Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

FLOW-TIME ESTIMATION BY SYNERGISTICALLY MODELING REAL AND SIMULATION DATA

Hoda Sabeti
Feng Yang

Department of Industrial and Management Systems Engineering
395 Evansdale Drive
Morgantown, West Virginia, USA

ABSTRACT

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. Quoting a precise and reliable lead time requires a good prediction for the flow time of a new order. 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.

1 INTRODUCTION

In today's competitive market, a manufacturer's ability to quote a competitive and reliable lead time for a new order is a key competitive advantage and plays a significant role in customer acquisition and satisfaction. Upon the arrival of a customer order, it is critical to accurately predict the flow time (the time needed to complete that job) and quote its lead time accordingly. 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.

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 (Wein 1991; Duenyas and Hopp 1995; Spearman and Zhang 1999; Savaşaneril, Griffin, and Keskinocak 2010; Altendorfer and Jodlbauer 2011). 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 (Lawrence 1995; Öztürk, Kayalıgil, and Özdemirel 2006; Pearn, Chung, and Lai 2007; Baykasoğlu, Göçken, and Unutmaz 2008). These surrogate models include classic linear regression (Hopp and Sturgis 2000; Vig and Dooley 1991; Hsu and Sha 2004; Sabuncuoglu and Comlekci 2002; Sha, Storch, and Liu 2007) as well as powerful models such as neural network (Hsu and Sha 2004; Philipoom, Rees, and Wiegmann 1994; Li et al. 2007).

In this stream of numerical work, Li et al. (2015) 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 adopts the stochastic kriging with qualitative factors (SKQ) (Wang et al. 2014) to surrogate the real process 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 highest quality.

2 Statement of the Research Problem

To assist the lead time quotation upon the arrival of a new order, an SKQ model, which quantifies the dependence of the expected job flow time upon the various shop-status factors, will be fitted 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.

The target manufacturing system may involve features such as random processing times, machine failures, batch processing, re-entrant flows, etc. As detailed in Li et al. (2015), 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 semiconductor fabrication system is considered as the example system in this paper, with the detailed configurations given in Appendix A for readers' convenience. For this example system, the original factors and factor numbers are provided in Table 1. The concept of buffers by Riano (2002) 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, Li et al. (2015) proposed a preliminary analytical analysis to find a smaller set of important variables, which can be classified as WIP and non-WIP variables. Table 2 provides for the example system the WIP and non-WIP variables included in Li et al. (2015) 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 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: 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

Table 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
	z_S the source of data	1

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

The SKQ developed in Wang et al. (2014) 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

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

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 (1) is arranged into two subsets

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

The replicated subset includes K ($0 \le K \le I$) 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

$$\begin{aligned} \mathscr{Y}_j(\mathbf{w}) &= \mathrm{E}[\mathscr{Y}(\mathbf{w})] + \varepsilon_j(\mathbf{w}) = \mathsf{Y}(\mathbf{w}) + \varepsilon_j(\mathbf{w}) & (3)\\ &= \mathbf{f}(\mathbf{w})^\top \beta + \mathsf{M}(\mathbf{w}) + \varepsilon_j(\mathbf{w}), & (4) \end{aligned}$$

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 (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 (Ankenman, Nelson, and Staum 2010). $\varepsilon_j(\mathbf{w})$ denotes the random error variability, and is referred to as the intrinsic variability (Ankenman, Nelson, and Staum 2010). The random errors $\varepsilon_1(\mathbf{w}), \varepsilon_2(\mathbf{w}), \ldots$, are assumed to be independent and identically distributed with mean zero.

