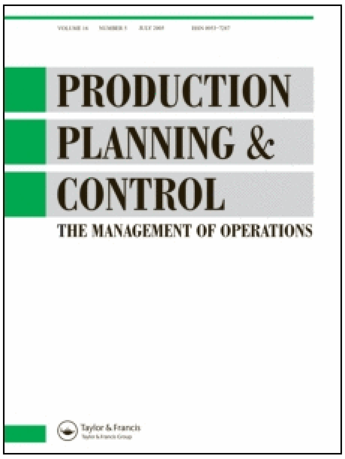


Production Planning & Control
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713737146

Metamodelling for cycle time-throughput-product mix surfaces using progressive model fitting

Feng Yang[a]; Jingang Liu[a]; Barry L. Nelson[b]; Bruce E. Ankenman[b]; Mustafa Tongarlak[b]
[a] Department of Industrial and Management Systems Engineering, West Virginia University, Morgantown, WV 26506, USA [b] Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208-3119, USA

Online publication date: 15 December 2010

To cite this Article Yang, Feng , Liu, Jingang , Nelson, Barry L. , Ankenman, Bruce E. and Tongarlak, Mustafa(2011) 'Metamodelling for cycle time-throughput-product mix surfaces using progressive model fitting', Production Planning & Control, 22: 1, 50 — 68

To link to this Article: DOI: 10.1080/09537287.2010.490026

URL: http://dx.doi.org/10.1080/09537287.2010.490026

Production Planning & Control
Vol. 22, No. 1, January 2011, 50–68

Metamodelling for cycle time-throughput-product mix surfaces using progressive model fitting

Feng Yang[a]*, Jingang Liu[a], Barry L. Nelson[b], Bruce E. Ankenman[b] and Mustafa Tongarlak[b]

[a]*Department of Industrial and Management Systems Engineering, West Virginia University, P.O. Box 6070, Morgantown, WV 26506, USA;* [b]*Department of Industrial Engineering and Management Sciences, Northwestern University, 2145 Sheridan Rd., Evanston, IL 60208-3119, USA*

(*Final version received 16 April 2010*)

A simulation-based methodology is proposed to map the mean of steady-state cycle time (CT) as a function of throughput (TH) and product mix (PM) for manufacturing systems. Nonlinear regression models motivated by queueing analysis are assumed for the underlying response surface. To ensure efficiency and control estimation error, simulation experiments are built up sequentially using a multi-stage procedure to collect data for fitting the models. The resulting response surface is able to provide a CT estimate for any TH and any PM, and thus allows the decision maker to instantly investigate options and trade offs regarding their production planning.

Keywords: discrete event simulation; response surface modelling; design of experiments; queueing; semiconductor manufacturing

1. Introduction

Planning for manufacturing, either at the factory or enterprise level, requires answering what-if questions involving (perhaps a very large number of) different scenarios for product mix (PM), production targets and capital expansion. Computer simulation is an essential tool for the design and analysis of complex manufacturing systems. Often, before a new system is deployed or changes are made to an existing system, a simulation model will be created to predict the system's performance. Even when no substantial changes are envisioned, simulation is used to allocate capacity among production facilities. In either case, simulation is faster and much more cost effective than experimenting with the physical system (when that is even possible). This is especially true in the semiconductor industry, which is the motivating application for this research (see, e.g. Schömig and Fowler 2000).

In semiconductor manufacturing, many man-hours are invested in developing and exercising simulation models of wafer fab systems. These models include critical details that are difficult or impossible to incorporate into simple load calculations or queueing approximations. Unfortunately, simulation models can be clumsy tools for planning or decision-making because even a few minutes per simulation run (which is optimistic) is too slow to allow what-if analysis in real time. In our research, we develop techniques to support strategic planning from a new perspective: we combine computing horsepower, adaptive statistical methods and queueing theory to make simulation a much more effective tool than before.

Our objective is to estimate the mean of steady-state cycle time (CT) as a function of the input decision variables throughput (TH) and PM. CT, technically is defined as a random variable representing the time required for a job or lot to traverse a given routing in a production system (e.g. Hopp and Spearman 2001). In the remainder of this article, we will use 'CT' to refer to the mean of this random variable. A semiconductor manufacturing system can control CT by controlling two decision variables, PM and release rate at which lots are started in the factory (lot-start rate or equivalently, TH rate). Hence, the CT–TH–PM surface can play an important role in strategic planning of semiconductor manufacturing including evaluating the mean of CT for a given TH and PM, determining the sensitivity of product CTs to changes in TH or PM, determining feasible THs that satisfy CT constraints, and finding a PM that maximises revenue subject to CT and TH constraints.

The proposed methodology is able to generate a complete CT–TH–PM surface (with the response of interest being the long-run average CT of products) like that provided by a tractable queueing model, but with the fidelity of simulation. In light of the

*Corresponding author. Email: feng.yang@mail.wvu.edu

ISSN 0953–7287 print/ISSN 1366–5871 online
© 2011 Taylor & Francis
DOI: 10.1080/09537287.2010.490026
http://www.informaworld.com

comprehensive nature of the CT–TH–PM surface, we integrate two different analysis approaches to map the desired response surface: (i) queueing theory and (ii) metamodelling, namely simulation-based response surface modelling. The former approximates system capacity and identifies bottleneck (BN) resources, which facilitates the definition and normalisation of the feasible TH–PM region (Section 2.2). Furthermore, the queueing analysis leads to the division of the feasible region into a number of subregions which allows for the fitting of a smooth CT–TH–PM surface within each subregion (Section 5.2). Metamodelling, which is the primary focus of this research, acts on a given simulation model of a manufacturing system by performing simulation experiments sequentially at selected settings of TH and PM until a desired precision has been achieved on the estimation of the response surface. Our CT–TH–PM model fitting is based on the assumption that the underlying surface can be captured by the proposed regression models, the forms of which are motivated by the analysis of simple queueing systems. Such a response surface is able to provide a CT estimate for any TH and any PM, and thus allows the decision maker to investigate options and trade offs almost instantly without running additional simulations.

The remainder of this article is organised as follows. Section 2 describes the research problem in precise terms and introduces some notation. Section 3 provides an overview of the methodology proposed for the generation of a CT–TH–PM surface, substantiated by the technical details presented in Sections 4 and 5. Section 6 describes the complete multi-stage procedure for estimating a CT–TH–PM surface via sequential simulation experiments. Section 7 provides an empirical evaluation of the proposed method.

2. Statement of the problem

As noted, our goal is to approximate the mean cycle time as a function of the manufacturing system TH and PM. For the generation of such a CT–TH–PM surface, we integrate analytical queueing analysis and simulation-based statistical modelling. In this section, we define most of the notation that will be used in these two analysis approaches, and state the research problem in more precise terms.

2.1. *Analytical formulation of the product system*

To perform the analytical queueing analysis as mentioned above, we represent the manufacturing system as a multi-product queueing network. Suppose that the system (e.g. wafer fab) consists of M different stations, and it is designed to process K types of products with each one following a different deterministic routing. We define the system in a generic way as follows.

- $\{s_j, j=1,2,\dots,M\}$: the number of parallel resources at station j.
- $\{u_{kj}, k=1,2,\dots,K; j=1,\ 2,\dots,M\}$: the effective service rate of each resource at station j for products of type k.
- $\{\delta_{kj}, k=1,2,\dots,K; j=1,\ 2,\dots,M\}$: the number of visits by product type k to station j.

The product flow is described by

- λ: the overall release rate of all the products into the system.
- $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\dots,\alpha_K)$: the product-mix vector with each element α_i representing the fraction of type k products in the flow, so that $\sum_{k=1}^{K}\alpha_k = 1$, $\alpha_k \in [0,1]$.
- $\lambda_k=\alpha_k\lambda$: the release rate of product type k to the system.

Under this formulation, capacity/BN analysis can be performed as follows. We can easily calculate ρ_j, the utilisation of station j $(j=1,2,\dots,M)$. Let $\rho_{kj}=\delta_{kj}/(s_j u_{kj})$, then $\rho_j = \lambda\sum_{k=1}^{K}\alpha_k\rho_{kj}$. The maximum utilisation $\rho_{\max}=\max_j \rho_j$ is called the system utilisation and is denoted by x in this article. A station, say station j_{BN}, that reaches $\rho_{\max}$ is called a BN station, that is,

$$j_{\mathrm{BN}} = \mathrm{argmax}_j\rho_j = \mathrm{argmax}_j\sum_{k=1}^{K}\alpha_k\rho_{kj}. \tag{1}$$

The stability constraint on the system requires $x=\lambda\sum_{k=1}^{K}\alpha_k\rho_{kj_{\mathrm{BN}}} < 1$, or equivalently,

$$\lambda < 1/\sum_{k=1}^{K}\alpha_k\rho_{kj_{\mathrm{BN}}} = u^*(\boldsymbol{\alpha}), \tag{2}$$

where $u^*(\boldsymbol{\alpha})$ is the system capacity, the upper limit on λ (or overall TH) for stability. Obviously, both capacity $u^*(\boldsymbol{\alpha})$ and the BN station j_{BN} depend on the system parameters as well as $\boldsymbol{\alpha}$.

Our metamodelling of the CT–TH–PM surface requires an analytical pre-analysis to obtain $u^*(\boldsymbol{\alpha})$ and identify the BN station as a function of $\boldsymbol{\alpha}$. The core of such capacity/BN analysis is the estimation of effective service rate $\{u_{kj}, k=1,2,\dots,K; j=1,2,\dots,M\}$ at each workstation of the system for any given PM. Methods for computing the effective service rates of realistic manufacturing systems are beyond the scope of this article, and we rely instead on existing queueing models and approximation methods to perform our capacity/BN pre-analysis. We refer interested readers to Kumar and Kumar (2001), Hopp *et al.* (2002), Meng and Heragu (2004), Morrison and Martin (2007) and Shantikumar *et al.* (2007).

In this article, we assume that the existing methods can provide reasonably good capacity/BN analysis. Nevertheless, as pointed out by Jacobs (2004) and Wu *et al.* (2007), queueing models can fall short of capturing many realistic features of real manufacturing systems, and this may lead to approximation errors in the estimated effective service rates. This inaccuracy will certainly affect our simulation-based CT–TH–PM modelling. However, we believe that simplified queueing models are more accurate for approximating the effective service rates rather than they are for estimating the expected CTs, which will be handled by simulation in our research. Limited by the scope of this article, we focus on the metamodelling aspect from here on. Note that in the case study of Section 7.2, we used the analytic engine provided by Factory Explorer (an integrated capacity, cost and discrete-event simulation software package particularly suitable for modelling wafer fabs) to estimate the system capacity and identify the BN station(s) as a function of the PM for a realistic semiconductor manufacturing system, and obtained highly accurate CT–TH–PM surfaces.

2.2. *CT–TH–PM surface*

As will become clearer later, after invoking the analytic engine to perform capacity/BN analysis for the normalisation and partition of the TH-PM region, simulation experiments will be performed to collect data for the fitting of the desired CT response surface. The response of interest is the mean of steady-state CT for products of type k ($k = 1, 2, \ldots, K$), denoted $c_k(\lambda, \boldsymbol{\alpha})$. Different types of products may follow different processing steps, and thus have different CT distributions. For each type of product, we seek to estimate its long-run average CT, which depends on the overall product flow through the system. The product flow is characterised by start rate/TH λ and PM $\boldsymbol{\alpha}$, and in our work, we consider λ and $\boldsymbol{\alpha}$ as independent variables that can be controlled by the production manager (equivalently, the decision variables are $\{\lambda_k = \alpha_k\lambda,\ k = 1,\ 2, \ldots, K\}$, the start rates of each product type). As established in Section 5.1, the stability condition of the system is such that λ has to be less than the system capacity $u^*(\boldsymbol{\alpha})$, which can be analytically approximated for given values of $\boldsymbol{\alpha}$. The constraints on PM $\boldsymbol{\alpha}$ will be discussed in Section 5.1. Hence, with a prior queueing analysis of the system, the feasible region for decision variables λ and $\boldsymbol{\alpha}$ is well-defined for the simulation-based CT–TH–PM model fitting.

For reasons that will become apparent in Section 3, instead of estimating $c_k(\lambda, \boldsymbol{\alpha})$ we normalise the range of λ across the PM region, and directly estimate $c_k(x, \boldsymbol{\alpha})$ where $x = \lambda / u^*(\boldsymbol{\alpha})$ is the fraction of system capacity in use and x is on the scale of $[0, 1)$ regardless of the value of $\boldsymbol{\alpha}$. Once we have obtained $c_k(x, \boldsymbol{\alpha})$, a simple transformation will give us $c_k(\lambda, \boldsymbol{\alpha})$. Again, system capacity $u^*(\boldsymbol{\alpha})$ obtained from queueing analysis is what makes this transformation possible.

To model the CT–TH–PM surface, the most straightforward way is to develop a response-surface model that incorporates x and $\boldsymbol{\alpha}$ as independent variables. However, our investigation of analytically tractable queueing network models convinces us that a general model for $c_k(x, \boldsymbol{\alpha})$ is unlikely to be successful because the correct form of the model depends on specifics of the network topology of the factory. Therefore, we propose a two-step methodology for the generation of the CT–TH–PM surface, which is described in the next section.

