Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, eds.

A METAMODELING-BASED APPROACH FOR PRODUCTION PLANNING

Minqi Li
Feng Yang

West Virginia University
P.O. Box 6070
Morgantown, WV 26505, USA

Jie Xu

George Mason University
4400 University Drive
Fairfax, VA 22030, USA

ABSTRACT

In production planning, one of the major challenges for plan optimization lies in quantifying the dependence of the objective criterion (typically total cost) upon the decision variables that specify a release plan of jobs. Existing methods either fall short in capturing such a relationship, which involves non-stationary stochastic processes of a manufacturing system (e.g., the number of jobs over time), or require discrete-event simulation (DES) to evaluate the objective criterion for each candidate decision, which is time-consuming. To enable the accurate and precise estimation of the objective for any decision plan within a reasonable time, this work proposed a metamodeling-based approach. The metamodels take the form of difference equations, embody the high-fidelity of DES, and can be used to address "what if" questions in a timely manner. When embedded in the optimization of production planning, the metamodels can help to improve the quality and responsiveness of decision making.

1 INTRODUCTION

Production planning can be loosely defined as the problem of determining a release plan of jobs into the manufacturing system so that the system outputs over time meet the customer demand as closely as possible (Missbauer and Uzsoy 2010). Production planning is typically approached as an optimization problem: the decision variables quantify the number of jobs released for processing over the planning horizon, and the objective is usually to minimize the total cost (or maximize the total profit) associated with a release plan. The total cost typically includes the cost for holding work in process (WIP), the cost for holding finished good inventory, and the penalty cost for not satisfying customer demand in time.

The prerequisite for solving the optimization problem of production planing is the ability to quantify the dependence of the cost objective upon the release plan. Quantifying this relationship is very difficult, because the total cost depends on the various outputs of the manufacturing system over the planning horizon. These outputs include the WIP (i.e., the number of jobs in the system), the departures of completed jobs from the system, etc; and they are non-stationary stochastic processes triggered by a input release plan over time. Hence, establishing the cost vs. release plan relationship boils down to quantifying for any release plan the non-stationary stochastic output processes of the manufacturing system, which is challenging.

Depending on how a manufacturing system's input-output relationship is evaluated, the existing plan optimization work can be divided into three streams. In the optimization framework of most of the production planning models (e.g., Hackman and Leachman 1989; Johnson and Montgomery 1974; Leachman 1993), the queueing (or stochastic) nature of manufacturing systems was completely disregarded, and the system outputs were calculated in terms of the lead times considered as exogenous parameters independent of the release plan. Recognizing this obvious flaw, more recent research efforts were made to build in their optimization loop the ability to quantify/constrain the system outputs by utilizing queueing models, simulation-based statistical models (i.e., metamodels such as clearing function models), or discrete-event

978-1-4799-7486-3/14/$31.00 ©2014 IEEE

simulation (Hung and Leachman 1996; Byrne and Bakir 1999; Kim and Kim 2001; Byrne and Hossain 2005; Asmundsson et al. 2006; Asmundsson et al. 2009). However, these methods still fall short of accurately capturing the input-output dynamics of manufacturing systems: they either rely on, to some extent, the steady-state assumption of the system; or they employ discrete-event simulation (DES) in such a way that the input-output relationship implied by the DES is not fully represented in the plan optimization. The simulation optimization approach adopted by Liu et al. (2011) provides so far the most accurate way to evaluate the objective of plan optimization: for a release plan, DES was performed to evaluate the cost objective. However, running DES in an optimization loop to evaluate each candidate plan is very time-consuming (and it could take days, weeks, or even months depending on the complexity of the system), whereas the production decision typically needs to be made responsively when the management need arises.

In light of the discussions above, this work intends to develop a metamodeling-based approach for production planning. The metamodels (i.e., transfer function models in this case) are difference equations estimated from DES; they embody the high-fidelity of DES to describe the non-stationary behavior of stochastic systems, and at the same time, they allow for addressing "what if" questions in a much more timely manner. When embedded in an optimization framework for production planning, these metamodels can help to provide improved decision making within a reasonable amount of time.

The remainder of this paper is organized as follows. Section 2 provides an overview of the proposed method for production planning. Section 3 briefs the metamodeling of system dynamics. In Section 4, the metamodels-based objective evaluation and plan optimization are presented in details. Section 5 demonstrates the proposed method by applying it on a general single-station system. A summary is given in Section 6.

2 METHODOLOGY OVERVIEW

For the optimization of production planning over a given horizon $(0,H]$, the evaluation of the cost objective is centered on two steps, as shown in Figure 1.

(1) Metamodeling via offline DES: From DES for the system being investigated, a number of transfer function models (TFMs) will be estimated, which describes the system's dynamic and stochastic behavior. As in Figure 1, these TFMs functionally relate the system's output characteristics (e.g., the moments of WIP and job departures from the system) to its input, which is the release plan of jobs over $(0,H]$. For any release plan, the TFMs can predict the system's output characteristics within a second.

(2) Cost evaluation via Monte Carlo simulation: For any release plan, the output characteristics predicted by the TFMs can be employed to simulate over $(0,H]$ the job departures. The job departures and customer demands (outside of the system) are both discrete-variate time series, and interact with each other leading to demand fulfillment/backlog. Based on the simulated scenarios of job departures and demands, the expected total cost can be numerically estimated for any input plan. Such Monte Carlo simulation for job departures is performed based on given TFMs, only involves a series of basic computations (as will be seen later in Section 4.2.2), and hence takes orders of less time than DES. Further, unlike DES, the time for this Monte Carlo simulation is hardly affected by the system's complexity, since it is solely based on statistical models.

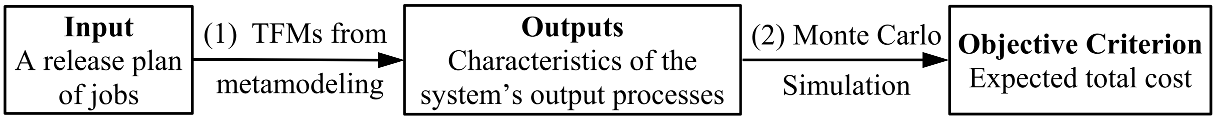

Figure 1: TFMs-based method for cost evaluation.