3.1 Extrinsic Variance Structure

The extrinsic variability model is inherited from Qian, Wu, and Wu (2008), and reviewed as follows. The covariance of $\mathsf{M}(\mathbf{w})$ can be written as

$$\mathrm{Cov}[\mathsf{M}(\mathbf{w}), \mathsf{M}(\mathbf{w}')] = \delta^2 \cdot \mathrm{Corr}[\mathsf{M}(\mathbf{w}), \mathsf{M}(\mathbf{w}')] = \delta^2 \cdot \left[\prod_{\ell=1}^{L} \tau^{(\ell)}_{z_\ell, z'_\ell}\right] \cdot K(\mathbf{x}, \mathbf{x}'), \tag{5}$$

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 (Santner, Williams, and Notz 2003;Qian, Wu, and Wu 2008). A popular function is the exponential correlation function

$$K(\mathbf{x}, \mathbf{x}') = \exp\left\{\sum_{h=1}^{d} -\theta_h |x_h - x'_h|^p\right\} \tag{6}$$

with $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$ being unknown parameters. In (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 Φ denotes the unknown parameters involved in the cross-category correlation model. Potential functional forms for $\tau^{(\ell)}_{z_\ell, z'_\ell}$ are given in Wang et al. (2014).

Given the data (1) collected at I distinct settings, the $I \times I$ variance-covariance matrix Σ_{M} is defined as

$$\Sigma_{\mathsf{M}} = \delta^2 \cdot \mathbf{R}(\theta, \Phi) \tag{7}$$

$$= \delta^2 \cdot \begin{pmatrix} 1 & \mathrm{Corr}[\mathsf{M}(\mathbf{w}_1), \mathsf{M}(\mathbf{w}_2)] & \cdots & \mathrm{Corr}[\mathsf{M}(\mathbf{w}_1), \mathsf{M}(\mathbf{w}_I)] \\ \mathrm{Corr}[\mathsf{M}(\mathbf{w}_2), \mathsf{M}(\mathbf{w}_1)] & 1 & \cdots & \mathrm{Corr}[\mathsf{M}(\mathbf{w}_2), \mathsf{M}(\mathbf{w}_I)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Corr}[\mathsf{M}(\mathbf{w}_I), \mathsf{M}(\mathbf{w}_1)] & \mathrm{Corr}[\mathsf{M}(\mathbf{w}_I), \mathsf{M}(\mathbf{w}_2)] & \cdots & 1 \end{pmatrix}, \tag{8}$$

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

$$\Sigma_{\mathsf{M}}(\mathbf{w}_0, \cdot) = \delta^2 \mathbf{v}(\mathbf{w}_0, \theta, \Phi) = \delta^2 \begin{pmatrix} \mathrm{Corr}[\mathsf{M}(\mathbf{w}_0), \mathsf{M}(\mathbf{w}_1)] \\ \mathrm{Corr}[\mathsf{M}(\mathbf{w}_0), \mathsf{M}(\mathbf{w}_2)] \\ \vdots \\ \mathrm{Corr}[\mathsf{M}(\mathbf{w}_0), \mathsf{M}(\mathbf{w}_I)] \end{pmatrix}, \tag{9}$$

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

3.2 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) is a diagonal matrix

$$\Sigma_\varepsilon = \mathrm{diag}\{\frac{\mathrm{Var}[\varepsilon(\mathbf{w}_1)]}{n(\mathbf{w}_1)}, \ldots, \frac{\mathrm{Var}[\varepsilon(\mathbf{w}_K)]}{n(\mathbf{w}_K)}, \mathrm{Var}[\varepsilon(\mathbf{w}_{K+1})], \ldots, \mathrm{Var}[\varepsilon(\mathbf{w}_I)]\}. \tag{10}$$

3.3 Iterative Procedure for Model Estimation

In the data set (2), the vector of sample averages for replicated data is denoted as

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

with

$$\overline{\mathscr{Y}}(\mathbf{w}_i) = \frac{1}{n(\mathbf{w}_i)} \sum_{j=1}^{n(\mathbf{w}_i)} \mathscr{Y}_j(\mathbf{w}_i) \qquad\qquad i = 1, \ldots, K. \tag{11}$$

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

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

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

$$\mathscr{Y} \sim \mathrm{MVN}[\beta_0 \mathbf{1}_I, \Sigma_{\mathsf{M}} + \Sigma_\varepsilon]$$

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

$$\ln L(\beta_0, \delta^2, \theta, \phi) = -\ln[(2\pi)^{\frac{I}{2}}] - \frac{1}{2}\ln[|\delta^2 R(\theta, \phi) + \Sigma_\varepsilon|] - \frac{1}{2}(\mathscr{Y} - \beta_0 \mathbf{1}_I)^\top [\delta^2 R(\theta, \phi) + \Sigma_\varepsilon]^{-1} (\mathscr{Y} - \beta_0 \mathbf{1}_I). \tag{12}$$

