the job flows [60]. The preliminary analytical analysis is used to obtain a smaller set of predictor variables, **w**, that have a more significant impact on flow time and are more possibly to be a major predictor. Usually, stations with high utilization rate are considered as IWs and we can use analytical methods to find high utilized stations. Assuming the arrival rate of jobs is λ and the utilization of each station is $\varphi_j(\lambda); i = 1,2,\ldots$, let $\varphi_{max}(\lambda) = Max\varphi_j(\lambda)$, and $\varphi_j(\lambda)/\varphi_{max}(\lambda)$ turns out to be a constant ratio. According to the rule of thumb adopted by [1] if $\varphi_j(\lambda)/\varphi_{max}(\lambda) > 0.8$, then station j is considered as highly utilized. Applying the analytical method by [35] on the example system shows that highly utilized stations are 3, 4, and 8 and considered as important workstations (IWs). For all the jobs that are in the same step of their production sequence, a virtual location called buffer is considered [26]. The buffers involve with IWs are referred as important buffers (IB). IBs include: $B2, B3, B6, B10, B13, B16, B17, andB21$ which are represented in 5.1 as shaded boxes. A server associated with IWs is considered as an important server (IS). The predictor variables **w** can be determined based on identified IBs and ISs. As mentioned before X_{ORIG} Includes: a. The shop status variables SVs; b. The size of a newly arrived order; and c. The arrival rate of future orders and predictor variables **w** can be derived from X_{ORIG}. For illustration, we divide variables into work in process (WIP) variables and non-WIP variables. First consider the WIP variables. In the example system, instead of the number of jobs at each buffer (SVs.A in X_{ORIG}), a set of stage WIP is included in **w**. Each stage involves the steps bounded by the IBs in a job sequence. In the example system, there are eight stages bounded by $IB = [B2, B3, B6, B10, B13, B16, B17, B21]$. The stage WIP variables are denoted as $x_{WIP} = \{x_1^{WIP}, x_2^{WIP}, , x_S^{WIP}\}$, where S is the number of stages and x_S^{WIP} the number of jobs in the sth stage. Based on the stage WIP definition, the 1st stage WIP, x_1^{WIP}, counts the number of jobs yet to be processed by the station at the first IB $B2$; the sth $(s = 2,3,\ldots,S)$ stage WIP x_S^{WIP} counts the number of jobs between the (s-1)th and sth IB, including those being processed at the (s-1)th IB and excluding those being processed at the sth IB. The number of jobs that have been processed by the station at the last IB are not considered as part of the stage WIP variables. The variables x_{WIP}

are important because they indicated the shop congestion, and are found to be the most important factors in flow time prediction [61].

5.5 Design of Simulation Experiments and Collecting Data

In this study due to the similarity of the research problems, we used the design of experiment method suggested by [1]. This method will be briefly discussed in this section. After designing the experiments, the simulation model will be used to collect the data. Duo to the large number of predictor variables, DOE for the flow time model is challenging. Besides predictor variables include different types of variables like qualitative variables, continuous quantitative, and discrete quantitative variables. Furthermore, a design in the space of $\mathbf{w}$ is not sufficient to specify the experimental condition of a simulation run which is represented as X_{ORIG}. Hence, a design in the space of $\mathbf{w}$ has to be converted to one in the space of X_{ORIG}. An overview of the DOE procedure suggested by [1] is presented in Figure 5.2. We apply this procedure sequentially in this study.

The first four steps of the procedure (Figure 5.2) involve the design experiments in the space of w. For WIP variables, [1] selected four levels of the total WIP, Q, and for each level generated design points for all $\mathbf{x}_{WIP}$ variables. They suggest that the lower and upper bounds, QL and QU, can be chosen based on the observed limits for real or simulation systems. [1] employed the uniform design algorithm for mixtures developed by [62] to find a vector of the proportions of the total WIP and generate candidate design points in the space of $\mathbf{x}_{WIP}$. In non-WIP, [1] applied a mixed-level fractional factorial design [63] with several (typically two or three) levels selected for each variable. The mixed-level fractional factorial design method is used to obtain a resolution-IV design for non-WIP variables where each variable has a lower level (LL), high level (HL), and possibly a medium level (ML). Table 5.4 specifies the levels for non-WIP variables.

In Step 3, [1] implemented the cross array method [63] to generate candidate design points in the joint space of $\mathbf{w} = (\mathbf{x}_{WIP}, \mathbf{x}_{non-WIP})$. The cross array method provides a candidate pool for D-optimal design in step 4. At each level of WIP, D-optimal method finds $K/4$ points that $Max_{\mathbf{w}} = |\mathbf{w}'\mathbf{w}|$. Consequently, with four total WIP levels, K design points in the space of $\mathbf{w}$ are generated by the procedure. For simulation data, multiple replications are required at each design point to quantify the relationship between flow time and independent variables. In step 5, [1] applied

Table 5.4: Levels of non-WIP variables.