3. Overview of the methodology

In this section, we provide an overview of the simulation-based response surface modelling for the CT–TH–PM surface. Our objective is to estimate the CT measure at any normalised TH x and for any feasible PM $\boldsymbol{\alpha}$. In light of the issues discussed in Section 2.2, we decided to utilise our success in estimating CT–TH curves with a fixed PM. We propose first using simulation to fit CT–TH curves for a carefully selected collection of PMs, and then perform model fitting across the $\boldsymbol{\alpha}$-space. Notationally, we define:

- $c_{k,x}(\boldsymbol{\alpha})$: the expected CT of product k at fixed TH x as a function of PM $\boldsymbol{\alpha}$.
- $c_{k,\boldsymbol{\alpha}}(x)$: the expected CT of product k for a given PM $\boldsymbol{\alpha}$ as a function of TH x.
- $\mathcal{A} = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_n\}$: a collection of n PMs.

Our methodology consists of two steps:

(i) Use the procedure developed in Yang *et al.* (2007) to estimate a collection of CT–TH curves $\{c_{k,\boldsymbol{\alpha}}(x),\ k = 1, 2, \ldots, K;\ \boldsymbol{\alpha} \in \mathcal{A}\}$ over a user-specified TH range $x \in [x_L, x_U]$ $(0 < x_L < x_U < 1)$. We take products of type 1 for example. Figure 1 shows the CT–TH curves for product 1 with each curve corresponding to a different PM vector $\boldsymbol{\alpha}_i \in \mathcal{A}$ $(i = 1, 2, 3)$. Note that by making the independent variable for each product's CT–TH curve the fraction

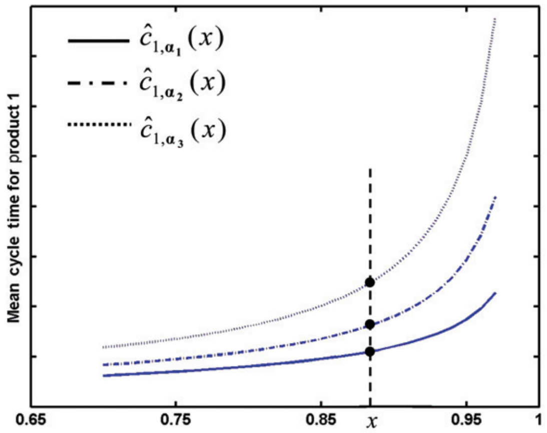


Figure 1. CT–TH curves for product 1 at different PMs.

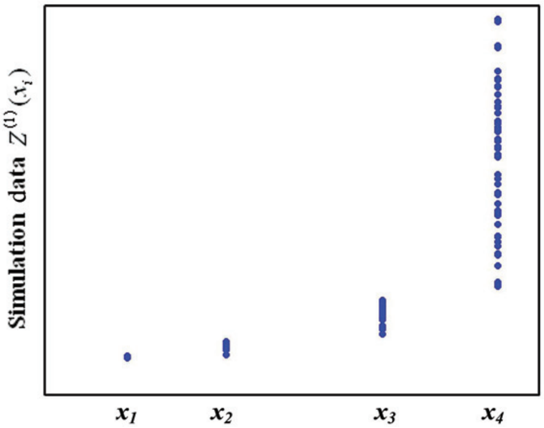


Figure 2. CT data from simulation replications for product 1 at different THs.

of system capacity at a given PM, all curves run from $[x_L, x_U]$ providing a common scale, which is the reason why we chose to estimate the CT response with respect to the normalised TH x rather than the actual TH λ. A brief review of the estimation of CT–TH curves is given in Section 4.

(ii) As illustrated in Figure 1, for products of type k $(k = 1, 2, \dots, K)$ a number of estimated CT–TH curves can be generated at a selected set of PMs $\{\widehat{c}_{k,\boldsymbol{\alpha}}(x), \boldsymbol{\alpha} \in \mathcal{A}\}$ with $x \in [x_L, x_U]$ (the fitted model is written as $\widehat{c}_{k,\boldsymbol{\alpha}}(x)$), and from these curves we can predict at any TH $x \in [x_L,\ x_U]$ the CTs $\widehat{c}_{k,\boldsymbol{\alpha}}(x) = \widehat{c}_{k,x}(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} \in \mathcal{A}$. Based on the 'data points' $\{\widehat{c}_{k,x}(\boldsymbol{\alpha}), \boldsymbol{\alpha} \in \mathcal{A}\}$ denoted as black dots in Figure 1, model fitting can be performed over the entire $\boldsymbol{\alpha}$-space to obtain $\widehat{c}_{k,x}(\boldsymbol{\alpha})$ for a given x. Section 5 provides the detailed methods for the fitting of CT–PM surface $c_{k,x}(\boldsymbol{\alpha})$ at a fixed TH x.

As will be illustrated empirically in Section 7, the method proposed above takes as input a simulation model representing a realistic manufacturing system, selects a collection of PMs $\mathcal{A}$ and generates via simulation a number of CT–TH curves $\{\widehat{c}_{k,\boldsymbol{\alpha}}(x), k = 1, 2, \dots, K; \boldsymbol{\alpha} \in \mathcal{A}\}$, $x \in [x_L,\ x_U]$ (Figure 1 displays such curves for product 1). Based on these curves, CT–PM fitting can be performed, and hence CT estimates for any PM at any TH level can be derived. Next, we discuss the technical details of the proposed method.

4. Review of the estimation of CT–TH Curves

Estimating CT–TH curves over a collection of PMs $\mathcal{A}$ is the primary step for generating the CT–TH–PM response surface, which provides the basis for the estimation of CT across PM space. In Yang *et al.* (2007), a simulation-based method was proposed for the generation of CT–TH curves at a fixed PM, which we briefly review as follows.

A nonlinear regression model (3), which is motivated by heavy-traffic queueing analysis, was developed to represent the underlying CT–TH curve

$$c_{k,\boldsymbol{\alpha}}(x) = \frac{\sum_{\ell=0}^{t} c_\ell x^\ell}{(1-x)^p}. \qquad (3)$$

For notational convenience we omit the subscript $\boldsymbol{\alpha}$ for the model in (3). Our task here is to generate CT–TH curves $\{c_{k,\boldsymbol{\alpha}}(x), k = 1, 2, \dots, K\}$ for a fixed PM $\boldsymbol{\alpha} \in \mathcal{A}$.

To estimate models (3) efficiently, Yang *et al.* (2007) proposed a multi-stage procedure which builds up simulation experiments until a desired precision has been achieved for the curve fitting. Specifically, with PM $\boldsymbol{\alpha}$ fixed, simulation is performed sequentially at different TH levels $\{x_1, x_2, \dots\}$ for data collection. At a certain TH x_i, CT data $\{Z^{(k)}(x_i),\ k = 1,\ 2, \dots, K\}$ (the average CT for all the type k products simulated in a simulation run) can be obtained from a simulation replication which is carried out with multiple types of product flows. Figure 2 gives an example of the x_i versus $Z^{(1)}(x_i)$ plots obtained for CT–TH fitting of product 1 with multiple replications performed at each x_i. Hence, based on a single set of simulation experiments performed at different THs, K data sets $\{Z^{(k)}(x_i), k = 1, 2, \dots, K\}$ can be obtained, from which K CT–TH curves $\{c_{k,\boldsymbol{\alpha}}(x), k = 1, 2, \dots, K\}$ can be fitted.

As will be articulated in Section 6, throughout the generation of CT–TH–PM surface, we assume that a user-specified type of product (say type 1) and a certain TH level x_0 are of particular interest to production planning. Hence in the CT–TH estimation, we control the estimation error on the CT estimates $\widehat{c}_{1,\boldsymbol{\alpha}}(x_0)$. As explained in Yang *et al.* (2007), the CT estimator $\widehat{c}_{1,\boldsymbol{\alpha}}(x_0)$ obtained from the fitted model (3) is approximately (asymptotically) normally distributed, that is, $\widehat{c}_{1,\boldsymbol{\alpha}}(x_0) \sim \mathrm{Norm}(c_{1,\boldsymbol{\alpha}}(x_0), \widehat{\mathrm{Var}}[\widehat{c}_{1,\boldsymbol{\alpha}}(x_0)])$. Note that $c_{1,\boldsymbol{\alpha}}(x_0)$ is the true unknown CT, and $\widehat{\mathrm{Var}}[\widehat{c}_{1,\boldsymbol{\alpha}}(x_0)]$ is the variance estimate of $\widehat{c}_{1,\boldsymbol{\alpha}}(x_0)$ which can be obtained from the nonlinear regression (3).

The quality of the curve fitting is evaluated in terms of $\sqrt{\widehat{\mathrm{Var}}[\widehat{c}_{1,\boldsymbol{\alpha}}(x_0)]})$, the precision achieved on $\widehat{c}_{1,\boldsymbol{\alpha}}(x_0)$. The sequential simulation is performed until

$$\sqrt{\widehat{\mathrm{Var}}[\widehat{c}_{1,\boldsymbol{\alpha}}(x_0)]}) \leq \sigma, \tag{4}$$

where σ is a user-specified parameter. Here we adopt the stopping rule (4) instead of the relative precision criterion used in Yang *et al.* (2007). As mentioned in Section 3, since the CT estimates obtained from CT–TH curves serve as 'data points' in the CT–PM modelling, (4) ensures that the variance of these 'data points' $\{\widehat{c}_{1,\boldsymbol{\alpha}}(x_0) = \widehat{c}_{1,x_0}(\boldsymbol{\alpha}), \boldsymbol{\alpha} \in \mathcal{A}\}$ is approximately σ^2 across the PM region, which justifies the constant–variance assumption for the least-squares fitting of the CT–PM model. Moreover, with the variance σ^2 being a known value we are able to derive valid statistical inference from the CT–PM nonlinear regression (Section 5.5.3).

5. CT–PM response surface

As outlined in Section 3, once a collection of CT–TH curves $\{c_{k,\boldsymbol{\alpha}}(x),\ k=1,2,\ldots,K;\ \boldsymbol{\alpha}\in\mathcal{A}\}$, $x\in[x_L, x_U]$ have been estimated, the next step is to fit the CT–PM surface. Namely, for a given system utilisation x, we estimate $c_{k,x}(\boldsymbol{\alpha})$ for products of type k $(k=1,2,\ldots,K)$.

5.1. *The feasible PM Space*

Obviously, PM $\boldsymbol{\alpha}$ has to satisfy:

$$\sum_{k=1}^{K} \alpha_k = 1, \quad \alpha_k \in [0, 1]. \tag{5}$$

Figure 3(a) illustrates the feasible PM region in a three-product case defined by constraint (5). In practice, the PM is usually subject to additional linear constraints imposed by realistic situations (e.g. lower bounds on release rates). We use the following notation to represent the linear constraints on PM

$$\mathbf{A}\boldsymbol{\alpha} \leq \mathbf{b}, \tag{6}$$

where $\mathbf{A}$ is a matrix of K columns with each row representing a constraint. Figure 3(b) gives an example of the more restricted PM region defined by (5) and (6).

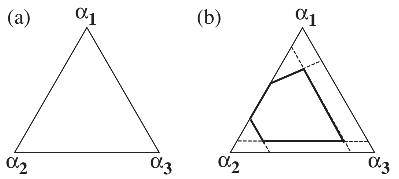


Figure 3. Feasible PM space: unconstrained (a) and constrained (b).

5.2. *Partitioning the PM space*

Production systems are usually constrained by one or more BN resources. A BN is usually a facility or resource which most constrains the production flow, and it plays a key role in determining the overall performance of the manufacturing system. As we change the PM, the BN may shift from one resource to another, which complicates the way that PM affects the CT. As will be seen in Section 5.3, within an $\boldsymbol{\alpha}$-region where no BN shift occurs, $c_{k,x}(\boldsymbol{\alpha})$ tends to be smooth and differentiable with respect to $\boldsymbol{\alpha}$. For the purpose of modelling the CT–PM surface, we divide the PM space into a number of subregions with each one dominated by a different BN station or stations, and fit the response surface for each subregion individually.

Suppose the feasible PM region of feasibility is defined as

$$\Omega = \{\boldsymbol{\alpha} | \boldsymbol{\alpha} \text{ satisfies constraints (5) and (6)}\}.$$

Following the definition of the BN station provided by (1), the subregion

$$\Omega_\nu = \{\boldsymbol{\alpha} | \boldsymbol{\alpha} \in \Omega \text{ and } \boldsymbol{\alpha} \text{ mix makes station } \nu \text{ a BN}\}$$

is given as the collection of $\boldsymbol{\alpha}$ that satisfies

$$\begin{aligned} &\sum_{k=1}^{K} \alpha_k = 1 \\ &\mathbf{A}\boldsymbol{\alpha} \leq \mathbf{b} \\ &\rho_\nu \geq \rho_j \quad j = 1, 2, \ldots, M. \end{aligned} \tag{7}$$

Following up on the three-product example discussed in Section 5.1, we further suppose that the system consists of three stations. It can be shown that for such a system the feasible region displayed in Figure 3(b) could be divided in three different ways as shown in Figure 4 depending on the system parameters. As explained in Section 2.1, queueing network models have been developed to obtain station utilisation ρ_j $(j=1,2,\ldots,M)$ as a function of PM $\boldsymbol{\alpha}$. Thus, the partition of the PM region into constant-BN subregions can be realised from analytical analysis prior to the running of any simulation experiments.