The procedure in Figure 1 allows for the evaluation of the expected total cost associated with any release plan, which serves as the basis for plan optimization. Figure 2 illustrates how the online plan optimization is performed given the TFMs fitted from extensive offline DES. Embedded in an optimization loop, the TFMs assist to find the optimum (or good) release plans.

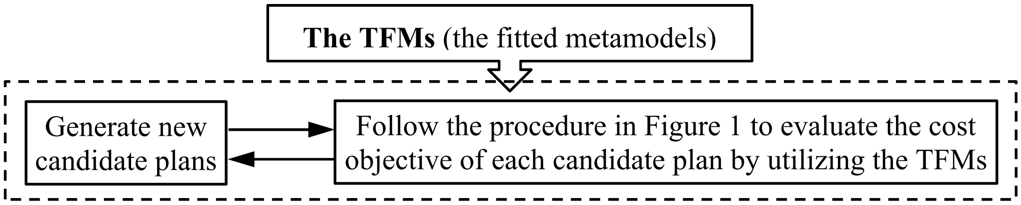
The TFMs (the fitted metamodels)
Generate new candidate plans
Follow the procedure in Figure 1 to evaluate the cost objective of each candidate plan by utilizing the TFMs

Figure 2: Optimization loop for production planning.

3 METAMODELING SYSTEM DYNAMICS

In this part, the metamodeling method developed by Yang and Liu (2012) are extended to obtain the TFMs able to capture the system's dynamic and stochastic behavior. The TFMs are difference equations estimated based on DES experiments.

For convenience of discussion, the following notations are defined:

- Δt: the time interval considered as the basic time unit 1.
- t: the time index measured in terms of the basic time unit.
- $A(t)$: the arrival process to the system which counts the number of arrivals during the time interval $(t-0.5, t+0.5]$. In the production planning context, $A(t)$ is also referred to as the release process for jobs.
- $Z(t)$: the departure process from the system which counts the number of finished jobs during the time interval $(t-0.5, t+0.5]$.
- $Q(t)$: the state process of the system which counts the number of work in process (WIP) in the system at time t.
- $x(t) = \mathrm{E}[A(t)]$: the first moment of the arrival process $A(t)$.
- $m_1(t) = \mathrm{E}[Q(t)]$: the first moment of the state process $Q(t)$.
- $d_i(t) = \mathrm{E}[Z^i(t)]$: the ith moment of the departure process $Z(t)$ $(i = 1, 2)$.
- $d_{1j}(t) = \mathrm{E}[Z(t)Z(t-j)]$: the first moment of $Z(t)Z(t-j)$ $(j = 1, 2, \ldots, J)$; J is the highest order of the time lag needed to describe the departures.
- $\mathbf{y}(t) = (m_1(t), d_1(t), d_2(t), d_{11}(t), d_{12}(t), \ldots, d_{1J}(t))$: the characteristics vector for the system's stochastic output processes.

As in Yang and Liu (2012), it is assumed that $x(t)$ completely characterizes the arrival process. Instead of being restricted to first-moment output measures as Yang and Liu (2012), this work includes in $\mathbf{y}(t)$ higher-moment metrics for the output processes. The TFMs can be written in general as

$$\mathbf{y}(t) = \mathbf{F}(x(t-1), ..., \mathbf{y}(t-1), \mathbf{y}(t-2), ...), \tag{1}$$

where $\mathbf{F}$ is a $(J+3) \times 1$ vector function of the same dimension size as $\mathbf{y}(t)$. Each component F_i specifies the dependence of the ith element of $\mathbf{y}(t)$ upon the given input $x(t)$ and the system's historical outputs. Each component model is assumed to take a polynomial form.

To fit the TFMs in (1), DES experiments are performed to collect data at a range of settings for the input $x(t)$. For a certain $x(t)$, N simulation replications are performed, and the paired input-output data are denoted as $(x(t), \widetilde{\mathbf{y}}(t))$ with

$$\widetilde{\mathbf{y}}(t) = \begin{pmatrix} \widetilde{m_1}(t) \\ \widetilde{d_1}(t) \\ \widetilde{d_2}(t) \\ \widetilde{d_{11}}(t) \\ \widetilde{d_{12}}(t) \\ \vdots \\ \widetilde{d_{1J}}(t) \end{pmatrix} = N^{-1} \sum_{i=1}^{N} \begin{pmatrix} Q_i(t) \\ Z_i(t) \\ Z_i(t)^2 \\ Z_i(t)Z_i(t-1) \\ Z_i(t)Z_i(t-2) \\ \vdots \\ Z_i(t)Z_i(t-J) \end{pmatrix}, \tag{2}$$

where the subscript i stands for the ith simulation replication. From the DES data, the TFMs can be estimated and denoted as

$$\widehat{\mathbf{y}}(t) = \widehat{\mathbf{F}}(x(t-1), ..., \widehat{\mathbf{y}}(t-1), \widehat{\mathbf{y}}(t-2), ...) \tag{3}$$

An example of estimated TFMs is given in Appendix A. The fitted TFMs (3) can be used to predict the system behavior. With the historical information $\{x(t), \mathbf{y}(t); t \leq 0\}$, which is typically known, the system's future performance $\{\widehat{\mathbf{y}}(t); t = 1, 2, \ldots, H\}$ can be obtained for an arbitrary input $\{x(t); t = 1, 2, \ldots, H\}$, through the recursive computation using (3).

4 METAMODEL-BASED PRODUCTION PLANNING

4.1 Problem Formulation of Production Planning

Following the existing literature (Missbauer and Uzsoy 2010), the optimization problem of production planning over a horizon $(0, H]$ is formulated as follows. Note that H is also given in terms of the number of time units Δt. The planning horizon is divided into P equal-length time buckets, and the pth bucket period is denoted as $[s_p, e_p]$ with $p = 1, 2, ..., P$. Within each bucket, the job release rate is assumed to be constant and equals to x_p. Thus, the release plan over $(0, H]$ can be specified by a $P \times 1$ vector

$$\mathbf{x} = (x_1, x_2, ..., x_P),$$

which is the vector of decision variables for plan optimization. The parameters, dependent variables, and objective function of the optimization problem are defined below.