Variable Notation	Variable Levels
$\mathbf{z}_{B_i}$	Two levels with LL and HL corresponding to the idle and busy status of server i.
$\mathbf{x}_{C_i}$	Three levels with LL being 0, HL the 95th percentile of the distribution for server i processing time, and ML the average of LL and HL.
$\mathbf{z}_{D_i}$	Two levels with LL and HL corresponding to the down and up status of server i.
$\mathbf{x}_{E_i}$	Three levels with LL being 0, HL the 95th percentile of the distribution for server i repair time, and ML the average of LL and HL.
$\mathbf{x}_{F_i}$	Three levels with LL being 0, HL the 95th percentile of the distribution for server i time between failure, and ML the average of LL and HL.
$\mathbf{x}_{G_i}$	Multiple levels with each one corresponding to a batch size allowed by server i.
$\mathbf{x}_O$	Multiple (typically two) levels can be selected based on the distribution of the order size $\mathbf{x}_O$. $\mathbf{x}_O$ follows a discrete uniform distribution over the range $[\mathbf{x}_O^L, \mathbf{x}_O^U] = [1,3]$, and its LL, ML, and HL are set as minimum, average, and maximum of the distribution, respectively.
$\mathbf{x}_R$	In this work, the future orders arrival rate $\mathbf{x}_R$ takes three levels, which correspond to a steady-state system utilization of 75%, 80%, and 90%, respectively. With selected utilizations, the arrival rates can be determined using the queueing analytical analysis, and for the example system, they turn out to be 0.129, 0.136, and 0.150 orders per hour.

the precision of the mean estimate, $\widehat{E}[Y(\mathbf{w}_k)]$, to obtain number of replications at each point, n_k, using a two-stage framework [64]. Where:

$$\widehat{E}[Y(\mathbf{w}_k)] = \frac{1}{n_k(0)} \sum_{j=1}^{n_k(0)} \mathscr{Y}_j(\mathbf{w}_k) \tag{5.1}$$

Stage 1:

At the first stage, [1] performed a relatively small number of simulation runs, $n_k(0)$, to collect initial data. They suggest to choose the number of initial runs based on the recommendations of [64] and for the example problem they used $n_k(0) = 50$. However, design points in the $\mathbf{w}$ space only involve important non-WIP variables and stage-WIP variables. To run the simulation model, all the experimental condition (X_{ORIG}) is required. So, for each design point in the $\mathbf{w}$ space, [1] used the design conversion process (discussed later in this section) to generate $n_k(0)$ points in the space of X_{ORIG}. These $n_k(0)$ points in X_{ORIG} space corresponds to replications at the $\mathbf{w}$ space. Based on the initial data gathered by $n_k(0)$ replications, [1] calculated the mean estimate $\widehat{E}[Y(\mathbf{w}_k)]$ and estimated variance of $Y(w_k)$, at each design point as:

$$\widehat{Var}[Y(\mathbf{w}_k)] = \frac{1}{n_k(0)} \sum_{j=1}^{n_k(0)} (\mathscr{Y}_j(\mathbf{w}_k) - \widehat{E}[Y(\mathbf{w}_k)])^2 \tag{5.2}$$

The standard error of $\widehat{E}[Y(\mathbf{w}_k)]$ is estimated as:

$$SE\{\widehat{E}[Y(\mathbf{w}_k)]\} = \sqrt{\frac{\widehat{Var}[Y(\mathbf{w}_k)]}{n_k}} \tag{5.3}$$

Denoting $p\%$ as the user-specified precision level, the condition below is used to obtain n_k:

$$\frac{SE\{\widehat{E}[Y(\mathbf{w}_k)]\}}{\widehat{E}[Y(\mathbf{w}_k)]} = \frac{\sqrt{\frac{\widehat{Var}[Y(\mathbf{w}_k)]}{n_k}}}{\widehat{E}[Y(\mathbf{w}_k)]} \leq p\% \tag{5.4}$$

Which leads to estimate the number of replications at each point as:

$$n_k = \frac{\widehat{Var}[Y(\mathbf{w}_k)]}{(p\%\widehat{E}[Y(\mathbf{w}_k)])^2} \tag{5.5}$$

In this research, we set $p\% = 1.5\%$ as [1].

Stage 2:

After determining n_k for each design point, extra $n_k - n_k(0)$ runs are performed. Then the design conversion process it applied to obtain $n_k - n_k(0)$ points in the X_{ORIG} space. The final data set including K data points, obtained from design of experiment and simulation model is denoted as:

$$\{(\mathbf{w}_k, \mathscr{Y}_j(\mathbf{w}_k)); k = 1,2,\ldots,K, j = 1,2,,\ldots,n(\mathbf{w}_k)\} \tag{5.6}$$

Next, a design conversion process is used by [1] to generate n_k points in the space of X_{ORIG} based on a point in $\mathbf{w}$ space. [1] declared that different process can be used to generate points in X_{ORIG} because only the important variables in $\mathbf{w}$ are used to predict flow time. In their proposed method they mentioned two issues to be solved. First, the number of jobs at each buffer considering variables in $\mathbf{w}_k$. Particularly, for a stage s, the given X_{WIP} needs to be allocated to each buffer within that stage. [1] used the uniform design mixtures by [62] to find n_k distinct mixtures and used them to allocate X_{WIP} into buffers. The second issue is non-WIP variables associated with unimportant stations. Considering SVs.A are already specified, SVs.B in X_{ORIG} can be determined. The station is busy if the WIP is not zero. For the unimportant SVs.C variables, if the server is busy according to the SVs.A, then set the elapsed processing time as the servers mean processing time. For the unimportant SVs.DF variables, set the status of unreliable servers to be up and the elapsed up time

the servers mean time between failures. For the unimportant SVs.G variables, if the batch processing server is busy according to the SVs.A, then set its batch size as the minimum [1].