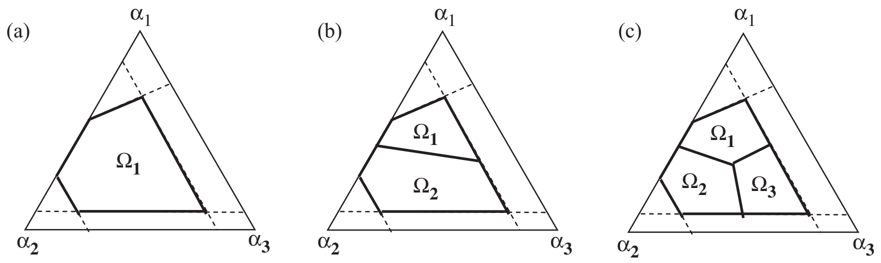


Figure 4. Division of the feasible PM region into constant BN regions.

For the case study in Section 7.2, we use the analytic engine in Factory Explorer to divide the PM region for a wafer fab.

In the remainder of this section, we discuss the statistical issues related to fitting the CT–PM surface.

5.3. *Form of the CT–PM model*

To estimate $c_{k,x}(\boldsymbol{\alpha})$ for product k, we developed a nonlinear regression model to represent the underlying CT–PM surface, the form of which is motivated by a simple open Jackson network.

Following the notation in Section 2.1, we consider a Jackson network in which each station has a single server having exponentially distributed service time with rate u_j (independent of product type). Given the system parameters for this network, the expected CT for each product type can be derived analytically as a function of PM $\boldsymbol{\alpha}$. Since all $c_{k,x}(\boldsymbol{\alpha})$ $(k=1,2,\ldots,K)$ functions have the same form, we consider the CT of product 1 without loss of generality:

$$c_{1,x}(\boldsymbol{\alpha}) = \sum_{j=1}^{M} \frac{\delta_{1j}}{u_j\left[1 - x\left(\frac{\sum_{k=1}^{K}\alpha_k\delta_{kj}/u_j}{\max_h \sum_{k=1}^{K}\alpha_k\delta_{kh}/u_h}\right)\right]} \qquad \boldsymbol{\alpha}\in\Omega. \tag{8}$$

Note that a station that achieves $\max_h \sum_{k=1}^{K}\alpha_k\delta_{kh}/u_h$ is a BN station. Within a subregion Ω_ν defined by (7) where station ν stays the BN, (8) can be written as

$$C_{1,x}(\boldsymbol{\alpha}) = \sum_{j=1}^{M} \frac{\delta_{1j}}{u_j\left[1 - x\left(\frac{\sum_{k=1}^{K}\alpha_k\delta_{kj}/u_j}{\sum_{k=1}^{K}\alpha_k\delta_{k\nu}/u_\nu}\right)\right]} \qquad \boldsymbol{\alpha}\in\Omega_\nu. \tag{9}$$

It is obvious from (9) that the CT is a continuous and differentiable function of $\boldsymbol{\alpha}$ within a constant-BN subregion. This motivates us to separately fit a regression model to each subregion Ω_ν.

Moreover, with simple mathematical manipulation, (9) can be rewritten as:

$$c_{1,x}(\boldsymbol{\alpha}) = \sum_{j=1}^{M}\frac{\sum_{k=1}^{K} a_{kj}\alpha_k}{\sum_{k=1}^{K} h_{kj}\alpha_k} = e_0 + \sum_{j=1}^{M}\frac{\sum_{k=1}^{K-1} e_{kj}\alpha_k}{h_{0j} + \sum_{k=1}^{K-1} h_{kj}\alpha_k}, \tag{10}$$

where α_K is eliminated by noting that $\alpha_K = 1 - \sum_{k=1}^{K-1}\alpha_k$. All the coefficients a_{kj}, h_{kj}, e_{kj} and e_0 depend on system parameters only.

Motivated by (10), we adopt a nonlinear regression model (11), which will be referred as the CT–PM model, to approximate the CT–PM surface for product of type k within a constant-BN region

$$\begin{aligned} c_{k,x}(\boldsymbol{\alpha}) = \mu(\boldsymbol{\alpha}, \boldsymbol{\theta}_R) &= \tau + \sum_{r=1}^{R} f(\boldsymbol{\alpha}, \mathbf{b}_r) \\ &= \tau + \sum_{r=1}^{R}\frac{\sum_{k=1}^{K-1} b_{kr}\alpha_k}{b_{0r} + \sum_{\ell=1}^{K-1} d_{\ell r}\alpha_\ell}. \end{aligned} \tag{11}$$

We make the following comments regarding (11).

- For notational convenience, we omit the subscripts and use $\mu(\boldsymbol{\alpha}, \boldsymbol{\theta}_R)$ to represent a CT–PM surface $c_{k,x}(\boldsymbol{\alpha})$ at a fixed TH x for product k.
- $(\alpha_1, \alpha_2, \ldots, \alpha_{K-1})$ are independent variables in (11), and $\alpha_K = 1 - \sum_{k=1}^{K-1}\alpha_k$ is eliminated from the model due to linear dependence. For notational convenience, we use vector $\boldsymbol{\alpha}$ to represent a PM setting $(\alpha_1, \alpha_2, \ldots, \alpha_K)$ and independent variables $(\alpha_1, \alpha_2, \ldots, \alpha_{K-1})$ interchangeably. One should bear in mind that in CT–PM modelling, there are only $K-1$ independent PM variables.
- The vector of unknown parameters $\boldsymbol{\theta}_R$ includes the constant term τ, and the coefficients $\mathbf{b}_r = (b_{0r}, b_{1r}, \ldots, b_{K-1,r})$ where $r = 1, 2, \ldots, R$.
- Model (11) is the sum of R ratio models

$$f(\boldsymbol{\alpha}, \mathbf{b}_r) = \frac{\sum_{k=1}^{K-1} b_{kr}\alpha_k}{b_{0r} + \sum_{\ell=1}^{K-1} d_{\ell r}\alpha_\ell} \qquad r = 1, 2, \ldots, R. \tag{12}$$

- R is an unknown parameter representing the number of ratio models $f(\boldsymbol{\alpha}, \mathbf{b}_r)$ included in (11).

The value of integer R depends on the true CT–PM surface, and the determination of R is the key model-selection issue for CT–PM fitting and will be discussed in Section 5.5.4.

- Vectors $\{\mathbf{d}_r = (d_{1r}, d_{2r}, \ldots, d_{K-1,r}), r = 1, 2, \ldots, R\}$ are parameters estimated prior to and independent of the fitting of the CT–PM model (11), and they are treated as known values in (11) (see Section 5.4).

The CT–PM model (11) is almost the same as formula (10) although it is expected that R is much smaller than M, the number of stations in the system. Next, we further examine the geometry of the CT–PM surface in Section 5.4 and detail a strategy for fitting the nonlinear response surface model (11) in Section 5.5.

5.4. *Curvature of CT–PM surface*

In this section, we discuss the curvature (or the bending) of the CT–PM surface based on Jackson networks, which is the queueing model that motivates our regression model (11). The form of (10) for a Jackson network clearly suggests an additive model which is the sum of a number of ratio functions. For convenience of discussion, we rewrite (10) as follows

$$\begin{aligned} c_{1,x}(\boldsymbol{\alpha}) = \mu(\boldsymbol{\alpha}) &= e_0 + \sum_{j=1}^{M} \frac{g_{1j}(\boldsymbol{\alpha})}{g_{2j}(\boldsymbol{\alpha})} \\ &= e_0 + \sum_{j=1}^{M} \frac{\sum_{k=1}^{K-1} e_{kj}\alpha_k}{h_{0j} + \sum_{k=1}^{K-1} h_{kj}\alpha_k}, \end{aligned} \tag{13}$$

where all the coefficients depend on system parameters only. For each ratio function $g_{1j}(\boldsymbol{\alpha})/g_{2j}(\boldsymbol{\alpha})$, both the numerator $g_{1j}(\boldsymbol{\alpha})$ and denominator $g_{2j}(\boldsymbol{\alpha})$ are linear functions of $\boldsymbol{\alpha}$. Geometrically speaking, $g_{2j}(\boldsymbol{\alpha})$ is a one-dimensional projection of the variable vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{K-1})$ onto the system parameter vector $\mathbf{h}_j = (h_{1j}, h_{2j}, \ldots, h_{K-1,j})$. Since the numerator $g_{1j}(\boldsymbol{\alpha})$ is linear with respect to $\boldsymbol{\alpha}$, for response surface (13) curvature is only induced to the surface along the projections defined by $\mathbf{h}_j$ $(j = 1, 2, \ldots, M)$. Consider a simple case with $M = 1$: the curvature of $\mu(\boldsymbol{\alpha})$ is most pronounced along vector $\mathbf{h}_1$ whereas there is no curvature in directions orthogonal to $\mathbf{h}_1$.

Real manufacturing systems could be composed of a large number of stations (e.g. M could be on the scale of hundreds), which implies response curvature on M directions $\{\mathbf{h}_j, j = 1, 2, \ldots, M\}$. However, it is reasonable to believe that using a substantially smaller number of, say R, carefully chosen directions, (13) could be well-approximated by the sum of ratio functions along those R directions. Identifying the curvature directions of the CT–PM surface plays an important role in determining R, the number of ratio functions incorporated in the CT–PM model, and in assisting the nonlinear fitting of (11) as will be seen in Section 5.5. In this article, a method based on a quadratic polynomial approximation is used for the identification of curvature directions, namely the determination of the vectors $\{\mathbf{d}_r, r = 1, 2, \ldots, R\}$ in model (11).

Suppose we approximate the CT–PM surface $\mu(\boldsymbol{\alpha})$ by a full quadratic model:

$$\begin{aligned} \mu_{QC}(\boldsymbol{\alpha}) &= \beta_0 + \boldsymbol{\alpha}'\beta + \boldsymbol{\alpha}'\mathbf{B}\boldsymbol{\alpha} \\ &= \beta_0 + \sum_{k=1}^{K-1} \beta_k\alpha_k + \sum_{k=1}^{K-1} \beta_{kk}\alpha_k^2 + \sum_{k=1}^{K-2} \sum_{\ell=k+1}^{K-1} \beta_{k\ell}\alpha_k\alpha_\ell, \end{aligned} \tag{14}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{K-1})$, and $\mathbf{B}$ is the $(K-1) \times (K-1)$ symmetric matrix

$$\mathbf{B} = \begin{pmatrix} \beta_{11} & \beta_{12}/2 & \cdots & \beta_{1,K-1}/2 \\ \beta_{12}/2 & \beta_{22} & \cdots & \beta_{1,K-1}/2 \\ \vdots & & \ddots & \vdots \\ \beta_{K-1,1} & \beta_{K-1,2} & \cdots & \beta_{K-1,K-1} \end{pmatrix}. \tag{15}$$

It is our empirical experience that a quadratic model, although inadequate to accurately characterise the CT–PM surface, provides good enough response surface approximation to determine curvature directions. We perform the curvature analysis based on the full quadratic model (14) following the approach in Myers and Montgomery (2002).

The curvature of the surface depends on the second-order coefficient matrix $\mathbf{B}$. Let $\mathbf{P}'\mathbf{B}\mathbf{P} = \Lambda$ where Λ is a diagonal matrix containing the eigenvalues of $\mathbf{B}$ as main diagonal elements, and $\mathbf{P}$ is the $(K-1) \times (K-1)$ matrix whose columns are the normalised eigenvectors associated with the eigenvalues of $\mathbf{B}$. Let $\mathbf{d}_{\max}$ be the $(K-1) \times 1$ eigenvector of $\mathbf{B}$ associated with the maximum absolute eigenvalue $\lambda_{\max}$. Then $\mathbf{d}_{\max}$ represents the projection direction of $\boldsymbol{\alpha}$ along which the curvature of the surface is most marked. In model (11), $\mathbf{d}_{\max}$ will be assigned to $\mathbf{d}_1$, and sequentially $\mathbf{d}_2, \mathbf{d}_3, \ldots, \mathbf{d}_R$ will be determined in the process of fitting (11) in a progressive manner. The detailed fitting method is described in Section 5.5.5.

5.5. *Estimation of the CT–PM model*

Obtaining a well-estimated CT–PM surface is difficult. For convenience of the discussion, we rewrite the

CT–PM model introduced in Section 5.3:

$$\mu(\boldsymbol{\alpha}, \boldsymbol{\theta}_R) = \tau + \sum_{r=1}^{R} f(\boldsymbol{\alpha}, \mathbf{b}_r) = \tau + \sum_{r=1}^{R} \frac{\sum_{k=1}^{K-1} b_{kr}\alpha_k}{b_{0r} + \sum_{\ell=1}^{K-1} d_{\ell r}\alpha_\ell}. \tag{16}$$

It is worth noting that model (16) falls into the category of multi-variate rational functions. In the literature, theories and techniques have been developed for constructing rational function models of general form, and we refer the interested reader to Cuyt and Verdonk (1988, 1989), Lehmensiek and Meyer (2001) and Cuyt *et al.* (2006). Here, in light of the particular form of (16), we developed our own model fitting strategy which provides a good estimation for the CT–PM surface within a constant-BN subregion of PM. Detailed discussions are given in the remainder of this section.

5.5.1. *Identifiability of model parameters*

Given the structure of the CT–PM model (16), obviously there are difficulties associated with the identifiability of model parameters. There are two characteristics of the CT–PM model which cause the non-identifiability: (i) The first one is due to symmetries in model (16), which lead to multiple equivalent parameters. For example, if the ratio functions in (16) swap places, the parameters would be permutated. (ii) The second reason is the mutual dependence among the model parameters. For a ratio function $f(\boldsymbol{\alpha}, \mathbf{b}_r)$ included in the model, if the parameters in the numerator equal zero, the parameters in the denominator can take any values and are thus not unique.