Parameters

D_p the customer demand in the pth time bucket ($p = 1, 2, \ldots, P$). The demand series $\{D_1, D_2, \ldots, D_P\}$ is allowed to be a general discrete-variate time series, the model of which is obtained from the forecasting efforts outside the scope of this work. Herein, it is assumed that the demand in each time bucket is realized at the end of that period, even though our method allows for the demand to occur as frequently as every time unit, which is typically much shorter than a time bucket.

w_p the unit holding cost of the WIP per time unit in period p.

h_p the unit holding cost of finished products (or jobs) per time unit in period p.

b_p the unit backlog cost per time unit in period p.

g_0 the initial number of finished goods at the beginning of the first time period.

Dependent variables

- The stochastic output processes defined in Section 3 are rewritten as follows to stress that they are dependent on the decision vector $\mathbf{x}$:

$$Z(t, \mathbf{x}), Q(t, \mathbf{x}); \quad t = 1, 2, ..., H \tag{4}$$

- Accordingly, the characteristics of (4) are denoted as:

$$\begin{aligned} m_1(t,\mathbf{x}) &= \mathrm{E}[Q(t,\mathbf{x})] \\ d_1(t,\mathbf{x}) &= \mathrm{E}[Z(t,\mathbf{x})] \\ d_2(t,\mathbf{x}) &= \mathrm{E}[Z^2(t,\mathbf{x})] \\ d_{1j}(t,\mathbf{x}) &= \mathrm{E}[Z(t,\mathbf{x})Z(t-j,\mathbf{x})];\;\; j=1,2,\ldots,J \end{aligned}$$

- $W_p(\mathbf{x})$: the cumulative number of WIP during the pth time bucket, which equals to $\sum_{t=s_p}^{e_p} Q(t,\mathbf{x})$.
- $G_p(\mathbf{x})$: the cumulative number of finished products at the end of the pth time bucket after the demand D_p has been fulfilled. A positive $G_p(\mathbf{x})$ indicates the inventory of finished products, while a negative $G_p(\mathbf{x})$ means the demand of time period p can not be satisfied with $-G_p(\mathbf{x})$ being the unmet quantity. Given the initial inventory g_0, $G_p(\mathbf{x})$ can be calculated recursively as:

$$G_p(\mathbf{x}) = G_{p-1}(\mathbf{x}) + \sum_{t=s_p}^{e_p} Z(t,\mathbf{x}) - D_p, \quad p=1,2,...,P \tag{5}$$

The inventory level and the amount of backlog at the end of period p can be denoted as:

$$G_p^+(\mathbf{x}) = \max\{0, G_p(\mathbf{x})\},\; G_p^-(\mathbf{x}) = -\min\{0, G_p(\mathbf{x})\}, \tag{6}$$

respectively. The backlog at the end of period p is to be fulfilled at the end of the following period(s).
- $V_p(t,\mathbf{x})$: the inventory of the finished products at time t during the pth period $[s_p, e_p]$, which is calculated as:

$$V_1(t,\mathbf{x}) = g_0 + \sum_{\tau=s_1=0}^{t} Z(\tau,\mathbf{x}) \text{ for },\; t=s_1,s_1+1,...,e_1 \tag{7}$$

$$V_p(t,\mathbf{x}) = G_{p-1}^+(\mathbf{x}) + \sum_{\tau=s_p}^{t} Z(\tau,\mathbf{x}) \text{ for } p=2,3,...,P,\; t=s_p,s_p+1,...,e_p \tag{8}$$

- $I_p(\mathbf{x})$: the cumulative amount of inventory during the pth period, which is given as:

$$I_p(\mathbf{x}) = \sum_{t=s_p}^{e_p} V_p(t,\mathbf{x})$$

Objective Function

The total cost associated with a release plan $\mathbf{x}$ can be written as

$$L(\mathbf{x}) = \sum_{p=1}^{P} w_p W_p(\mathbf{x}) + \sum_{p=1}^{P} h_p I_p(\mathbf{x}) + \sum_{p=1}^{P} b_p(e_p - s_p + 1)G_p^-(\mathbf{x}),\; p=1,2,...,P \tag{9}$$

including three types of cost: (i) the WIP holding cost, (ii) the finished-good inventory cost, and (iii) the backlog cost. Clearly, $L(\mathbf{x})$ is a random variable which is dependent on $\mathbf{x}$. The objective of the production planning in this work is to minimize the expected total cost with respect to the release plan $\mathbf{x}$, which can be formulated as:

$$\begin{aligned} \min_{x_1,x_2,\ldots,x_P} \quad & \mathrm{E}[L(\mathbf{x})] \\ \text{Subject to} \quad & x_p \geq 0,\; p=1,2,...,P \end{aligned} \tag{10}$$

4.2 Cost Evaluation

To solve the optimization problem (10), we need to be able to evaluate the cost objective $\mathrm{E}[L(\mathbf{x})]$ for any release plan $\mathbf{x}$. The cost evaluation is performed by utilizing the fitted TFMs obtained for a manufacturing system of interest. The expected total cost $\mathrm{E}[L(\mathbf{x})]$ can be decomposed as the sum of three types of cost.

The expected WIP holding cost $\mathrm{E}[\sum_{p=1}^{P} w_p W_p(\mathbf{x})] = \sum_{p=1}^{P} w_p \mathrm{E}[W_p(\mathbf{x})]$ can be estimated as

$$\widehat{E}[\sum_{p=1}^{P} w_p W_p(\mathbf{x})] = \sum_{p=1}^{P}\sum_{t=s_p}^{e_p} w_p \widehat{m}_1(t,\mathbf{x}),\ p = 1,2,...P \tag{11}$$

For any $\mathbf{x}$, the TFMs allow for the prediction of $\{\widehat{m}_1(t,\mathbf{x}); t = 1,2,\ldots,H]$, and hence the expected cost for holding WIP during $(0,H]$.