Finally, the simulation model for the system is coded in Microsoft Visual C^{++}. The simulation experiments are carried out following the DOE method mentioned before to develop the models and quantify the dependence of flow time distribution upon the predictor variables **w**. For each simulation run, X_{ORIG} is specified according to the DOE : a simulation run is initiated with a designed shop status and a newly arrived order of a designed size at time 0; as the simulation proceeds, orders are fed into the system at the designed arrival rate; the simulation run is terminated once the new order generated at time 0 is completed, and its flow time is recorded.

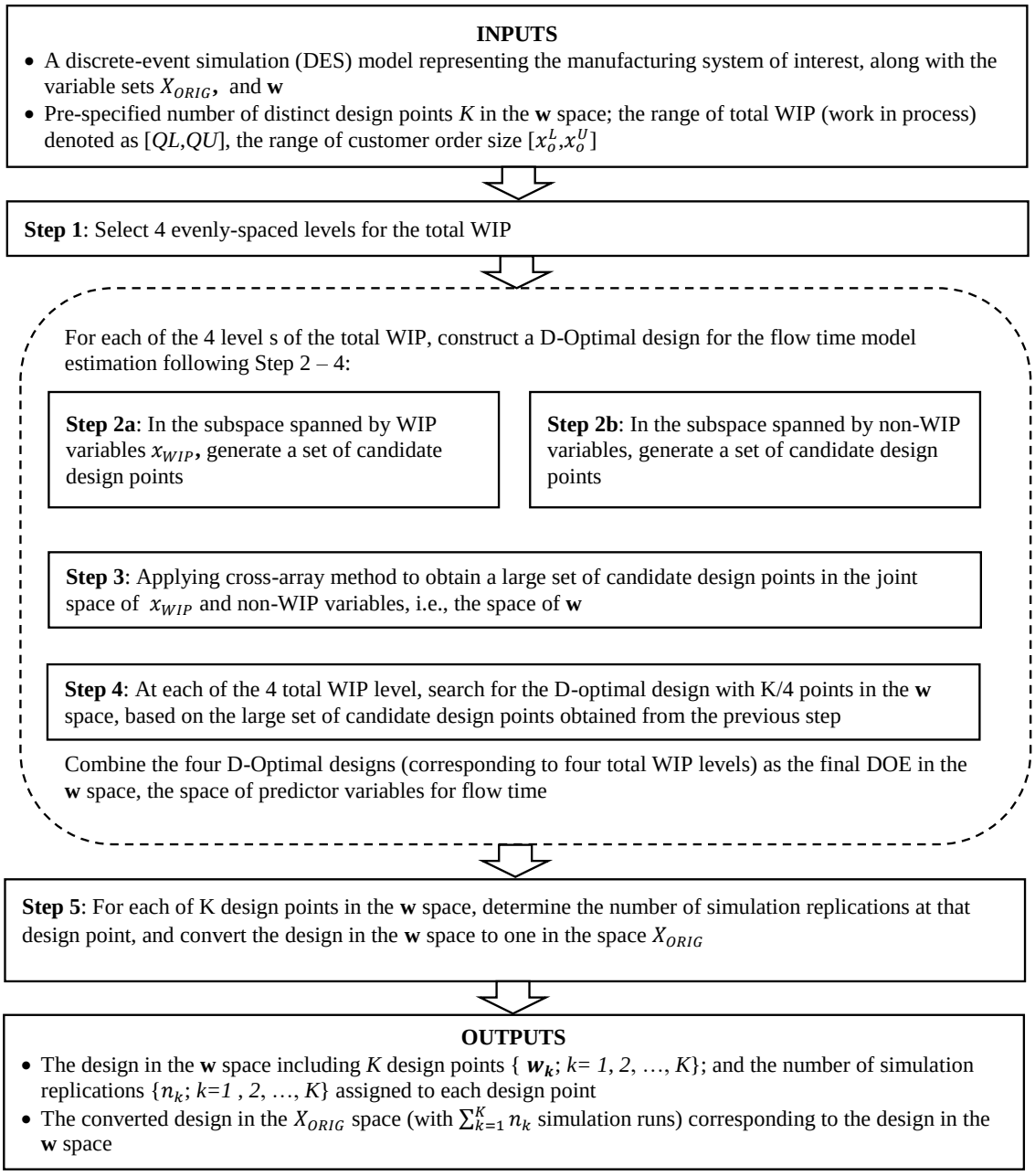


Figure 5.2: The design of experiments procedure suggested by [1].