Determining the curvature directions $\{\mathbf{d}_r, r = 1, 2, \ldots, R\}$ (with $\mathbf{d}_r$ being a normalised vector) independently of the nonlinear regression fitting essentially fixes problem (i), and alleviates the non-identifiability difficulty (ii). In addition, eliminating $\{\mathbf{d}_r, r = 1, 2, \ldots, R\}$ as unknowns in (16) helps with the ill-posedness problem in parameter optimisation. Here, the ill-posedness is due to the fact that the optimisation objective function is not continuous with respect to the model parameters to be optimised. The denominator of each ratio function $f(\boldsymbol{\alpha}, \mathbf{b}_r)$ incorporated in the model needs to be non-zero over the entire PM subregion Ω_ν being investigated. With $\mathbf{d}_r$ given, feasibility constraints on model parameters can be easily derived forcing the denominator to be at least ϵ (a small positive value) away from 0. Thus, for the nonlinear fitting, the unknown parameter $b_{0,r}$ has to satisfy either of the following constraints:

$$\begin{aligned}
&\text{Constr1}: b_{0r} + \mathbf{d}_r'\boldsymbol{\alpha} > \epsilon \quad \text{for any } \boldsymbol{\alpha} \in \Omega_\nu \Leftrightarrow b_{0r} \\
&\qquad > \epsilon - \min_{\boldsymbol{\alpha}\in\Omega_\nu}\{\mathbf{d}_r'\boldsymbol{\alpha}\}, \\
&\text{Constr2}: b_{0r} + \mathbf{d}_r'\boldsymbol{\alpha} < -\epsilon \quad \text{for any } \boldsymbol{\alpha} \in \Omega_\nu \Leftrightarrow b_{0r} \\
&\qquad < -\epsilon - \max_{\boldsymbol{\alpha}\in\Omega_\nu}\{\mathbf{d}_r'\boldsymbol{\alpha}\}.
\end{aligned} \tag{17}$$

Both $\boldsymbol{\alpha}$ and $\mathbf{d}_r$ are $(K-1) \times 1$ vectors and $\mathbf{d}_r'\boldsymbol{\alpha}$ denotes the inner product of them. Since Ω_ν is a simplex region as illustrated in Figure 4, the minimum and maximum of $\mathbf{d}_r'\boldsymbol{\alpha}$ can be easily obtained for given $\mathbf{d}_r$.

In our method, we are not seeking to obtain parameter estimates that may converge to the true values of the underlying CT–PM surface, which might be practically impossible to accomplish given the structure of the model. Rather, we develop a systematic fitting strategy which leads to a well-estimated response surface that adequately describes the CT–PM experiment data.

5.5.2. *Least-square fitting of the CT–PM model*

Suppose that for product k and fixed TH x, N data points $\{(\boldsymbol{\alpha}_1, y_1), (\boldsymbol{\alpha}_2, y_2), \ldots, (\boldsymbol{\alpha}_N, y_N)\}$ have been collected within a BN-constant region Ω_ν for the CT–PM fitting. Recall that in CT–PM modelling, the CT response y_i is actually an estimate from the fitted CT–TH curve $\widehat{c}_{k,\boldsymbol{\alpha}_i}(x)$, and as established in Section 4 we have $y_i = \widehat{c}_{k,x}(\boldsymbol{\alpha}_i) = \widehat{c}_{k,\boldsymbol{\alpha}_i}(x) \sim \text{Norm}(c_{k,\boldsymbol{\alpha}_i}(x), \sigma^2)$. We assume that

$$y_i = \mu(\boldsymbol{\alpha}_i, \boldsymbol{\theta}_R) + \varepsilon_i \quad i = 1, 2, \ldots, N,$$

where $\varepsilon_i \sim \text{Norm}(0, \sigma^2)$ and σ^2 is a user-specified precision target in the CT–TH curve fitting (Section 4). The independence of errors can be justified by appealing to the fact that $\{y_1, y_2, \ldots, y_N\}$ are CT estimates obtained from different CT–TH curves fitted from independent simulation data sets. With the assumptions satisfied for nonlinear regression, we use the least-square method to estimate model (16).

Note that when the nonlinear fitting is performed on (16), the number of ratio functions R is assumed given, and the number of data points N is assumed sufficiently large to allow for the fitting of model (16) including R ratio functions. The constraints on unknown parameters are provided in (17). When performing the constrained nonlinear regression, only one of the two alternative constraints can be imposed on the unknown parameter b_{0r}. Let `ActiveConstr` be a $R \times 1$ array defined as: `ActiveConstr`$(r) = 1$ if $b_{0,r}$ is subject to Constr1; `ActiveConstr`$(r) = 2$ otherwise. For a specified `ActiveConstr` array, the constrained nonlinear least-squares fitting can be formalised as:

$$\min_{\boldsymbol{\theta}_R} \sum_{i=1}^{N} [y_i - \mu(\boldsymbol{\alpha}_i, \boldsymbol{\theta}_R)]^2$$

$$\text{where } \boldsymbol{\theta}_R = (\tau, \mathbf{b}_1, \ldots, \mathbf{b}_R) \qquad (18)$$

$$\text{Subject to: } b_{0,r} \text{ satisfies constraint of type } \texttt{ActiveConstr}(r) \text{ for } r = 1, 2, \ldots, R.$$

Obtaining good starting values for the unknown parameters is important to ensure a successful nonlinear optimisation. As will be seen in Section 5.5.5, where we provide a complete description of the fitting process, a sequential model-fitting scheme is used to provide good starting values for (18).

5.5.3. *Statistical inference*

Under the assumption that the error terms are normally distributed with known variance σ^2, we do not need to resort to a large-sample justification for statistical inference on the CT–PM model. However, to apply conventional statistical procedures (e.g. Seber and Wild 2003, Chapter 5), we have to cope with the difficulties caused by parameter non-identification.

As pointed out in Section 5.5.1, for a ratio function

$$f(\boldsymbol{\alpha}, \mathbf{b}_r) = \frac{\sum_{k=1}^{K-1} b_{kr}\alpha_k}{b_{0r} + \sum_{\ell=1}^{K-1} d_{\ell r}\alpha_\ell} \qquad r = 1, 2, \ldots, R \qquad (19)$$

incorporated in model (16), if $b_{1,r} = b_{2,r} = \cdots = b_{K-1,r} = 0$ then b_{0r} is not identified. When such identification failure occurs, Phillips (1989) has shown that the distribution of the estimated parameters is no longer normal and instead belongs to the family of 'mixed Gaussian distribution'. Therefore, to use the conventional method to carry out parameter inference, we must ensure that a given CT–PM model contains no irrelevant ratio functions. In Section 5.5.4, model selection strategies will be described in detail which are expected to lead to a fully identified CT–PM model. If the model is indeed identified, valid statistical inference can be made based on an approximate normal distribution of the model parameters, which is detailed in Appendix A.1.

5.5.4. *Model selection*

In the CT–PM fitting, the major model selection issue is the determination of R, the number of ratio functions included in the CT–PM model for a given data set. The complexity of the model is characterised by the value of R, and a desirable value of R is the smallest possible integer that is able to generate a good approximation of the true response surface, taking into account the trade-off between estimation bias and variability due to estimation errors.

Suppose that the set of curvature directions in the current estimated model $\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_R)$ is $\mathbf{D}_R = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_R\}$ with $R \geq 1$ being the number of ratio functions included so far. The question is: can the approximation of the true CT–PM surface through $\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_R)$ be improved by adding one additional ratio function $f(\boldsymbol{\alpha}, \mathbf{b}_{R+1})$ to capture some neglected nonlinearities? If the answer is yes, the data can be explained more accurately by $\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_{R+1})$; otherwise, we declare $\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_R)$ as the best fitted model. In our fitting process (Section 5.5.5), we embed three model selection schemes which are which are performed following the sequence presented below.

5.5.4.1. *Quadratic model-based pre-selection*.

To incorporate an additional ratio function $f(\boldsymbol{\alpha}, \mathbf{b}_{R+1})$, we propose to first determine its corresponding curvature direction $\mathbf{d}_{R+1}$ by performing the curvature analysis described in Section 5.4. Specifically, we calculate the estimated residuals $\{e_i = y_i - \mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_R), i = 1, 2, \ldots, N\}$, and fit a full quadratic model to $\{e_i,\ i = 1, 2, \ldots, N\}$ to identify the direction $\mathbf{d}_{R+1}$ along which curvature is most pronounced.

There are two situations that may suggest that the additional function $f(\boldsymbol{\alpha}, \mathbf{b}_{R+1})$ should be omitted: (i) $\mathbf{d}_{R+1}$ turns out to be collinear to any of the normal vector directions already in $\mathbf{D}_R$ indicating that this additional ratio function $f(\boldsymbol{\alpha}, \mathbf{b}_{R+1})$ would not provide much information that is not present in $\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_R)$; and (ii) the absolute value of the eigenvalue λ_{R+1} corresponding to $\mathbf{d}_{R+1}$ is small suggesting little curvature along the direction of $\mathbf{d}_{R+1}$. We use the following criteria for collinearity and curvature checking:

$$\mathbf{d}'_{R+1}\mathbf{d}_r > \ell_0 \qquad r = 1, 2, \ldots, R$$

$$\frac{\lambda_{R+1}}{\max_{r=1,2,\ldots,R} \lambda_r} < c_0. \qquad (20)$$

If either condition of (20) is satisfied, we reject model $\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_{R+1})$. The parameters $\ell_0 \in (0, 1)$ and $c_0 \in (0, 1)$ are user-specified. In our experiments, we used $\ell_0 = 0.98$ and $c_0 = 0.02$. The selected values of ℓ_0 and c_0 are of limited importance and could be specified in a conservative manner (i.e. setting ℓ_0 close to 1 and c_0 close to 0) since this quadratic model-based checking serves only as a pre-screen for the hypothesis testing below.

5.5.4.2. *Hypothesis testing*.

For model selection in our nonlinear regression analysis, the two model specifications that we want to discriminate between are:

$$H_0 : y = \mu(\boldsymbol{\alpha}, \boldsymbol{\theta}_R) + \varepsilon, \tag{21}$$

$$H_1 : y = \mu(\boldsymbol{\alpha}, \boldsymbol{\theta}_R) + f(\boldsymbol{\alpha}, \mathbf{b}_{R+1}) + \varepsilon. \tag{22}$$

The specific method for testing whether an additional ratio function should be added to the CT–PM response model is given in Appendix A.2.

5.5.4.3. *Comparison with a reference surface.* Another difficulty involved with estimating a model of the form (16) is related to ill-posedness. In the nonlinear fitting (18), constraint (17) is imposed to bind the denominators of ratio functions $f(\boldsymbol{\alpha}, \mathbf{b}_r)$ $(r = 1, 2, \ldots, R)$ at least ϵ (e.g. 10^{-5}) away from 0 over Ω_ν, the entire constant–BN subregion. Although it rarely occurs in our CT–PM fitting process (described in Section 5.5.5), we could obtain estimated parameters (from solving (18)) that make the denominator of $f(\boldsymbol{\alpha}, \mathbf{b}_r)$ close to ε at some feasible PM points in Ω_ν. When this occurs, the resulting model may fit the design points but deviate dramatically from the true response surface.

Fortunately, this kind of nonlinear peculiarity can be easily circumvented by performing a safety check on the new expanded model $\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_{R+1})$ after it survives the hypothesis test described above. Specifically, we carry out a simple comparison between $\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_{R+1})$ and the quadratic model $\widehat{\mu}_{QC}(\boldsymbol{\alpha})$ defined in (14), which as mentioned in Section 5.4 is able to provide a reasonably good approximation, though not a sufficiently accurate representation, of the CT–PM surface. Given a data sample, both models (14) and (16) will be estimated, and the maximum relative deviation between these two fitted models over the Ω_ν region will be calculated. If it is unreasonably large, say,

$$\max_{\boldsymbol{\alpha} \in \Omega_\nu} \frac{\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_{R+1}) - \widehat{\mu}_{QC}(\boldsymbol{\alpha})}{\widehat{\mu}_{QC}(\boldsymbol{\alpha})} > 100\%, \tag{23}$$

we reject $\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_{R+1})$ and declare $\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_R)$ as the best fitted model.

5.5.5. *CT–PM fitting process*

Given a sample $\{(\boldsymbol{\alpha}_1, y_1), (\boldsymbol{\alpha}_2, y_2), \ldots, (\boldsymbol{\alpha}_N, y_N)\}$, we propose to fit the CT–PM model in a progressive manner. We start with the simplest model including only one ratio function, and then sequentially expand the model by incorporating one additional ratio function at a time. Throughout this progressive fitting process, strategies are embedded to provide good starting values for the unknown parameters in the nonlinear regression. Admittedly, the process described below is fairly complicated, but all the efforts are necessary to ensure a successful nonlinear fitting

Initially, we set $R = 0$ (the number of ratio functions included in the CT–PM model is 0); and the set of curvature directions $\mathbf{D} = \emptyset$.

Step 1: (1) Fit the full quadratic model (14) to the data $\{(\boldsymbol{\alpha}_i, y_i),\ i = 1, 2, \ldots, N\}$, and determine the curvature direction $\mathbf{d}_{\max}$ and the corresponding eigenvalue $\lambda_{\max}$. Denote the estimated quadratic model as $\widehat{\mu}_{QC}(\boldsymbol{\alpha})$.