The other two types of cost are: the expected inventory holding cost $\mathrm{E}[\sum_{p=1}^{P} h_p I_p(\mathbf{x})]$, and the expected backlog cost $\mathrm{E}[\sum_{p=1}^{P} b_p(e_p - s_p + 1)G_p^-(\mathbf{x})]$. However, these two types of expected cost cannot be estimated directly even with the TFMs, for two reasons: First, $I_p(\mathbf{x})$ and $G_p^-(\mathbf{x})$ depend on the departure process $Z(t,\mathbf{x})$ through the nonlinear functional relationships (5)-(8); Second, the departures $Z(t,\mathbf{x})$ and the demand series are both time series counts over $(0,H]$, and they interact with each other throughout the planning horizon. In light of this, for the evaluation of

$$\mathrm{E}[\sum_{p=1}^{P} h_p I_p(\mathbf{x})] + \mathrm{E}[\sum_{p=1}^{P} b_p(e_p - s_p + 1)G_p^-(\mathbf{x})],$$

the TFMs-based Monte Carlo simulation procedure is developed and outlined in Figure 3. For a release plan $\mathbf{x}$ over $(0,H]$, the fitted TFMs are able to predict the characteristics of the departures for the future planning horizon: $\{\widehat{d}_1(t,\mathbf{x}), \widehat{d}_2(t,\mathbf{x}), \widehat{d}_{1j}(t,\mathbf{x}); j = 1,2,\ldots,J; t = 1,2,\ldots,H\}$. These departure characteristics can then be used to fit the time-series count models which are able to capture the departures over the planning horizon. Using the fitted time-series models for departures, we can further carry out Monte Carlo simulation to simulate the departures over $(0,H]$, and hence evaluate the expected inventory and backlog cost via multiple simulation replications. The computation time requested by the procedure (Figure 3) is estimated as follows: On a computer with Inter(R) Core(TM) i7 CPU and 8G RAM, it takes about 0.2 second to complete the steps (a)-(c); one replication of the Monte Carlo simulation for the case in Section 5 takes about 0.05 second. As pointed out earlier, the computation time for the procedure remains almost the same regardless of the scale and complexity of the manufacturing system.

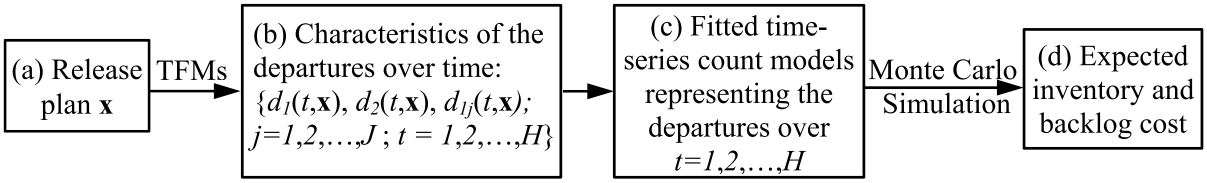
(a) Release plan x
TFMs
(b) Characteristics of the departures over time: {d1(t,x), d2(t,x), d1j(t,x); j=1,2,…,J ; t = 1,2,…,H}
(c) Fitted time-series count models representing the departures over t=1,2,…,H
Monte Carlo Simulation
(d) Expected inventory and backlog cost

Figure 3: Evaluation of the expected inventory and backlog cost

In the following two subsections, we discuss in details the methods involved in Step (b)-(d) of the procedure: How to fit the time-series count models from the departure characteristics provided by the TFMs (Section 4.2.1), and how to perform the Monte Carlo simulation based on the fitted time-series model (Section 4.2.2).

4.2.1 Time-Series Count Model

Given the TFMs-predicted departure characteristics $\{\widehat{d}_1(t,\mathbf{x}), \widehat{d}_2(t,\mathbf{x}), \widehat{d}_{1j}(t,\mathbf{x}); j = 1,2,\ldots,J; t = 1,2,\ldots,H\}$, we seek to estimate a time-series count model, which corresponds to the discrete-variate stochastic process

$\{\widehat{Z}(t,\mathbf{x});t=1,2,\ldots,H\}$. And $\{\widehat{Z}(t,\mathbf{x});t=1,2,\ldots,H\}$ mimics the real departures $\{Z(t,\mathbf{x});t=1,2,\ldots,H)\}$ in the sense that they share the same mean, variance, and lag-j autocorrelations: $\{\mu(t,\mathbf{x}),\sigma(t,\mathbf{x}),\rho_j(t,\mathbf{x});j=1,2,\ldots,J;t=1,2,\ldots,H\}$. Thus, the characteristics of $\{\widehat{Z}(t,\mathbf{x});t=1,2,\ldots,H)\}$ can be estimated from the TFMs as follows:

$$\begin{aligned}
\widehat{\mu}(t,\mathbf{x}) &= \widehat{d_1}(t,\mathbf{x}) && (12)\\
\widehat{\sigma}^2(t,\mathbf{x}) &= \widehat{d_2}(t,\mathbf{x})-(\widehat{d_1}(t,\mathbf{x}))^2 && (13)\\
\widehat{\rho}_j(t,\mathbf{x}) &= \frac{\widehat{d_{1j}}(t,\mathbf{x})-\widehat{d_1}(t,\mathbf{x})\widehat{d_1}(t-j,\mathbf{x})}{\widehat{\sigma}(t,\mathbf{x})\widehat{\sigma}(t-j,\mathbf{x})} \text{ for } j=1,2,...,J && (14)
\end{aligned}$$

with $t=1,2,\ldots,H$. The departure characteristics at $t=0,-1,\ldots,-J+1$ are the seed values needed to estimate $\widehat{\rho}_j(t,\mathbf{x})$ in (14) for $t=1,2,\ldots J$, and they can be obtained from historical data.