Since the data set (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 Wang et al. (2014) 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 $\mathrm{Var}[\varepsilon(\mathbf{w}_i)]$ $(i = 1,2,\ldots,K)$ by

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

- For non-replicated data, set $\widehat{\mathrm{Var}}[\varepsilon(\mathbf{w}_i)] = v_0$ for $i = K+1,\ldots,I$. The initial variance estimate v_0 can be set as the median of the sample variances $\{\widehat{\mathrm{Var}}[\varepsilon(\mathbf{w}_i)]; i = 1,2,\ldots,K\}$.
- Assemble $\{\frac{\widehat{\mathrm{Var}}[\varepsilon(\mathbf{w}_i)]}{n(\mathbf{w}_i)}; i = 1,2,\ldots,K\}$ 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 (12), 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

$$\frac{\partial \ln L(\beta_0,\delta^2,\theta,\phi)}{\partial\beta_0} = 0, \tag{13}$$

and expressed as

$$\widehat{\beta}_0(\delta^2,\theta,\phi) = (1_I^T[\delta^2 R(\theta,\phi)+\widehat{\Sigma}_\varepsilon]^{-1}1_I)^{-1}(1_I^T[\delta^2 R(\theta,\phi)+\widehat{\Sigma}_\varepsilon]^{-1}\mathscr{Y}) \tag{14}$$

- Substitute $\widehat{\beta}_0(\delta^2,\theta,\phi)$ into (12) and maximize

$$\begin{aligned}\ln L(\delta^2,\theta,\phi) = &-\ln[(2\pi)^{\frac{I}{2}}] - \frac{1}{2}\ln[|\delta^2 R(\theta,\phi)+\widehat{\Sigma}_\varepsilon|] \\ &-\frac{1}{2}(\mathscr{Y}-\widehat{\beta}_0(\delta^2,\theta,\phi)1_I)^T[\delta^2 R(\theta,\phi)+\widehat{\Sigma}_\varepsilon]^{-1}(\mathscr{Y}-\widehat{\beta}_0(\delta^2,\theta,\phi)1_I)\end{aligned} \tag{15}$$

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

$$\widehat{\mathrm{Y}}(\mathbf{w}_i) = \widehat{\beta}_0 + \widehat{\delta}^2 v(\mathbf{w}_i,\widehat{\theta},\widehat{\phi})^T[\widehat{\Sigma}_M+\widehat{\Sigma}_\varepsilon]^{-1}(\mathscr{Y}-\widehat{\beta}_0\mathbf{1}_I)). \tag{16}$$

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

- Based on the estimates obtained from (16), calculate the squared residuals:

$$\widehat{e}^2(\mathbf{w}_i) = (\mathscr{Y}(\mathbf{w}_i) - \widehat{Y}(\mathbf{w}_i))^2 \quad i = K+1,K+2,\ldots,I \tag{17}$$

- 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 $(\widehat{\beta}_0,\widehat{\delta^2},\widehat{\theta},\widehat{\phi})$.

4 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 semiconductor wafer fabrication system, and only differ in some processing-time parameters at certain workstations. The DES model specified in Appendix B 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 A 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 proposed by Li et al. (2015) is first performed to identify $\mathbf{w}$, a set of relatively important variables. In this case, $\mathbf{w}$ includes 16 quantitative and 6 qualitative variables.

4.1 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 $\mathbf{w}$ 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 well-designed simulation experiments. 120 distinct design points were generated in the space of $\mathbf{w}$ following the experimental design method developed by Li et al. (2015), which seeks to optimize the D-criterion while providing a decent coverage of the design space. 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 Li et al. (2015) method, and ranges from 50 to 400 among the 120 design points.

4.2 Validation Data (VD)

The goal is to obtain a prediction model relating the mean CT to $\mathbf{w}$ 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 $\mathbf{w}$ space providing a dense and fairly even coverage of the design space. These check points are all different from the points in *ED_Real* or *ED_Sim*. At a check point $\mathbf{w}$, 500 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.