(2) Set $R = 1$, $\mathbf{d}_R = \mathbf{d}_{\max}$, and $\mathbf{D} = \mathbf{D} \bigcup \mathbf{d}_R$.

(3) Fit $\mu(\boldsymbol{\alpha}, \boldsymbol{\theta}_1) = \tau + f(\boldsymbol{\alpha}, \mathbf{b}_1)$, the CT–PM model with the current value of $R = 1$ and curvature direction $\mathbf{d}_1$. Two different least squares problems (18) will be solved subjecting $b_{0,1}$ to Constr1 and Constr2, respectively. Compare the sum of squared error (SSE) resulting from the two nonlinear fittings. If SSE1 < SSE2, set `ActiveConstr`$(R) = 1$; otherwise, `ActiveConstr`$(R) = 2$. For determination of the starting values of the unknown parameters in this nonlinear fitting, see Appendix A.3.

Step 2: Check to see if there are a sufficient number of PM points for the fitting of the expanded model $\mu(\boldsymbol{\alpha}, \boldsymbol{\theta}_{R+1})$. If N is less than $1 + (R + 1) \times K$ (the number of unknown parameters in $\mu(\boldsymbol{\alpha}, \boldsymbol{\theta}_{R+1})$), stop; otherwise, continue.

(1) From the latest estimated CT–PM model $\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_R)$, compute the estimated residuals $\{(\boldsymbol{\alpha}_i, e_i),\ i = 1, 2, \ldots, N\}$, and perform quadratic linear regression on them to identify the curvature direction $\mathbf{d}_{\max}$ and the corresponding eigenvalue $\lambda_{\max}$. If conditions (20) are satisfied, then stop and declare $\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_R)$ as the best fitted model from the current data set; otherwise, continue.

(2) Perform the model selection analysis based on the hypothesis test described in Section 5.5.4. If the new expanded model $\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_{R+1})$ is rejected, then stop and declare $\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_R)$ as the best fitted model from the current data set. Otherwise, continue.

(3) Set $R = R + 1$, $\mathbf{d}_R = \mathbf{d}_{\max}$, and $\mathbf{D} = \mathbf{D} \bigcup \mathbf{d}_R$.

(4) Fit the partial model $E[e_i] = f(\boldsymbol{\alpha}, \mathbf{b}_R)$ to $\{(\boldsymbol{\alpha}_i, e_i),\ i = 1, 2, \ldots, N\}$ with parameter b_{0R} subject to constraint (17). As in Step 1(3), two different nonlinear fittings will be performed subjecting b_{0R} to two types of constraints. Again, if SSE1 < SSE2, set `ActiveConstr`(R)=1; otherwise, `ActiveConstr`(R)=2. Appendix A.3 provides a way to determine the starting values of the unknown parameters in this partial fitting.

Step 3: (1) Estimate the CT–PM model $\mu(\boldsymbol{\alpha}, \boldsymbol{\theta}_R)$ by solving (18) with the current values of R, curvature directions $\mathbf{D}$, and `ActiveConstr` array specified in the

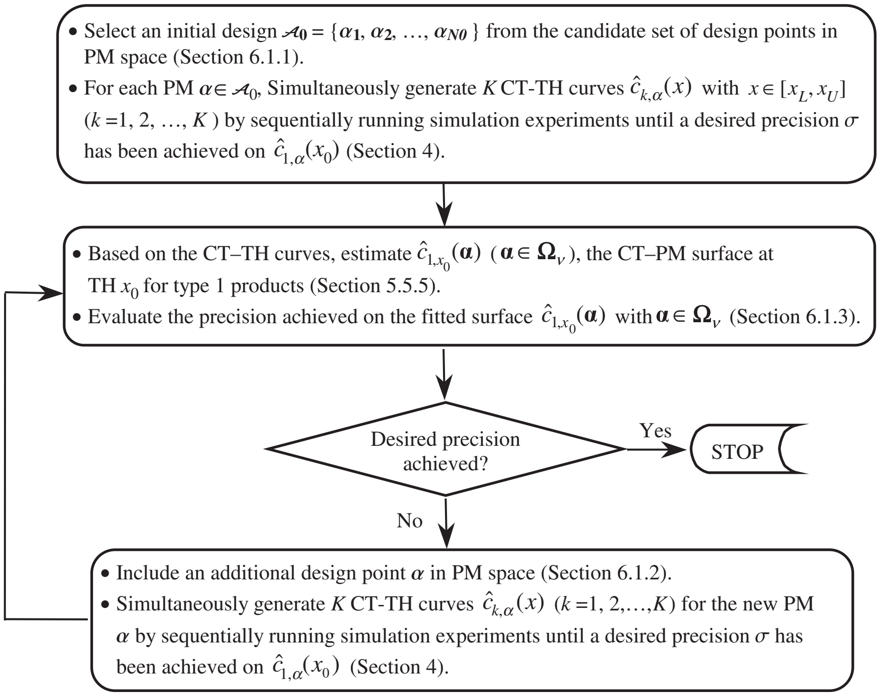

• Select an initial design $\mathcal{A}_0 = \{\alpha_1, \alpha_2, \ldots, \alpha_{N0}\}$ from the candidate set of design points in PM space (Section 6.1.1).
• For each PM $\alpha \in \mathcal{A}_0$, Simultaneously generate K CT-TH curves $\hat{c}_{k,\alpha}(x)$ with $x \in [x_L, x_U]$ (k =1, 2, …, K) by sequentially running simulation experiments until a desired precision σ has been achieved on $\hat{c}_{1,\alpha}(x_0)$ (Section 4).
• Based on the CT–TH curves, estimate $\hat{c}_{1,x_0}(\boldsymbol{\alpha})$ ($\boldsymbol{\alpha} \in \boldsymbol{\Omega}_\nu$), the CT–PM surface at TH x_0 for type 1 products (Section 5.5.5).
• Evaluate the precision achieved on the fitted surface $\hat{c}_{1,x_0}(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} \in \boldsymbol{\Omega}_\nu$ (Section 6.1.3).
Desired precision achieved?
Yes
STOP
No
• Include an additional design point α in PM space (Section 6.1.2).
• Simultaneously generate K CT-TH curves $\hat{c}_{k,\alpha}(x)$ (k =1, 2,…,K) for the new PM α by sequentially running simulation experiments until a desired precision σ has been achieved on $\hat{c}_{1,\alpha}(x_0)$ (Section 4).

Figure 5. Flow chart for the procedure.

previous steps. The latest estimates of the unknown parameters τ, $\mathbf{b}_1, \ldots, \mathbf{b}_R$ obtained from $\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_{R-1})$ and $f(\boldsymbol{\alpha}, \widehat{\mathbf{b}}_R)$ will be used as the starting values.

(2) Compare $\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_R)$ obtained in Step 3(1) with the quadratic model $\widehat{\mu}_{QC}(\boldsymbol{\alpha})$ fitted from Step 1(1), and check to see if the condition (23) is satisfied. If yes, then stop and declare $\mu(\boldsymbol{\alpha}, \boldsymbol{\theta}_{R-1})$ as the best fitted model from the current data set; delete $\mathbf{d}_R$ from $\mathbf{D}$ and set $R = R - 1$. Otherwise, go back to Step 2.

6. Procedure for estimating the CT–TH–PM response surface

This section is devoted to the construction of the experiment design and issues related to computational efficiency. To provide context, a high-level description of the procedure is provided in Figure 5 for estimating the CT–TH–PM surface within a constant-BN PM subregion Ω_ν (ν is a station that can serve as BN of the system). The procedure integrates the design of experiments, simulation and statistical modelling, and we will refer to this procedure as DSSM procedure in the remainder of this article.

The inputs/outputs of the procedure are given as follows.

Inputs: Simulation model of the system being investigated, a TH range of interest $[x_L, x_U]$, a TH level x_0 that the production is targeting (or is likely to operate at), the product type (assumed to be type 1) of particular interest to the user, absolute precision level σ for the estimated CT–TH curves (see Section 4 for the definition of σ), precision level $100\gamma\%$ which is defined as the relative error on $\widehat{c}_{1,x_0}(\boldsymbol{\alpha})(\boldsymbol{\alpha} \in \Omega_\nu)$, the CT estimates of particular interest (Section 6.1.3).

Outputs: A set of CT–TH curves $\{c_{k,\boldsymbol{\alpha}}(x), k = 1, 2, \ldots, K;\ \boldsymbol{\alpha} \in \mathcal{A}\}$, $x \in [x_L, x_U]$ where $\mathcal{A}$ is a collection of PM vectors. Based on these curves, for products of any type k, we can derive the CT estimates at any TH x and any PM $\boldsymbol{\alpha}$.

Assuming that $c_{1,x_0}(\boldsymbol{\alpha})$, $\boldsymbol{\alpha} \in \Omega_\nu$ (the expected CTs at TH x_0 for product 1) are of primary interest to the user, our goal is to obtain via simulation the CT estimates $\widehat{c}_{1,x_0}(\boldsymbol{\alpha})(\boldsymbol{\alpha} \in \Omega_\nu)$ with a specified relative precision $100\gamma\%$ while still well-estimating the CT–TH–PM surface for all $x_L \le x \le x_U$ and all types of products.

To generate the CT–TH–PM response surface, simulation experiments have to be carried out at a number of TH–PM combinations for data collection. Our approach is to first select the factor levels in the PM space, and then for each PM, apply the procedure proposed by Yang *et al.* (2007) to decide at what TH rates the simulation should be carried out. As illustrated in Figure 5, the experimentation is initiated with a pilot design $\mathcal{A}_0$ consisting of N_0 PMs. For each $\boldsymbol{\alpha} \in \mathcal{A}_0$, the CT–TH curves $\{c_{k,\boldsymbol{\alpha}}(x),\ k = 1,$

$2, \ldots, K$; $x \in [x_L, x_U]\}$ are generated by running simulation at different THs. Based on these curves, we can estimate CT for any product type at any TH and PM, and evaluate the relative error obtained for the CT estimates. As far as efficiency issues are concerned, we use the estimation quality of $\widehat{c}_{1,x_0}(\boldsymbol{\alpha})$, $\boldsymbol{\alpha} \in \Omega_\nu$ to drive the DSSM procedure: the design of simulation experiments aims at achieving good estimation of $\widehat{c}_{1,x_0}(\boldsymbol{\alpha})$, $\alpha \in \Omega_\nu$, and the sequential experimentation is continued until the desired precision is achieved. Technical details of the procedure are given in the remainder of this section.

6.1. *Experiment design*

When PM is fixed, Yang *et al.* (2007) developed an experiment design strategy which, tailored to the present context, targets efficient estimation of the CT–TH curve $\widehat{c}_{1,\boldsymbol{\alpha}}(x)$, $\boldsymbol{\alpha} \in \mathcal{A}$, and which sequentially allocates simulation runs to different levels of TH x until a desired precision σ has been achieved on $\widehat{c}_{1,\boldsymbol{\alpha}}(x_0)$. To estimate the complete CT–TH–PM surface, the design also includes $\mathcal{A}$, the collection of PM settings at which we fit the CT–PM surface. Since Yang *et al.* (2007) has already addressed the design issues for estimating the CT–TH curves, in this section we focus on the design in the PM space.

The fitting of the CT–PM surface is based on model (11) over a constant-BN region Ω_ν. Hence, we discuss the allocation of the PM design points within Ω_ν for the purpose of achieving a well-fitted CT–PM model. Each subregion Ω_ν is a simplex defined by linear constraints (7), so what we have is a K-component (K is the number of product types) mixture design problem within Ω_ν for the estimation of $c_{1,x_0}(\boldsymbol{\alpha})$, $\boldsymbol{\alpha} \in \Omega_\nu$, the CT measures of particular interest.

6.1.1. *Initial design*

For such constrained mixture designs, Myers and Montgomery (2002) recommend selecting design points from a candidate set, say $\mathcal{C} = \{\boldsymbol{\alpha}_1^*, \boldsymbol{\alpha}_2^*, \ldots, \boldsymbol{\alpha}_{\overline{N}}^*\}$, which provides good coverage of the feasible space. They claim that the set of candidate points to use for designing experiments should depend upon the form of the model the experimenter wishes to fit, and they recommended three different sets for linear, quadratic and cubic models based on their practical experience. Our model (11) does not fall into the category of polynomials with which they have experimented. However, our empirical experience with the CT–PM surface suggests that a quadratic model is able to provide an approximate fit for the response surface, although obviously inferior compared to (11). Thus, in our experiments, we chose to use the set Myers and Montgomery (2002) recommended for quadratic models, that is, the candidate set of design should include the following points of the simplex Ω_ν: extreme vertices, edge centres, constraint plane centroids, overall centroid and axial points.

Given the constraints (7) that define Ω_ν, we can use the CONVRT and CONAEV algorithms developed by Piepel (1988) to find the vertices, edge centres, and all other centroids of the simplex. In our procedure, the initial design points will be selected as a subset of these candidate points in $\mathcal{C}$. Let $\mathcal{A}_0$ denote the set of initial design points of size N_0 within the constant-BN region Ω_ν. We propose some additional requirements on the initial set of design points:

- To avoid extrapolation, $\mathcal{A}_0$ must include all the N_ν extreme vertices of Ω_ν.
- The number of initial design points N_0 should be sufficiently large to allow for the fitting of the full quadratic model given by (14), that is, $N_0 \geq 1 + (K-1)(K+2)/2$.
- In addition, we recommend $N_0 \geq 1 + 2 \times K$ so that the initial sample is large enough to estimate a CT–PM model (11) including two ratio functions.