What time-series count models can describe $\{\widehat{Z}(t,\mathbf{x});t=1,2,\ldots,H)\}$ with the given characteristics (12)-(14)? In this work, the INAR model developed by Weiß (2013) is chosen to describe system departures, due to its ability to handle both under-dispersed and over-dispersed time series data and to accommodate higher-order autocorrelations. In addition, the parameters of an INAR model can be easily established from the characteristics of the underlying time-series data, as shown below.

For the convenience of presentation, take $J=3$ as an example. Then the parameters needed for the corresponding INAR model include: the autoregressive coefficients $\alpha(t,\mathbf{x})=(\alpha_1(t,\mathbf{x}),\alpha_2(t,\mathbf{x}),\alpha_3(t,\mathbf{x}))$; and $\mu_\varepsilon(t,\mathbf{x})$ and $\sigma_\varepsilon(t,\mathbf{x})$, the mean and variance of the innovation $\varepsilon(t,\mathbf{x})$.

Following Weiß (2013), the parameters $\alpha(t,\mathbf{x})$ are estimated as

$$\widehat{\alpha}(t,\mathbf{x})=\begin{pmatrix}\widehat{\alpha}_1(t,\mathbf{x})\\ \widehat{\alpha}_2(t,\mathbf{x})\\ \widehat{\alpha}_3(t,\mathbf{x})\end{pmatrix}=\mathbf{A}(t,\mathbf{x})^{-1}\mathbf{b}(t,\mathbf{x}),$$

where $\mathbf{A}(t,\mathbf{x})$ is a 3×3 matrix:

$$\begin{pmatrix}
\widehat{\sigma}^2(t-1,\mathbf{x}) & \widehat{\rho}_1(t-1,\mathbf{x})\widehat{\sigma}(t-1,\mathbf{x})\widehat{\sigma}(t-2,\mathbf{x}) & \widehat{\rho}_2(t-1,\mathbf{x})\widehat{\sigma}(t-1,\mathbf{x})\widehat{\sigma}(t-3,\mathbf{x})\\
\widehat{\rho}_1(t-1,\mathbf{x})\widehat{\sigma}(t-1,\mathbf{x})\widehat{\sigma}(t-2,\mathbf{x}) & \widehat{\sigma}^2(t-2,\mathbf{x}) & \widehat{\rho}_1(t-2,\mathbf{x})\widehat{\sigma}(t-2,\mathbf{x})\widehat{\sigma}(t-3,\mathbf{x})\\
\widehat{\rho}_2(t-1,\mathbf{x})\widehat{\sigma}(t-1,\mathbf{x})\widehat{\sigma}(t-3,\mathbf{x}) & \widehat{\rho}_1(t-2,\mathbf{x})\widehat{\sigma}(t-2,\mathbf{x})\widehat{\sigma}(t-3,\mathbf{x}) & \widehat{\sigma}^2(t-3,\mathbf{x})
\end{pmatrix},$$

and

$$\mathbf{b}(t,\mathbf{x})=\begin{pmatrix}\widehat{\rho}_1(t,\mathbf{x})\widehat{\sigma}(t,\mathbf{x})\widehat{\sigma}(t-1,\mathbf{x})\\ \widehat{\rho}_2(t,\mathbf{x})\widehat{\sigma}(t,\mathbf{x})\widehat{\sigma}(t-2,\mathbf{x})\\ \widehat{\rho}_3(t,\mathbf{x})\widehat{\sigma}(t,\mathbf{x})\widehat{\sigma}(t-3,\mathbf{x})\end{pmatrix}.$$

The mean and variance of the innovation are estimated as

$$\widehat{\mu}_\varepsilon(t,\mathbf{x})=\widehat{\mu}(t,\mathbf{x})-\sum_{i=1}^{J}\widehat{\alpha}_i(t,\mathbf{x})\widehat{\mu}(t-i,\mathbf{x})$$

$$\begin{aligned}
\widehat{\sigma}_\varepsilon^2(t,\mathbf{x})=\widehat{\sigma}^2(t,\mathbf{x})-\sum_{i=1}^{J}\widehat{\alpha}_i^2(t,\mathbf{x})\widehat{\sigma}^2(t-i,\mathbf{x})-\sum_{j=1}^{J}\widehat{\alpha}_j(t,\mathbf{x})(1-\widehat{\alpha}_j(t,\mathbf{x}))\widehat{\mu}(t-j,\mathbf{x})\\
-2\sum_{i=1}^{J}\sum_{j>i}^{J}\widehat{\alpha}_i(t,\mathbf{x})\widehat{\alpha}_j(t,\mathbf{x})\widehat{\rho}_{j-i}(t-i,\mathbf{x})\widehat{\sigma}(t-i,\mathbf{x})\widehat{\sigma}(t-j,\mathbf{x}).
\end{aligned}$$

Any release plan $\mathbf{x}$ in $(0,H]$ corresponds to a non-stationary departure process, which is described by an INAR model with fitted parameters $\{\widehat{\alpha}(t,\mathbf{x}),\widehat{\mu}_\varepsilon(t,\mathbf{x}),\widehat{\sigma}_\varepsilon^2(t,\mathbf{x});t=1,2,\ldots,H\}$.

4.2.2 Monte Carlo Simulation for Departures

The job departures from the system over $(0,H]$ can be simulated using the INAR model established above. The departure process described by the INAR model is denoted as $\{\widehat{Z}(t,\mathbf{x});t=1,2,\ldots,H\}$, and generated as follows.

$$\widehat{Z}(t,\mathbf{x})=\sum_{i=1}^{J}\alpha_i(t,\mathbf{x})\circ\widehat{Z}(t-i,\mathbf{x})+\varepsilon(t,\mathbf{x});\ \ t=1,2,\ldots,H \tag{15}$$

where $\{\varepsilon(t,\mathbf{x});t=1,2,\ldots,H\}$ are independent random variables with mean and standard deviation being $\{\mu_\varepsilon(t,\mathbf{x}),\sigma_\varepsilon(t,\mathbf{x});t=1,2,\ldots,H\}$. In this work, the generalized Poisson distribution (Consul 1989) is used to model the distribution of $\varepsilon(t,\mathbf{x})$ because of its flexible probability characteristics in modeling dispersions. The history $\{\widehat{Z}(t,\mathbf{x});t=0,-1,\ldots-J+1\}$ serves as the seed to initiate (15) and can be obtained from historical data.