4.3 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):

$$\text{MAPE} = \frac{100\%}{2400}\sum_{i=1}^{2400}\left|\frac{\widehat{\mathsf{Y}}(\mathbf{w}_i)-\mathsf{Y}_T(\mathbf{w}_i)}{\mathsf{Y}_T(\mathbf{w}_i)}\right| \tag{18}$$

The estimated root mean squared error(ERMSE):

$$\text{ERMSE} = \sqrt{\frac{1}{2400}\sum_{i=1}^{2400}(\widehat{\text{Y}}(\mathbf{w}_i) - \text{Y}_T(\mathbf{w}_i))^2} \tag{19}$$

In (18) and (19), $\widehat{\text{Y}}(\mathbf{w}_i)$ is the estimated mean FT at a check point $\mathbf{w}_i$.

4.4 Comparison of Modeling Methods

Three prediction models are respectively obtained through three different venues.

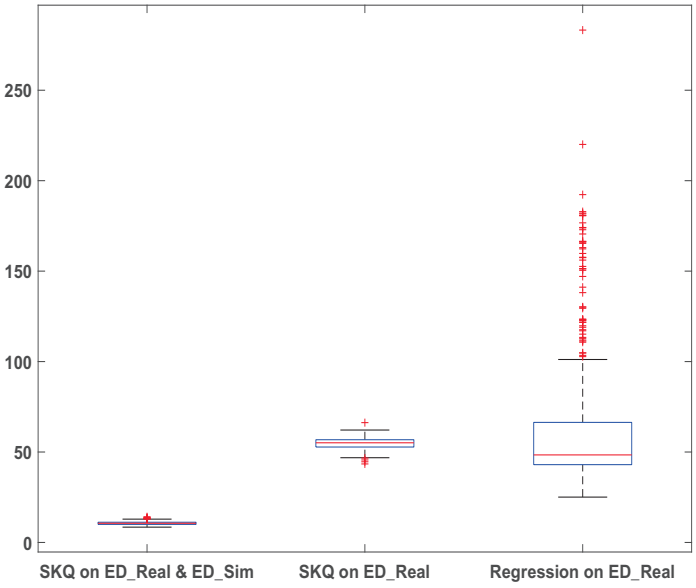
- Iterative SKQ on *ED_Real* and *ED_Sim*: The iterative SKQ procedure (Section 3.3) was applied to model the data ensemble of *ED_Real* and *ED_Sim* with the predictors being **w** including the qualitative variable that has two categorical levels, Real or Simulation.
- Iterative SKQ on *ED_Real*: On *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 Li et al. (2015) was applied with the predictors being **w** excluding the qualitative variable for real or simulation data.

To statistically compare these three approaches, 500 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 4.1, 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 500 fitted models (e.g., regression models), and 500 MAPEs and ERMSEs (4.3).

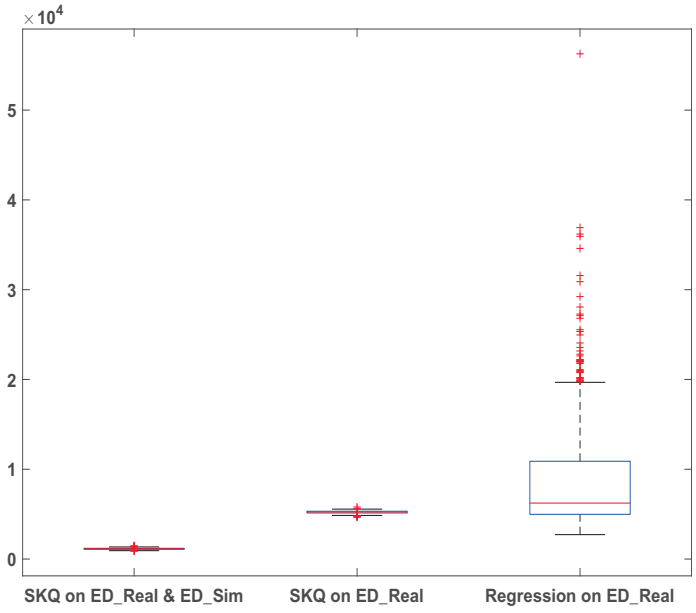
Figure 1(a) and (b) display the MAPE and ERMSE box plots respectively for the three approaches. Each box is plotted from the 500 MAPEs or ERMSEs for the corresponding approach. The medians of the boxes are also given in Table 3. Clearly, by borrowing information from the well-designed simulation data *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 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 3

Table 3: Medians of MAPEs and ERMSEs from macro-replications.