Thus, $N_0 = \max\{1 + (K-1)(K+2)/2,\ 1 + 2 \times K\}$. The additional $N_0 - N_\nu$ non-vertex points are selected from $\mathcal{C}$ using a *maxmin* criterion which maximises the minimum distance between any two points.

6.1.2. *Design augmentation*

As illustrated in Figure 5, we initiate the experiments with a pilot design in the PM region as discussed in Section 6.1.1. The design points will then be sequentially added one at a time until the stopping rule is satisfied. In the DSSM procedure, the design augmentation is again guided by achieving good estimation for $\widehat{c}_{1,x_0}(\boldsymbol{\alpha})$, $\boldsymbol{\alpha} \in \Omega_\nu$.

We adopt the method in Seber and Wild (2003) to determine which PM point should be incorporated into the current design. Given that N design points have been allocated in the PM region, the $(N+1)^{st}$ additional design point $\boldsymbol{\alpha}_{N+1}$ is selected by minimising the determinant of the variance–covariance matrix of estimated unknown parameters (D-optimality criterion) in model $\widehat{c}_{1,x_0}(\boldsymbol{\alpha})$. For details see, Appendix A.4.

6.1.3. *Stopping criterion*

We allow the user to specify two precision levels, σ for the fitted CT–TH curves, and $\gamma\%$ for the fitted

CT–PM surface. As can be seen from the DSSM procedure (Figure 5), computation allocation is ultimately driven by $\gamma\%$, the desired relative error on $\widehat{c}_{1,x_0}(\boldsymbol{\alpha})$, $\boldsymbol{\alpha}\in\Omega_\nu$. The sequential simulation is continued until the following stopping criterion is satisfied:

$$\max_{\boldsymbol{\alpha}\in\Omega_\nu} \frac{2\times\sqrt{\widehat{\mathrm{Var}}[\widehat{c}_{1,x_0}(\boldsymbol{\alpha})]}}{\widehat{c}_{1,x_0}(\boldsymbol{\alpha})} < \gamma\%. \tag{24}$$

Statistical inference issues on the CT–PM modelling have been discussed in Section 5.5.3, where the notation $\mu(\boldsymbol{\alpha},\boldsymbol{\theta})=c_{k,x}(\boldsymbol{\alpha})$ is used to represent a CT–PM surface at a fixed TH for type k products. It has been established that $\widehat{c}_{1,x_0}(\boldsymbol{\alpha})$ is approximately normally distributed and the calculation of $\widehat{\mathrm{Var}}[\widehat{c}_{1,x_0}(\boldsymbol{\alpha})]$ is derived in Appendix A.1. The stopping rule (24) requires that the half-width for the confidence interval on $c_{1,x_0}(\boldsymbol{\alpha})$ is within $\gamma\%$ in a relative sense for any $\boldsymbol{\alpha}\in\Omega_\nu$. The left side of (24) can be easily obtained by evaluating $2\times\sqrt{\widehat{\mathrm{Var}}[\widehat{c}_{1,x_0}(\boldsymbol{\alpha})]}/\widehat{c}_{1,x_0}(\boldsymbol{\alpha})$ over a dense grid of $\boldsymbol{\alpha}$-values within Ω_ν.

However, as detailed in Section 5.5.2, the CT–PM surface is estimated based on the 'data points' $\{(\boldsymbol{\alpha}_1, y_1), (\boldsymbol{\alpha}_2, y_2), \ldots, (\boldsymbol{\alpha}_N, y_N)\}$ obtained from the CT–TH curve fitting, and we have $y_i \sim \mathrm{Norm}(c_{1,x_0}(\boldsymbol{\alpha}), \sigma^2)$. The prespecified value of σ definitely has an impact on the allocation of simulation effort in our experiments. If σ is large, it can be expected that more design points will be needed in the PM region in order to satisfy (24), and hence extra computation in design augmentation and model refitting will be required. Even worse, with highly variable sample the resulting fitted model may vary substantially on a sample-to-sample basis due to the complexity of the nonlinear model (11).

In light of these issues, we strongly recommend using a high precision level (small value for σ) and a minimum number of PM design points in the CT–TH–PM generation. The simple thumb rule adopted in our experiments is that σ is set in such a way that

$$\frac{2\sigma}{\min_{\boldsymbol{\alpha}\in\Omega}\widehat{c}_{1,x_0}(\boldsymbol{\alpha})} \approx \gamma\%/2. \tag{25}$$

The denominator in (25) could be replaced by a rough CT estimate within the feasible PM region.

7. Empirical evaluation

In this section, we demonstrate the effectiveness of our modelling method through empirical evaluation. Two different systems, an analytically tractable Jackson network and a real fab model, are explored. The proposed procedure is able to characterise a system processing multiple types of jobs with its CT–TH–PM response surface, however, in both cases considered here, we restrict the number of different types of jobs to be three for the sake of presentation: with PM $\boldsymbol{\alpha}$ being three-dimensional, the partition of the feasible PM region and the target response surface can be well-illustrated graphically.

In our experiments, we assume that the PM is not subject to additional linear constraints (6) imposed by practical considerations of production planning, which makes the feasible PM region a triangle in a three-dimension space as illustrated in Figure 3(a). Therefore, we are coping with a response surface over a larger input region, which is more challenging from the perspective of response surface modelling.

For both cases, user-specified parameters for the DSSM procedure are given as follows.

- TH range of interest is set as $[x_L, x_U]=[0.75, 0.85]$. This is motivated by the fact that semiconductor manufacturers typically run their facility in a utilisation range of [0.75, 0.85] (Hopp 2007).
- Products of type 1 and TH level of $x_0=0.8$ are of particular interest to production planning.
- The desired relative error is set at $\gamma\%=5\%$, which is to be achieved on $\widehat{c}_{1,x_0}(\boldsymbol{\alpha})$, $\boldsymbol{\alpha}\in\Omega_\nu$, the CT estimates of product 1 at $x_0=0.8$. The precision σ is set following (25).

7.1. *A Jackson network model*

We consider a three-product and three-station Jackson network, for which the true CT–TH–PM surface is known from queueing theory and hence provides a benchmark to evaluate the numerical results obtained from our method. The system configuration is specified in Table 1 following the notation defined in Section 2.1. Here we illustrate the application of the proposed method on this Jackson network and generate its CT–TH–PM surface via simulation.

7.1.1. *Preliminary queueing analysis of the model*

First, analytical queueing analysis is performed to partition the PM region into the constant-BN

Table 1. Three-station Jackson queueing model.

Station 1	Station 2	Station 3
$s_1=1$	$s_2=1$	$s_3=1$
$u_1=4$	$u_2=3$	$u_3=2.8$
$\delta_{11}=1$	$\delta_{12}=2$	$\delta_{13}=3$
$\delta_{21}=3$	$\delta_{22}=2$	$\delta_{23}=1$
$\delta_{31}=2$	$\delta_{32}=1$	$\delta_{33}=1$

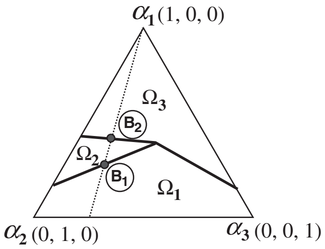


Figure 6. Division of the PM space and a constant-ratio PM path.

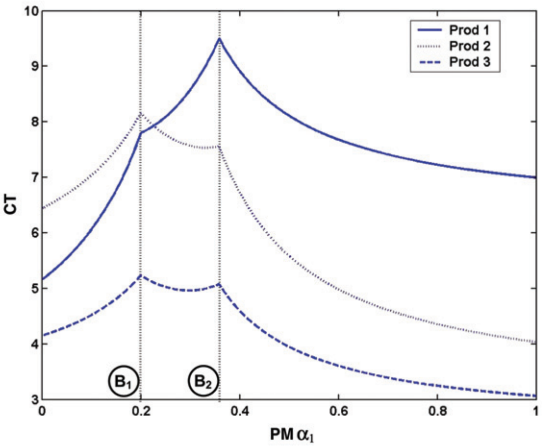


Figure 7. True CT–PM curves.

subregions. Figure 6 shows the division of PM space for this example. Each station can serve as a BN and the PM region is divided into three subregions with Ω_ν being dominated by the BN-station ν ($\nu = 1, 2, 3$). In addition, system capacities $u^*(\boldsymbol{\alpha})$ are derived from analytical analysis so that TH rates can be normalised to the scale of $[0, 1)$.

As a digression from applying the proposed method on the Jackson network for CT–TH–PM generation, we take some effort to examine the true CT–PM surface for this model. For an open Jackson network, the CT–PM surface (at a given x) is given by (8). If we fix $\alpha_2 : \alpha_3 = 3 : 1$, and vary α_1 from 0 to 1, we obtain a PM path as the dotted line in Figure 6. Along this path, we plot $c_{k,0.8}(\boldsymbol{\alpha})$, $k = 1, 2, 3$, the CT at TH $x = 0.8$, against α_1, and the resulting CT–PM curves are given in Figure 7. Obviously in Figure 7, the CT–PM curves are smooth and differentiable except at BN-shift points B_1 and B_2, which are also marked in Figure 6. We can change the ratio of $\alpha_2 : \alpha_3$, and plot the CT–PM curves similar to those obtained in Figure 7. This graphically demonstrates our conclusion in Section 5.3, which motivates us to model each subregion Ω_ν separately.

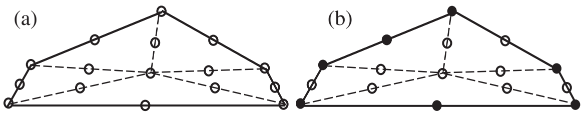


Figure 8. Initial design.

7.1.2. *Applying DSSM procedure on the Jackson network*

Returning to the generation of the CT–TH–PM surface, now that the PM region has been partitioned into three constant-BN subregions, the second step is to apply the DSSM procedure on each subregion and estimate a smooth CT–TH–PM surface within Ω_ν, $\nu = 1, 2, 3$. Considering region Ω_1 for example, we next follow the DSSM procedure and detail the process for generating the target response surface. The inputs to the procedure are set up as given in the beginning of Section 7. In addition, for this toy example, the CT roughly ranges from 5 to 10. Following the rule in (25), σ is set at $0.07 \approx CT_{\min} \times \gamma\%/4$ with $CT_{\min} = 5$ and $\gamma\% = 5\%$.

The initial design set $\mathcal{A}_0$ in PM space is determined as in Section 6.1. Specifically, the candidate set of design points is given in Figure 8(a), and the seven selected points comprising the initial design $\mathcal{A}_0$ are represented by black dots in Figure 8(b). For each $\boldsymbol{\alpha} \in \mathcal{A}_0$, simulation experiments are carried out at different TH levels and the CT–TH curves $\{c_{k,\boldsymbol{\alpha}}(x), k = 1, 2, 3\}$ are generated simultaneously for products of any type (as illustrated in Figure 1).

The simulation procedure is then driven by the estimation error on $\widehat{c}_{1,x_0}(\boldsymbol{\alpha})$, $\boldsymbol{\alpha} \in \Omega_1$, the CT estimates for product 1 at $x_0 = 0.8$. Based on the fitted curves $\{\widehat{c}_{1,\boldsymbol{\alpha}}(x), k = 1, 2, 3; \boldsymbol{\alpha} \in \mathcal{A}_0\}$, the CT–PM surface $\widehat{c}_{1,x_0}(\boldsymbol{\alpha})$, $\boldsymbol{\alpha} \in \Omega_1$ is estimated and then evaluated using the conventional statistical inference method (Section 5.5.3), and additional PM design points are added sequentially to the experiment design until the desired precision is achieved (Section 6.1.3).

Figure 9 shows the evolution of this sequential procedure. The horizontal axis represents the number of PM design points included in the experiments as the procedure progresses, and the vertical axis represents the corresponding estimation error for $\widehat{c}_{1,x_0}(\boldsymbol{\alpha})$, $\boldsymbol{\alpha} \in \Omega_1$. In Figure 9, each dot denotes

$$\max_{\boldsymbol{\alpha} \in \Omega_1} \frac{\widehat{c}_{1,x_0}(\boldsymbol{\alpha}) - c_{1,x_0}(\boldsymbol{\alpha})}{c_{1,x_0}(\boldsymbol{\alpha})}, \qquad (26)$$

the maximum relative deviation of the estimated $\widehat{c}_{1,x_0}(\boldsymbol{\alpha})$ from the true underlying surface evaluated at over 5000 grid points in Ω_1. Each star represents the maximum relative estimation error obtained from the statistical

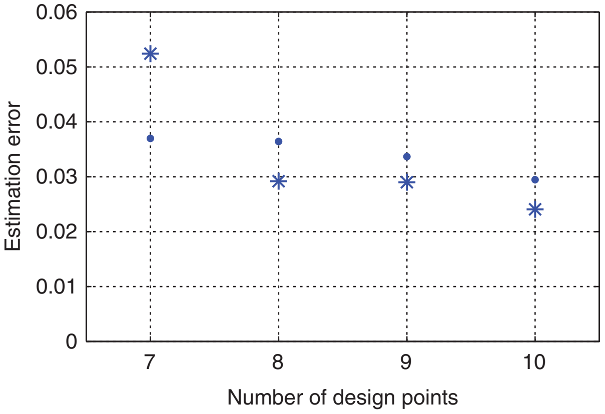


Figure 9. The evolution of the DSSM procedure applied on the Jackson network.

inference in the procedure, that is,

$$\max_{\boldsymbol{\alpha}\in\Omega_1} \frac{2 \times \sqrt{\widehat{\mathrm{Var}}[\widehat{c}_{1,x_0}(\boldsymbol{\alpha})]}}{\widehat{c}_{1,x_0}(\boldsymbol{\alpha})}. \tag{27}$$

The procedure is initiated with seven design points, and is terminated with eight points where the estimated error (27) satisfies the stopping criterion (24). We include two further steps with 9 and 10 design points to provide a more complete picture. As can be seen from Figure 9, the procedure could underestimate/overestimate the deviation of the estimated model from the true surface, but it does provide a rough idea of the size of the deviation.