In (15), the binomial thinning operator "$\circ$" is defined as (Weiß 2013):

$$\alpha\circ n=\sum_{i=1}^{n}B_i(\alpha)$$

where $\{B_i(\alpha)\}$ is a sequence of independent and identically distributed Bernoulli random variables independent of n, with $\mathrm{P}\{B_i=1\}=\alpha$.

For any $\mathbf{x}$ over $(0,H]$, a realization of the departure process can be simulated via (15), and used to evaluate the inventory and backlog cost associated with that realization; Based on multiple simulation replications (realizations), the expected cost $\mathrm{E}[\sum_{p=1}^{P}h_pI_p(\mathbf{x})]+\mathrm{E}[\sum_{p=1}^{P}b_p(e_p-s_p+1)G_p^-(\mathbf{x})]$ can be estimated.

As to the question of how many replications should be carried out for each candidate plan $\mathbf{x}$, a good answer is given in a simulation optimization algorithm (e.g., Xu et al. (2010)), which takes into account the uncertainty of cost estimates at candidate plans. In this work, which is a preliminary exploration, a fixed number for replications is used for all the candidate plans.

4.3 Optimization of Production Planning

To solve the optimization problem (10) for production planning, the cost evaluation ability rendered by the TFMs-based Monte Carlo simulation (Section 4.2) needs to be embedded in a optimization scheme, as shown in Figure 2. Simulation optimization algorithms such as the "Industrial Strength COMPASS" by Xu et al. (2010) are preferred, since the cost evaluation relies on Monte Carlo simulation. In this early investigation, the genetic algorithm (GA) provided by Matlab Optimization Toolbox is adopted for the optimization search with $\mathbf{x}$ being decision variables and the expected total cost in (9) the objective function. For the empirical case presented in Section 5, it takes 2.3 hours for the computer used (Inter(R) Core(TM) i7 CPU and 8G RAM) to complete the plan optimization via GA; in the GA search, about 800 candidate solutions were evaluated. As noted already, the evaluation time for each candidate is about the same regardless of the complexity of the system being investigated, since the evaluation is based on the metamodels.

5 EMPIRICAL RESULTS

A single-station system with five servers is used to demonstrate the metamodeling-based approach for production planing. The processing time of each server follows gamma distribution with a mean of 1 time unit and the coefficient of variation equals to 0.2. The job arrivals are assumed to be Poisson with the arrival rate varying over time.

The planning horizon is given as $(0,H]=(0,200]$, and is divided into $P=4$ time buckets with each one being 50 time units long. The demand over the planning horizon is allowed to be random, but is set as deterministic here. Three cases of demand $\{D_p;p=1,2,3,4\}$ considered are given in Table 1. The release plan over $(0,H]$ is specified by $\mathbf{x} = (x_1, x_2, x_3, x_4)$, with x_p being the arrival rate of jobs in the pth time bucket.

Following the notations in Section 4.1, the cost parameters are set as follows: $w_p = 1$, $h_p = 2$, and $b_p = 5$ for $p = 1,2,3,4$. These values represent the typical cost ratios for semiconductor manufacturing, as pointed out in Liu et al. (2011). The initial number of finished goods $g_0 = 0$.

5.1 Cost Estimation Results

The TFMs fitted from DES is given in (16)-(21), Appendix A. The TFMs allow for the estimation of the outputs $\{\widehat{m}_1(t,\mathbf{x}), \widehat{d}_1(t,\mathbf{x}), \widehat{d}_2(t,\mathbf{x}), \widehat{d}_{11}(t,\mathbf{x}), \widehat{d}_{12}(t,\mathbf{x}), \widehat{d}_{13}(t,\mathbf{x}); t = 1,2,\ldots,H\}$ for any $\mathbf{x}$. (Note that these outputs are written as $\widehat{m}_1(t), \widehat{d}_1(t), \widehat{d}_2(t), \widehat{d}_{11}(t), \widehat{d}_{12}(t)$, and $\widehat{d}_{13}(t)$ in (16)-(21)).

As explained in Section 4.2, the expected WIP cost was calculated following (11) with the predicted $\{\widehat{m}_1(t,\mathbf{x}); t = 1,2,\ldots,H\}$. The inventory and backlog cost was obtained following the procedure in Figure 3 via the TFMs-based Monte Carlo simulation. Specifically, the predicted departure characteristics $\{\widehat{d}_1(t,\mathbf{x}), \widehat{d}_2(t,\mathbf{x}), \widehat{d}_{11}(t,\mathbf{x}), \widehat{d}_{12}(t,\mathbf{x}), \widehat{d}_{13}(t,\mathbf{x}); t = 1,2,\ldots,H\}$ were used to fit the INAR time-series count models for the job departures. The fitted INAR was used to generate the simulated departures $\{\widehat{Z}(t,\mathbf{x}); t = 1,2,\ldots,H\}$, which was employed to estimate the expected inventory and backlog cost.

Table 1 evaluates the cost estimation provided by our metamodeling-based method for a range of demand cases and with different release plans. In Table 1, the expected total costs estimated from the metamodeling approach are given in the column marked "Metamodeling". DES were performed to provide the "true" cost estimates, which are given in the column "DES". The last column provides the relative deviations of the metamodeling-estimated costs from their "true" values, and evidences high accuracy of the expected costs estimated by our metamodeling approach.