Method	MAPE	ERMSE
Iterative SKQ on *ED_Real* and *ED_Sim*	10.60%	1141.17
Iterative SKQ on *ED_Real*	55.10%	5242.91
Regression on *ED_Real*	48.40%	6233.71



(a) MAPE box plots.



(b) ERMSE box plots.

Figure 1: Comparison of model quality.

5 SUMMARY

An iterative procedure was developed to enable the SKQ (stochastic kriging with qualitative factors) fitting of both replicated simulation data, which can be well designed, and non-replicated real data, which are observed passively. Through the empirical case of estimating expected flow times, it is evident that the estimation quality is substantially improved by synergistically modeling real and simulation data. In this paper, the design of simulation experiments method was borrowed from a previous work and not tailored to the SKQ fitting of both real and simulation data. An immediate future work is to develop design of

simulation experiments so that the simulation data can best complement the real data in modeling the target response surface for a real system.

A CONFIGURATION OF THE EXAMPLE SYSTEM

The example system investigated in this paper is developed by Kayton et al. (1997) and also used in Li et al. (2015). This example system is a scaled-down semiconductor fabrication system. Customer order to the system is a homogeneous compound Poisson process with a rate ranging within $[0.194, 0.238]$ 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 2 shows the sequence of required processing steps and the stations that a job has to visit.

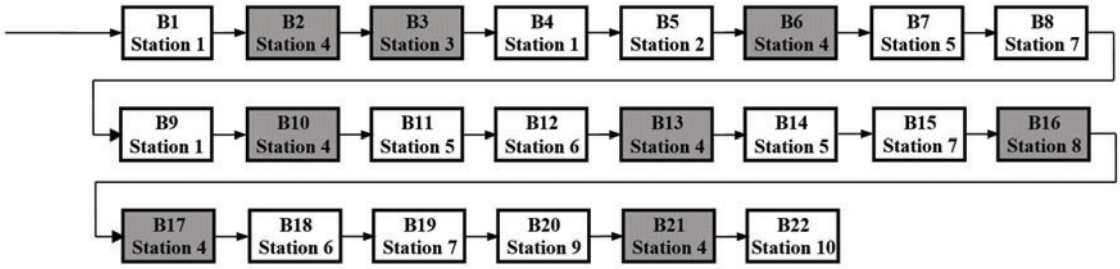


Figure 2: Job processing sequence and important workstations (Li et al. 2015).

As shown in Figure 2, Stations 1, 4, 5, 6 and 7 are revisited by jobs. Table 4 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 2 and Table 4, this manufacturing system involves major features 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 4: System parameters (Li et al. 2015).

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)	80	220	80	40	25	22	40	100	50	50
Stdev PT(min)	7	16	7	4	2	2.4	4	12	4	5

B CONFIGURATION OF THE SYSTEM FOR "REAL" DATA

The "real" system follows the same configuration as the one in Appendix A, and its parameters are provided in Table 5.

Table 5: Configuration of each workstation for collecting the actual 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)	84	235	110	64	30	25	42	130	54	54
Stdev PT(min)	8	18	10	60	3	3	5	15	5	7

REFERENCES

Altendorfer, K., and H. Jodlbauer. 2011. "An Analytical Model for Service Level and Tardiness in a Single Machine MTO Production System". *International Journal of Production Research* 49 (7): 1827–1850.

Ankenman, B., B. L. Nelson, and J. Staum. 2010. "Stochastic Kriging for Simulation Metamodeling". *Operations Research* 58 (2): 371–382.

Baykasoğlu, A., M. Göçken, and Z. D. Unutmaz. 2008. "New Approaches to Due Date Assignment in Job Shops". *European Journal of Operational Research* 187 (1): 31–45.

Duenyas, I., and W. J. Hopp. 1995. "Quoting Customer Lead Times". *Management Science* 41 (1): 43–57.