The outputs of the DSSM procedure are the CT–TH curves $\{\widehat{c}_{k,\boldsymbol{\alpha}_i}(x), k = 1, 2, 3; \boldsymbol{\alpha}_i \in \mathcal{A}\}$. From these curves, for any $\boldsymbol{\alpha} \in \Omega_1$ and any $x \in [x_L, x_U]$, the CT estimates can be obtained. The estimation error for $\widehat{c}_{1,x_0}(\boldsymbol{\alpha})$ is controlled in the procedure to be within $\gamma\% = 5\%$ while other CT estimates are expected to be reasonably good. Here the fitted and true CT–PM surfaces for product 1 at $x_0 = 0.8$ are given in (28) and (29), respectively. The maximum deviation of the fitted model $\widehat{c}_{1,0.8}(\boldsymbol{\alpha})$ from the true model $c_{1,0.8}(\boldsymbol{\alpha})$ is 4% within the pre-specified precision 5%. Evidently the fitted model is able to approximate the true response surface to a desired precision although the estimated parameters are not the same as the parameters of the true surface model, as can be seen from comparing (28) and (29).

$$\begin{aligned}\widehat{c}_{1,0.8}(\boldsymbol{\alpha}) = 5.2562 &+ \frac{5.9886\alpha_1 - 1.1809\alpha_2}{0.3579 - 0.9889\alpha_1 + 0.1487\alpha_2} \\ &+ \frac{-2.0462\alpha_1 + 0.8385\alpha_2}{0.4498 - 0.9944\alpha_1 - 0.1054\alpha_2},\end{aligned} \tag{28}$$

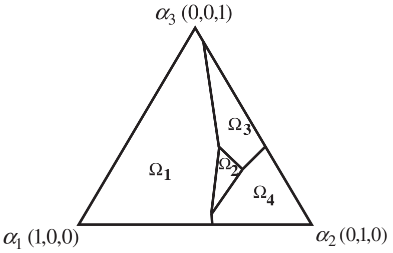


Figure 10. Division of PM region for the wafer fab model.

$$\begin{aligned}c_{1,0.8}(\boldsymbol{\alpha}) = 5.1780 &+ \frac{1.1053\alpha_1 + 0.3684\alpha_2}{0.4514 - 0.9995\alpha_1 - 0.0322\alpha_2} \\ &+ \frac{2.0794\alpha_1 - 0.4159\alpha_2}{0.2496 - 0.9568\alpha_1 - 0.2912\alpha_2}.\end{aligned} \tag{29}$$

From the outputs of the procedure, other CT estimates $c_{k,x}(\boldsymbol{\alpha})$, $\boldsymbol{\alpha} \in \Omega_1$ for $k = 1, 2, 3$ and $x \in [x_L, x_U]$ can also be obtained. Based on our experience, the estimation of $c_{1,0.85}(\boldsymbol{\alpha})$ appeared to be a most difficult case where the maximum deviation of the estimated model $\widehat{c}_{1,0.85}(\boldsymbol{\alpha})$ from the true surface is 6.7%. Similar results have been obtained for the other two subregions Ω_2 and Ω_3.

Jackson networks consisting of more than three workstations have also been considered in our experiments, and the DSSM procedure has demonstrated its ability to provide good CT–TH–PM estimation. It is worth mentioning that in our experience, increasing the size of the system hardly introduces additional difficulties for the procedure to accurately model the surface.

7.2. *A semiconductor manufacturing system*

In this section, we apply the proposed method to a semiconductor wafer fab simulation and characterise the manufacturing system by its CT–TH–PM surface. We consider a model describing a real wafer fab, provided by the Modeling and Analysis for Semiconductor Manufacturing Lab at Arizona State University (www.eas.asu.edu/~masmlab/).

The model is able to process three types of jobs: type 1, type 2 and type 3. Jobs of different types follow different process steps, and thus have different expected CTs.

7.2.1. *Analytical queueing analysis of the system*

We use the analytical engine provided by Factory Explorer to perform the capacity/BN analysis of the fab model. As shown in Figure 10, the PM region is divided into four constant-BN subregions with each one defined by a number of linear constraints.

For each PM vector $\boldsymbol{\alpha}$, Factory Explorer gives an analytical estimate for system capacity $u^*(\boldsymbol{\alpha})$, which allows us to normalise the TH levels to the scale of $[0, 1)$ and to transform the fitted response surface $\widehat{c}(x, \boldsymbol{\alpha})$ into $\widehat{c}(\lambda, \boldsymbol{\alpha})$ (Section 2.2).

7.2.2. *Applying the DSSM procedure on the wafer fab*

Again, the inputs to the procedure are set up as given in the beginning of Section 7. In this case, the CT measures roughly ranges from 300 to 450 h. Following the rule in (25), σ is set at $4 \approx CT_{\min} \times \gamma\%/4\,\mathrm{h}$ with $CT_{\min} = 300$ and $\gamma\% = 5\%$.

Since the true underlying surface for the fab is unknown, grid points evenly distributed over the PM region and the TH range were selected to check lack of fit in the fitted model at those locations. At these check points defined in terms of $(x, \boldsymbol{\alpha})$, substantial additional data were collected to obtain the 'nearly true' estimates for expected CTs. About 25 times as much computational effort as used in the DSSM procedure was invested in these check points. At each check point, simulation experiments were performed until the standard error of the expected CT estimate was essentially zero.

We present the estimation results for subregion Ω_1. In our experiments, a total of nine design points were employed, and the fitted model $\widehat{c}_{1,0.8}(\boldsymbol{\alpha})$ $(\boldsymbol{\alpha} \in \Omega_1)$ is given as

$$\widehat{c}_{1,0.8}(\boldsymbol{\alpha}) = 400.8383 + \frac{84.7793\alpha_1 - 34.5376\alpha_2}{-0.0736 - 0.9337\alpha_1 + 0.3581\alpha_2} + \frac{11.6227\alpha_1 - 29.7463\alpha_2}{-0.9403 + 0.3272\alpha_1 + 0.9449\alpha_2}. \qquad (30)$$

In Ω_1, 45 PM check points evenly distributed over the subregion are used, and the CT estimates $\widehat{c}_{1,0.8}(\boldsymbol{\alpha})$ obtained from (30) are compared with the 'nearly true' CTs. Figure 11 shows the histogram of the relative deviations of the CT estimates from their 'true' values. Among these 45 check points, almost all the relative deviations fall within the range of [−5%, 5%] with a few slight exceptions.

8. Summary

In a multi-product environment, mapping the expected CT as a function of TH and PM is difficult due to the complex nature of the underlying CT–TH–PM surface. In this article, a new metamodelling methodology, coupled with preliminary queueing analysis, is proposed for generating the CT–TH–PM response surface via sequential simulation experiments. It has been shown through experiments on Jackson networks and a real semiconductor manufacturing simulation that the proposed method provides good estimation for the CT–TH–PM surface characterising the system being investigated.

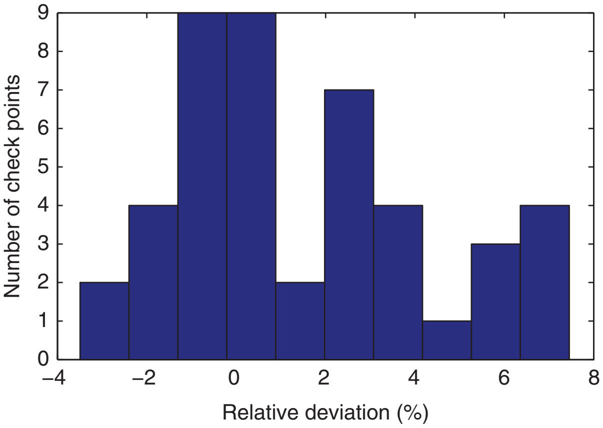
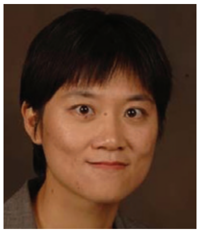


Figure 11. Relative deviation of $\widehat{c}_{1,0.8}(\boldsymbol{\alpha})$ from the 'true' CTs at 45 check points.

Notes on contributors



Feng Yang is an Assistant Professor in the Industrial and Management Systems Engineering Department at West Virginia University. Her research interests include simulation and metamodeling, design of experiments, and applied statistics.

Jingang Liu is a PhD student in the Industrial and Management Systems Engineering Department at West Virginia University. His research work has been focused on simulation and metamodeling.

Barry L. Nelson is the Charles Deering McCormick Professor of Teaching Excellence and Chair of the Department of Industrial Engineering and Management Sciences at Northwestern University. His research focus is on the design and analysis of computer simulation experiments on models of discrete-event, stochastic systems, including methodology for simulation optimization, variance reduction, metamodelling and multivariate input modelling. His application areas are manufacturing, services, finance and transportation. He has published numerous papers and two books, including *Discrete-Event System Simulation*, 5th edition (Prentice Hall, 2010) which has been adopted by over 60 universities. Nelson is a Fellow of INFORMS. In 2006, he received the Outstanding Simulation Publication Award from the INFORMS Simulation Society for his work on simulation optimization, and in 2009, he was awarded the Best Paper-Operations Award from *IIE Transactions*. He has also received the Northwestern University Alumni Association Excellence in Teaching Award, has twice been named McCormick Teacher of the Year, and has received the IIE Operations Research Division Award for Excellence in the Teaching of Operations Research.

Bruce E. Ankenman is an Associate Professor, Department of Industrial Engineering and Management Sciences, McCormick School of Engineering, Northwestern University. Prof. Ankenman's primary area of expertise is in Applied Statistics. For over 12 years he has taught courses and published a substantial number of journal articles in the areas of statistical quality control and design of industrial and simulation experiments. He has recently been involved with research on adapting design of experiment techniques for use in simulation experiments. These experiments are run on simulation models of manufacturing or other large scale systems and help engineers and managers to make tactical and strategic decisions based on realistic computer models. In the past 5 years, he has also dedicated himself to the teaching of the engineering design process. He directs Engineering Design and Communication (EDC) which is Northwestern's leading edge freshmen design course. In EDC, students design products that help people with disabilities to do everyday tasks like put on shoes, tie a pony tail, to participate in sports and recreational activities, such as water skiing and fishing. He is the director of undergraduate programs at the Segal Design Institute, an institute that offers a degree, a certificate and many upper level courses in engineering design. He served for 3 years as a Department Editor for *IIE Transactions* and is a current Associate Editor for *Naval Research Logistics*. Professor Ankenman is also the director of the Masters of Engineering Management Program.

Mustafa Tongarlak is a PhD candidate in the Department of Industrial and Management Sciences at Northwestern University. His research interests focus on the design and analysis of computer simulation experiments.

References

Cuyt, A. and Verdonk, B., 1988. Different techniques for the construction of multivariate rational interpolants. *In*: A. Cuyt, ed. *Nonlinear numerical methods and rational approximation*. Dordrecht, Netherlands: Springer, 167–190.

Cuyt, A. and Verdonk, B., 1989. Rational interpolation on general data sets in C^n. *In*: W.F. Ames and C. Brezinski, eds. *Numerical and applied mathematics*. Vol. 1, Basel, Switzerland: J.C. Baltzer AG, 415–429.

Cuyt, A., *et al.*, 2006. Adaptive multivariate rational data fitting with applications in electromagnetics. *IEEE Transactions on Microwave Theory and Techniques*, 54 (5), 2265–2274.

Gallant, A.R., 1987. *Nonlinear statistical models*. New York: John Wiley.

Hopp, W.J., 2007. *Supply chain science*. New York: McGraw-Hill/Irwin.

Hopp, W.J. and Spearman, M.L., 2001. *Factory physics: foundations of manufacturing management*. 2nd ed. New York, USA: Irwin.

Hopp, W.J., *et al.*, 2002. Using an optimized queueing network model to support wafer fab design. *IIE Transactions*, 34, 119–130.

Jacobs, J.H., 2004. *Performance quantification and simulation optimization of manufacturing Flow lines*. Thesis (PhD). Eindhoven University of Technology.

Kumar, S. and Kumar, P.R., 2001. Queueing network models in the design and analysis of semiconductor wafer fabs. *IEEE Transactions on Robotics and Automation*, 17 (5), 548–561.

Lehmensiek, R. and Meyer, P., 2001. Creating accurate multivariate rational interpolation models of microwave circuits by using efficient adaptive sampling to minimize the number of computational electromagnetic analyses. *IEEE Transactions on Microwave Theory and Techology*, 49 (8), 1419–1430.

Meng, G. and Heragu, S., 2004. Batch size modelling in a multi-item, discrete manufacturing system via an open queueing network. *IIE Transactions*, 36, 743–753.