Table 1: Evaluation of the estimated cost provided by the metamodeling-based method

		$E[L(\mathbf{x})]$		
Demand case	Release plan $\mathbf{x}$	Metamodeling	DES	Relative deviation
{ 200 200 200 200 }	{ 3.80 3.80 3.80 3.80 }	51785	51951	−0.32%
	{ 4.00 4.00 4.00 4.00 }	48131	48749	−1.27%
	{ 4.20 4.20 4.20 4.20 }	48625	48833	−0.43%
{ 150 175 200 225 }	{ 2.85 3.33 3.80 4.28 }	49090	49472	−0.77%
	{ 3.00 3.50 4.00 4.50 }	45576	45932	−0.78%
	{ 3.15 3.68 4.20 4.73 }	45329	45959	−1.39%
{ 200 225 150 175 }	{ 3.80 4.28 2.85 3.33 }	48593	49440	−1.71%
	{ 4.00 4.50 3.00 3.50 }	46127	46341	−0.45%
	{ 4.20 4.73 3.15 3.68 }	47451	46689	1.63%

5.2 Plan Optimization Results

For each demand case, the planning optimization problem (10) was solved as described in Section 4.3, with the GA being the optimization algorithm and the cost objective evaluated by the metamodeling method.

In Table 2, the solution $\mathbf{x}^*$ from the GA search and the corresponding cost objective $\mathrm{E}[L(\mathbf{x}^*)]$ are given. To evaluate the GA solutions, an exhaustive cost evaluation was performed on the 4096 evenly-spaced grid points in the space of $\mathbf{x}$; $\widetilde{\mathbf{x}}$ is the grid point with the minimum expected cost among all the grids, and $\mathrm{E}[L(\widetilde{\mathbf{x}})]$ the expected cost at $\widetilde{\mathbf{x}}$. As can be seen from Table 2, the GA leads to close or even slightly superior solutions, compared to the grids-based exhaustive search.

Table 2: Evaluation of the optimization solution for the production planning.

Demand case	$\mathbf{x}^*$	$\mathrm{E}[L(\mathbf{x}^*)]$	$\tilde{\mathbf{x}}$	$\mathrm{E}[L(\tilde{\mathbf{x}})]$
{ 200 200 200 200 }	{ 4.07 4.27 4.07 4.27 }	47799	{ 4.13 4.13 4.13 4.25 }	47776
{ 150 175 200 225 }	{ 3.05 3.56 4.05 4.71 }	44583	{ 3.13 3.63 4.13 4.75 }	44756
{ 200 225 150 175 }	{ 4.09 4.62 2.88 3.74 }	44691	{ 4.13 4.56 2.88 3.75 }	45120

6 SUMMARY

To support responsive decision making for production planning, a metamodeling-based Monte Carlo simulation approach is developed to relate the objective criterion (typically total cost) to the decision plan. The Monte Carlo simulation is built on the metamodels, which are fitted from offline discrete-event simulation data, and allows for high-quality and timely estimation of the objective criterion for any decision plan, which lays a solid foundation for plan optimization. Empirical results for a general single-station system indicates the potential of our metamodeling-based optimization for production planning. In the immediate next step, we will adapt and apply our methods on realistic manufacturing systems. It is worth mentioning that despite its complex simulation model, a real system's WIP and departures (output processes) are non-stationary time series as well. We have collected and examined the output time series from a realistic simulation model, and detected no features that cannot be captured by the proposed metamodels from a statistical perspective. Efforts will also be made to explore the use of simulation optimization algorithms for plan optimization.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation Grant No. CMMI-1068131.

A FITTED TFMS FOR THE EMPIRICAL EXAMPLE

The TFMs fitted from the discrete-event simulation for the single-station system is given in (16)–(21):

$$\begin{aligned}\widehat{m}_1(t) =& 0.0174+0.6615m_1(t-1)+0.0268d_{11}(t-1)+0.3227x(t-1)+0.0680m_1(t-1)x(t-1)\\ &+0.2446d(t-1)x(t-1)-0.2909d_1(t-1)^2+0.0351d(t-1)d_{13}(t-1)\\ &-0.0337d_{13}(t-1)x(t-1)-0.0469x(t-1)^2\end{aligned} \tag{16}$$

$$\begin{aligned}\widehat{d}_1(t) =& -0.1399+0.3548m_1(t-1)-0.0681d_1(t-1)+0.8122x(t-1)+0.0638d_1(t-1)^2\\ &-0.0050m_1(t-1)d_2(t-1)-0.0456m_1(t-1)x(t-1)-0.0018d_{13}(t-1)x(t-1)\\ &+0.0203x(t-1)^2+0.0006d_{11}(t-1)d_{13}(t-1)\end{aligned} \tag{17}$$

$$\begin{aligned}\widehat{d}_2(t) =& -6.1084+2.4646m_1(t-1)+5.7952d_1(t-1)-1.2925d_2(t-1)+2.9555x(t-1)\\ &-0.0152m_1(t-1)^2-0.1548m_1(t-1)d_1(t-1)-0.0151m_1(t-1)d_{12}(t-1)\\ &-0.1943m_1(t-1x(t-1)-5.2057d_1(t-1)^2+1.8838d_1(t-1)d_2(t-1)\\ &+1.3715d_1(t-1)x(t-1)-0.1496d_2(t-1)^2-0.2645d_2(t-1)x(t-1)\\ &+0.0021d_{11}(t-1)^2+0.0050d_{12}(t-1)^2\end{aligned} \tag{18}$$

$$\begin{aligned}\widehat{d}_{11}(t) = &-2.0489+1.6974m_1(t-1)+2.8051d_1(t-1)-0.5052d_2(t-1)-0.6258x(t-1)\\ &-0.0062m_1(t-1)^2-0.1045m_1(t-1)d_1(t-1)-0.0068m_1(t-1)d_{13}(t-1)\\ &-0.1739m_1(t-1)x(t-1)-2.6427d_1(t-1)^2+1.0055d_1(t-1)d_2(t-1)\\ &+1.4903d_1(t-1)x(t-1)-0.0803d_2(t-1)^2-0.1720d_2(t-1)x(t-1)\\ &+0.0008d_{11}(t-1)d_{12}(t-1)+0.0022d_{13}(t-1)^2\end{aligned} \tag{19}$$