Hopp, W. J., and M. L. R. Sturgis. 2000. "Quoting Manufacturing Due Dates Subject to a Service Level Constraint". *IIE Transactions* 32 (9): 771–784.

Hsu, S., and D. Sha. 2004. "Due Date Assignment Using Artificial Neural Networks under Different Shop Floor Control Strategies". *International Journal of Production Research* 42 (9): 1727–1745.

Kayton, D., T. Teyner, C. Schwartz, and R. Uzsoy. 1997. "Focusing Maintenance Improvement Efforts in a Wafer Fabrication Facility Operating Under the Theory of Constraints". *Production and Inventory Management Journal* 38 (4): 51.

Lawrence, S. R. 1995. "Estimating Flowtimes and Setting Due-dates in Complex Production Systems". *IIE Transactions* 27 (5): 657–668.

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

Li, S., Y. Li, Y. Liu, and Y. Xu. 2007. "A GA-based NN Approach for Makespan Estimation". *Applied Mathematics and Computation* 185 (2): 1003–1014.

Öztürk, A., S. Kayalıgil, and N. E. Özdemirel. 2006. "Manufacturing Lead Time Estimation Using Data Mining". *European Journal of Operational Research* 173 (2): 683–700.

Pearn, W., S. Chung, and C. Lai. 2007. "Due-date Assignment for Wafer Fabrication under Demand Variate Environment". *IEEE Transactions on Semiconductor Manufacturing* 20 (2): 165–175.

Philipoom, P. R., L. P. Rees, and L. Wiegmann. 1994. "Using Neural Networks to Determine Internally-Set Due-Date Assignments for Shop Scheduling". *Decision Sciences* 25 (5-6): 825–851.

Qian, P. Z. G., H. Wu, and C. J. Wu. 2008. "Gaussian Process Models for Computer Experiments with Qualitative and Quantitative Factors". *Technometrics* 50 (3): 383–396.

Riano, G. 2002. *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.

Sabuncuoglu, I., and A. Comlekci. 2002. "Operation-based Flowtime Estimation in a Dynamic Job Shop". *Omega* 30 (6): 423–442.

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

Savaşaneril, S., P. M. Griffin, and P. Keskinocak. 2010. "Dynamic Lead-time Quotation for an M/M/1 Base-stock Inventory Queue". *Operations Research* 58 (2): 383–395.

Sha, D., R. Storch, and C.-H. Liu. 2007. "Development of a Regression-based Method with Case-based Tuning to Solve the Due Date Assignment Problem". *International Journal of Production Research* 45 (1): 65–82.

Spearman, M. L., and R. Q. Zhang. 1999. "Optimal Lead Time Policies". *Management Science* 45 (2): 290–295.

Vig, M. M., and K. J. Dooley. 1991. "Dynamic Rules for Due-date Assignment". *The International Journal of Production Research* 29 (7): 1361–1377.

Wang, K., X. Chen, F. Yang, D. W. Porter, and N. Wu. 2014. "A New Stochastic Kriging Method for Modeling Multi-Source Exposure–Response Data in Toxicology Studies". *ACS Sustainable Chemistry & Engineering* 2 (7): 1581–1591.

Wein, L. M. 1991. "Due-date Setting and Priority Sequencing in a Multiclass M/G/1 Queue". *Management Science* 37 (7): 834–850.

AUTHOR BIOGRAPHIES

Hoda Sabeti is a Ph.D. candidate in the Industrial and Management Systems Engineering Department at West Virginia University. Her research interests include discrete-event simulation, metamodeling, and their applications in manufacturing, service and transportation. Her e-mail address is Hoda.Sabeti@gmail.com.

Feng Yang is currently an Associate Professor in the Industrial and Management Systems Engineering Department at West Virginia University. She received her Ph.D. degree in Industrial Engineering and Management Sciences from Northwestern University in 2006. Her research has been focusing on metamodeling, design of experiments and applied statistical methods, and she is interested in developing methodologies for various applications including production planning for manufacturing, service management, nanotoxicology, and sensor-based environment monitoring. Her email address is Feng.Yang@mail.wvu.edu.