Morrison, J.R. and Martin, D.P., 2007. Practical extensions to cycle time approximations for the G/G/m-queue with applications. *IEEE Transactions on Automation Science and Engineering*, 4 (4), 523–532.

Myers, R.H. and Montgomery, D.C., 2002. *Response surface methodology: process and product optimization using designed experiment*. 2nd ed. New York: Wiley-Interscience.

Phillips, P.C.B., 1989. Partially identified econometric models. *Econometric Theory*, 5, 181–240.

Piepel, G.F., 1988. Programs for generating extreme vertices and centroids of linearly constrained experimental regions. *Journal of Quality Technology*, 20, 125–139.

Schömig, A. and Fowler, J.W., 2000. Modelling semiconductor manufacturing operations. *In*: *Proceedings of the 9th ASIM simulation in production and logistics conference*. Berlin, Germany, 55–64.

Seber, G.A.F. and Wild, C.J., 2003. *Nonlinear regression*. New York: John Wiley & Sons.

Shantikumar, J.G., Ding, S., and Zhang, M.T., 2007. Queueing theory for semiconductor manufacturing systems: a survey and open problems. *IEEE Transactions on Automation Science and Engineering*, 4 (4), 513–522.

Wu, K., Mcginnis, L.F., and Zwart, B., 2007. Compatibility of queueing theory, manufacturing systems, and SEMI Standards. *In*: *Proceedings of the 3rd annual IEEE conference on automation science and engineering*. Scottsdale, AZ, USA, 22–25.

Yang, F., Ankenman, B.E., and Nelson, B.L., 2007. Efficient generation of cycle time-throughput curves through simulation and metamodelling. *Naval Research Logistics*, 54, 78–93.

Appendix A

A.1. *Statistical inference on the CT–PM model*

The notation used in Section 5 is inherited. Suppose that N 'data points' have been obtained for the estimation of CT–PM model (16) which includes R ratio functions. Using $\boldsymbol{\theta}$ to denote all the unknown parameters in the CT–PM model, we define the additional notation as follows:

- $\mathbf{f}(\widehat{\boldsymbol{\theta}}) = (\mu(\boldsymbol{\alpha}_1, \widehat{\boldsymbol{\theta}}), \mu(\boldsymbol{\alpha}_2, \widehat{\boldsymbol{\theta}}), \ldots, \mu(\boldsymbol{\alpha}_N, \widehat{\boldsymbol{\theta}}))'$ is a $N \times 1$ vector function of $\boldsymbol{\theta}$.
- $\mathbf{F}.(\widehat{\boldsymbol{\theta}}) = \partial \mathbf{f}(\widehat{\boldsymbol{\theta}})/\partial \boldsymbol{\theta}'$ is a $N \times (1 + K \times R)$ first-derivative matrix. Note that $1 + K \times R$ is the number of unknown parameters in the CT–PM model (16).
- $\mathbf{C}_N = \mathbf{F}.(\widehat{\boldsymbol{\theta}})'\mathbf{F}.(\widehat{\boldsymbol{\theta}})$.
- $\mathbf{f}_{\boldsymbol{\alpha}} = \partial\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}$ is a $(1 + K \times R) \times 1$ vector.

Assume that the error term in the CT–PM model is normally distributed with variance σ^2. Then the estimated parameters $\widehat{\boldsymbol{\theta}}$ is approximately normally distributed with variance:

$$\widehat{\mathrm{Var}}[\widehat{\boldsymbol{\theta}}] = \sigma^2 \mathbf{C}^{-1}. \tag{A1}$$

The CT estimator at PM $\boldsymbol{\alpha}$ from the fitted CT–PM model is approximately normally distributed as well, and its estimated variance is:

$$\widehat{\mathrm{Var}}[\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}})] = \sigma^2 \mathbf{f}'_{\boldsymbol{\alpha}} \mathbf{C}^{-1} \mathbf{f}_{\boldsymbol{\alpha}}. \tag{A2}$$

A.2. *Hypothesis test-based model selection*

For model selection in our nonlinear regression analysis, the two model specifications that we want to discriminate between are:

$$H_0 : y = \mu(\boldsymbol{\alpha}, \boldsymbol{\theta}_R) + \varepsilon, \tag{A3}$$

$$H_1 : y = \mu(\boldsymbol{\alpha}, \boldsymbol{\theta}_R) + f(\boldsymbol{\alpha}, \mathbf{b}_{R+1}) + \varepsilon. \tag{A4}$$

The hypothesis test above is equivalent to

$$H_0 : b_{1,R+1} = b_{2,R+1} = \cdots = b_{K-1,R+1} = 0, \tag{A5}$$

$$H_1 : b_{k,R+1} \neq 0 \text{ for some } k = 1, 2, \ldots, K-1. \tag{A6}$$

Given the hypothesis (A5) versus (A6), it would be natural to employ the test based on the approximate normality of the least-square parameter estimators $(\widehat{\boldsymbol{\theta}}_R, \widehat{\mathbf{b}}_{R+1})$. However, when H_1 is true, as explained in Section 5.5.3, we have the non-identifiability problem with $b_{0,R+1}$, and hence the distribution of estimated parameters is no longer approximately normal.

To circumvent these difficulties, we adopt the method proposed by Gallant (1987) which is based on the linear approximation of the additive nonlinear function $f(\boldsymbol{\alpha}, \mathbf{b}_{R+1})$. Instead of testing (A6), an approximate alternative hypothesis is employed:

$$\widetilde{H}_1 : y = \mu(\boldsymbol{\alpha}, \boldsymbol{\theta}_R) + \mathbf{z}' \times \mathbf{w} + \varepsilon, \tag{A7}$$

where $\mathbf{w}$ is a $(K-1) \times 1$ unknown (linear) parameter vector. The additional regressor $\mathbf{z}$ ($(K-1) \times 1$ vector) is defined as

$$z_k = \frac{\alpha_k}{\widetilde{b}_{0,R+1} + \sum_{\ell=1}^{K-1} d_{\ell,R+1}\alpha_\ell} \quad k = 1, 2, \ldots, K-1. \tag{A8}$$

Note that $\widetilde{b}_{0,R+1}$ in the denominator of (A8) is a predetermined value, and hence $\mathbf{z}$ is independent of any unknown parameters. Comparing (A4) and (A7), it can be seen that by eliminating $b_{0,R+1}$ as an unknown, $f(\boldsymbol{\alpha}, \mathbf{b}_{R+1})$ can be approximated by a linear function which is free of the non-identifiability problem.

To test $\widetilde{H}_1$ against H_0, standard procedures for testing parameter significance can be utilised. Plus, in CT–PM modelling, we benefit fully from the information on the error distribution (normal with known variance σ^2). From Appendix A.1, the estimated parameters $\boldsymbol{\theta} = (\widehat{\boldsymbol{\theta}}_R, \mathbf{w})$ in model (A7) is approximately normally distributed with variance $\sigma^2\mathbf{C}^{-1}$. Partitioning $\mathbf{C}^{-1}$ according to the partition of $\boldsymbol{\theta}$ into $(\widehat{\boldsymbol{\theta}}_R, \mathbf{w})$, and we have

$$\mathbf{C}^{-1} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}.$$

Under H_0, the statistic

$$\widehat{\chi}_w = \widehat{\mathbf{w}}'(C_{22})^{-1}\widehat{\mathbf{w}}/\sigma^2$$

is approximately distributed as the central chi-square distribution χ^2_{K-1} ($\widehat{\mathbf{w}}$ is a $(K-1) \times 1$ parameter vector). Using $\chi^2_{K-1}(a)$ (e.g. $a = 0.05$) to represent the $100(1-a)$th percentile of the χ^2_{K-1} distribution, the decision rule for testing $\widetilde{H}_1$ against H_0 is given as: H_0 is rejected if $\widehat{\chi}^2_w > \chi^2_{K-1}(a)$; otherwise, accept H_0.

Gallant (1987) provides theoretical guidance on how to choose the additional regressors $\mathbf{z}$ that maximise the power of the test when H_1 is true. In our case, the choice of $\mathbf{z}$ amounts to the determination of $\widetilde{b}_{0,R+1}$. In our model selection methods, to alleviate the computational burden, we simply use the estimated $b_{0,R+1}$ obtained from regressing the residuals $\{e_i = y_i - \mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_R), i = 1, 2, \ldots, N\}$ onto PM $\boldsymbol{\alpha}$ (see Appendix A.3 for specifics).

A.3. *Determining starting values for the CT–PM fitting*

As detailed in Section 5.5.5, in the fitting of the CT–PM surface, model (16) is sequentially expanded by including one additional ratio function at a time. Given the current estimated model $\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_R)$, the $(R+1)$th ratio function is written as

$$f(\boldsymbol{\alpha}, \mathbf{b}_{R+1}) = \frac{\sum_{k=1}^{K-1} b_{k,R+1}\alpha_k}{b_{0,R+1} + \sum_{\ell=1}^{K-1} d_{\ell,R+1}\alpha_\ell}. \tag{A9}$$

There are two issues to be addressed before we estimate the expanded model $\mu(\boldsymbol{\alpha}, \boldsymbol{\theta}_{R+1}) = \mu(\boldsymbol{\alpha}, \boldsymbol{\theta}_R) + f(\boldsymbol{\alpha}, \mathbf{b}_{R+1})$. (i) Would $\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_{R+1})$ represent a significant improvement over $\mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_R)$? (ii) If the answer to (i) is yes, then how to determine the starting values of $\mathbf{b}_{R+1}$ for the nonlinear fitting of $\mu(\boldsymbol{\alpha}, \boldsymbol{\theta}_{R+1})$?

Question (i) is a model selection issue which is discussed in Section 5.5.4, where, to perform the hypothesis test comparing $\mu(\boldsymbol{\alpha}, \boldsymbol{\theta}_R)$ and $\mu(\boldsymbol{\alpha}, \boldsymbol{\theta}_{R+1})$, a preliminary estimate of $b_{0,R+1}$ is needed to formulate the regressors $\mathbf{z}$ defined in (A8). Therefore, to answer both questions (i) and (ii), rough parameter estimates for $f(\boldsymbol{\alpha}, \mathbf{b}_{R+1})$ need to be provided before fitting $\mu(\boldsymbol{\alpha}, \boldsymbol{\theta}_{R+1})$ is performed, and we propose the following method to obtain such preliminary estimates.

First, calculate the residuals $\{e_i = y_i - \mu(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}_R), i = 1, 2, \ldots, N\}$, the variability of which is expected to be explained by $f(\boldsymbol{\alpha}, \mathbf{b}_{R+1})$:

$$e_i = f(\boldsymbol{\alpha}, \mathbf{b}_{R+1}) + \varepsilon_i = \frac{\sum_{k=1}^{K-1} b_{k,R+1}\alpha_k}{b_{0,R+1} + \sum_{\ell=1}^{K-1} d_{\ell,R+1}\alpha_\ell} + \varepsilon_i. \quad \text{(A10)}$$

Multiplying both sides of (A10) by the denominator of $f(\boldsymbol{\alpha}, \mathbf{b}_{R+1})$, we have

$$e_i^* = \sum_{k=1}^{K-1} b_{k,R+1}\alpha_k + b_{0,R+1}e_i + \varepsilon_i^*, \quad \text{(A11)}$$

where $e_i^* = e_i \times (\sum_{\ell=1}^{K-1} d_{\ell,R+1}\alpha_\ell)$ and $\varepsilon_i^* = \varepsilon_i \times (\sum_{\ell=1}^{K-1} d_{\ell,R+1}\alpha_\ell)$. Performing a linear regression analysis on (A10) will generate parameter preliminary estimates for $f(\boldsymbol{\alpha}, \mathbf{b}_{R+1})$.

A.4. *Design augmentation*

Following the notation defined in Appendix A.1, we write $\mathbf{f}_{\boldsymbol{\alpha}_{N+1}} = \partial f(\boldsymbol{\alpha}_{N+1}, \widehat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}$ which is a $(1 + K \times R) \times 1$ vector.

Given that N design points have been incorporated in the PM region, the location of the $(N+1)$th design point $\boldsymbol{\alpha}_{N+1}$ is determined by maximising

$$\begin{aligned} \mathbf{C}_{N+1} &= \left| \begin{pmatrix} \mathbf{F}.(\widehat{\boldsymbol{\theta}}) \\ \mathbf{f}'_{\boldsymbol{\alpha}_{N+1}} \end{pmatrix}' \begin{pmatrix} \mathbf{F}.(\widehat{\boldsymbol{\theta}}) \\ \mathbf{f}'_{\boldsymbol{\alpha}_{N+1}} \end{pmatrix} \right| \\ &= |\mathbf{C}_N|(1 + \mathbf{f}'_{\boldsymbol{\alpha}_{N+1}}\mathbf{C}_N^{-1}\mathbf{f}_{\boldsymbol{\alpha}_{N+1}}). \end{aligned} \quad \text{(A12)}$$

Since $\mathbf{C}_N$ does not involve $\boldsymbol{\alpha}_{N+1}$, the criterion (A12) reduces to maximising

$$\mathbf{f}'_{\boldsymbol{\alpha}_{N+1}}\mathbf{C}_N^{-1}\mathbf{f}_{\boldsymbol{\alpha}_{N+1}} \quad \text{(A13)}$$

with respect to $\boldsymbol{\alpha}_{N+1}$. The additional design point $\boldsymbol{\alpha}_{N+1}$ can be easily found by evaluating (A13) over a fine grid of $\boldsymbol{\alpha}$-values.