$$\begin{aligned}\widehat{d}_{12}(t) = &-1.6035+2.1939d_1(t-1)-0.3141d_2(t-1)+0.8714d_{11}(t-1)+0.6223x(t-1)\\ &-0.1652m_1(t-1)d_2(t-1)-0.0059m_1(t-1)^2+0.8260m_1(t-1)d_1(t-1)\\ &-3.5571d_1(t-1)^2+0.0372m_1(t-1)d_{11}(t-1)-0.1333m_1(t-1)x(t-1)\\ &-0.0298d_2(t-1)d_{11}(t-1)+0.7539d_1(t-1)d_2(t-1)+1.7727d_1(t-1)x(t-1)\\ &-0.0903x(t-1)^2-0.0903x(t-1)^2-0.3766d_2(t-1)x(t-1)+0.1184d_{11}(t-1)x(t-1)\end{aligned} \tag{20}$$

$$\begin{aligned}\widehat{d}_{13}(t) = &-1.9206+2.4480d_1(t-1)-0.2471d_2(t-1)+0.8279d_{12}(t-1)+0.6543x(t-1)\\ &-0.1475m_1(t-1)d_2(t-1)-0.0088m_1(t-1)^2+0.8435m_1(t-1)d_1(t-1)\\ &-3.4641d_1(t-1)^2+0.0315m_1(t-1)d_{12}(t-1)-0.2012m_1(t-1)x(t-1)\\ &-0.0374d_2(t-1)d_{12}(t-1)+0.7305d_1(t-1)d_2(t-1)+1.4605d_1(t-1)x(t-1)\\ &-0.3568d_2(t-1)x(t-1)+0.0029d_{11}(t-1)d_{12}(t-1)+0.1543d_{12}(t-1)x(t-1)\end{aligned} \tag{21}$$

REFERENCES

Asmundsson, J. M., R. L. Rardin, and R. Uzsoy. 2006. “Tractable Nonlinear Production Planning Models for Semiconductor Wafer Fabrication Facilities”. *Nonlinear Production Planning Models for Semiconductor Wafer Fabrication Facilities* 19 (1): 95–111.

Asmundsson, J. M., R. L. Rardin, and R. Uzsoy. 2009. “Production Planning Models for Resources Subject to Congestion”. *Naval Research Logistics* 56:142–157.

Byrne, M. D., and M. A. Bakir. 1999. “Production Planning Using a Hybrid SimulationAnalytical Approach”. *International Journal of Production Economics* 59 (1): 305–311.

Byrne, M. D., and M. M. Hossain. 2005. “Production Planning: An Improved Hybrid Approach”. *International Journal of Production Economics* 93–94:225–229.

Consul, P. C. 1989. *Generalized Poisson Distributions: Properties and Applications*. New York and Basel: M. Dekker.

Hackman, S. T., and R. C. Leachman. 1989. “A General Framework for Modeling Production”. *Management Science* 35 (4): 478–495.

Hung, Y. F., and R. C. Leachman. 1996. “A Production Planning Methodology for Semiconductor Manufacturing Based on Iterative Simulation and Linear Programming Calculations”. *IEEE Transactions on Semiconductor Manufacturing* 9 (2): 257–269.

Johnson, L. A., and D. C. Montgomery. 1974. *Operations Research in Production Planning, Scheduling, and Inventory Control*. New York: Wiley.

Kim, B., and S. Kim. 2001. “Extended Model for a Hybrid Production Planning Approach”. *International Journal of Production Economics* 73 (2): 165–173.

Leachman, R. C. 1993. “Modeling Techniques for Automated Production Planning in the Semiconductor Industry”. In *Optimization in Industry: Mathematical Programming and Modelling Techniques in Practice*, edited by T. A. Ciriani and R. C. Leachman, 1–30. New York: Wiley.

Liu, J. G., C. H. Li, F. Yang, H. Wan, and R. Uzsoy. 2011. “Production Planning for Semiconductor Manufacturing via Simulation Optimization”. In *Proceedings of the Winter Simulation Conference*, edited by S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, 3617–3627. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Missbauer, H., and R. Uzsoy. 2010. “Optimization Formulations of Production Planning Problems”. In *Planning Production and Inventories in the Extended Enterprise: A State of the Art Handbook*, edited by K. G. Kempf, P. Keskinocak, and R. Uzsoy, 437 – 508. New York: Springer.

Weiß, C. H. 2013. “Integer-Valued Autoregressive Models for Counts Showing Underdispersion”. *Journal of Applied Statistics* 40 (9): 1931–1948.

Xu, J., B. L. Nelson, and J. Hong. 2010. “Industrial Strength COMPASS: A Comprehensive Algorithm and Software for Optimization via Simulation”. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 20:1–29.

Yang, F., and J. G. Liu. 2012. “Simulation-Based Transfer Function Modeling for Transient Analysis of General Queueing Systems”. *European Journal of Operational Research* 223:150–166.

AUTHOR BIOGRAPHIES

MINQI LI is a Ph.D. student in the Department of Industrial and Management Systems Engineering at West Virginia University. His research interests include modeling/simulation of semiconductor manufacturing, design of experiments, and applied statistics. His email address is mli6@mix.wvu.edu.

FENG YANG is an Associate Professor in the Department of Industrial and Management Systems Engineering at West Virginia University. She received her Ph.D. degree from the Department of Industrial Engineering and Management Sciences at Northwestern University. Her research interests include computer simulation, applied statistics, and applications in manufacturing and service industry. Her email address and web page are, respectively, Feng.Yang@mail.wvu.edu and http://www2.statler.wvu.edu/~yang/.

JIE XU is an Assistant Professor in the Department of Systems Engineering and Operations Research at George Mason University. He received his Ph.D. from the Department of Industrial Engineering and Management Sciences of Northwestern University. His research interests include Monte Carlo simulation, stochastic optimization, computational intelligence, and applications in risk management and aviation. He is a member of INFORMS and IEEE and is a senior member of IIE. His email address is jxu13@gmu.edu and his web page is http://mason.gmu.edu/~jxu13/.