2011

A Simulation-Based Transfer Function Modeling Approach for Responsive Production Planning

Jingang Liu
West Virginia University

A Simulation-Based Transfer Function Modeling Approach for Responsive Production Planning

Jingang Liu

Dissertation submitted to the
College of Engineering and Mineral Resources
at West Virginia University
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in
Industrial Engineering

Feng Yang, Ph. D., Chair
Wafik H. Iskander, Ph. D.
Majid Jaraiedi, Ph. D.
Bhaskaran Gopalakrishnan, Ph. D.
Robert M. Mnatsakanov, Ph. D.

Department of Industrial and Management Systems Engineering

Morgantown, West Virginia
2011

Keywords: Queueing System; Transient Analysis; Transfer Function; Production Planning; Simulation; Metamodeling

ABSTRACT

A Simulation-Based Transfer Function Modeling Approach for Responsive Production Planning

Jingang Liu

In this thesis, a novel statistical metamodeling method is proposed to study the input-output dynamic relations of general queueing system, especially when the system going through transient state. This metamodeling approach incorporates discrete even simulation, statistical inference, analytical queueing analysis to estimate a set of transfer function models (TFMs), which fully describe the system dynamic in terms of outputs that are interested : first and second moment of work-in-process (WIP) and first moment of departure rate.

Empirical queueing examples with non-Markov property are studied, including single station-single server system, single-station multi-server with failure system and multi-station multi-server with reentrant flow system and failure, following the proposed approach. TFMs for such systems are obtained. Predictions of the system dynamics is estimated from these TFMs under given system. Cross validation of the predicted outputs with simulation results show that the proposed approach is very accurate in describing the system evolution under both transient and steady states of general queueing system. Also, the proposed TFMs has been applied to the general Jackson-network models, cross validation result shows that the estimation results are promising as well.

The above proposed TFMs is integrated into a production planning model for a 5 station in-tandem system with reentrant flow and machine failures. The demands are independent and random with mean and distribution known. The production planning model estimates both mean and variance of the expected total cost based on TFMs. Genetic algorithm (GA)

is employed to find the Pareto Front of such a production planning model. Objectives from these Pareto Front points are compared with simulation results and proved the accuracy of the production planning model based on TFMs.

Acknowledgement

I am extremely grateful to my advisor Dr. Feng Yang. She has been a great support for me throughout my study. It is my great honor to work under her guidance. This dissertation could not have been written without her talent, encouragement, patient and financial support. I would also like to thank Dr. Iskander, Dr. Jaraiedi, Dr. Gopalakrishnan and Dr. Mnatsakanov for being my committee members and for their suggestions.

Final, I want to thank my parents for their constant support and unconditional love. I could not have been where I am without them.

Table of Contents

List of Tables

List of Figures

Glossary

ACV = Interarrival time coefficient of variation

ARMA = Autoregressive moving average

BNS = Bottleneck station

CF = Clear function

CT = Cycle time

DOE = Design of experiments

EDS = Estimation dataset

FG = Finished good

GA = Genetic algorithm

HUS = Heavily utilized station

LP = Linear programming

MOP = Multi-objective problem

MRP = Material requirements planning

MRP-C = Capacitated material requirements planning

MRP-II = Manufacturing resources planning

MTTF = Mean time to failure

MTTR = Mean time to repair

ODE = Ordinary differential equation

RR = Release rate

SCV = Service time coefficient of variation

TFM = Transfer function model

TH = Throughput

VDS = Validation dataset

WIP = Work in process

List of Symbols

$a(t)$: expected arrival rate at time t. $a(t) = \lim_{\delta \to 0^+} \delta^{-1} \mathrm{E}\{A(t, t+\delta) > 0\}$

$a_n(t)$: expected arrival rate at time t when there are n jobs in the system. $a_n(t) = \lim_{\delta \to 0^+} \delta^{-1} \Pr\{A(t, t+\delta) = 1, Q(t) = n\}$

$A(t)$: number of arrivals into the system within time interval $(0, t], t \in (-\infty, \infty)$

$A(u, v)$: number of arrivals into the system within time interval $(u, v], u < v \in (-\infty, \infty)$

b_p: backlogging cost per time unit in period p

B_p: number of backlogged products in period p

$\mathbf{d}$: mean demand vector

$d(t)$: expected departure rate at time t. $d(t) = \lim_{\delta \to 0^+} \delta^{-1} \mathrm{E}\{D(t, t+\delta) > 0\}$

$d_n(t)$: expected departure rate at time t when there are n jobs in the system. $d_n(t) = \lim_{\delta \to 0^+} \delta^{-1} \Pr\{D(t, t+\delta) = 1, Q(t) = n\}$

D_p: demand quantity in period p

$D(t)$: number of departures from the system within time interval $(0, t], t \in (-\infty, \infty)$

$D(u, v)$: number of departures from the system within time interval $(u, v], u < v \in (-\infty, \infty)$

e_p: ending time of planning period p

h_p: FG inventory hodling cost per time unit in period p

I_p: number of FG at the number of period p

$m_1(t)$: the expectation (first moment) of the number of jobs in the system at time t.

$m_1(t) = \mathrm{E}[Q(t)]$

$p_n(t)$: probability of n jobs in the system at time t

$Q(t)$: number of work in process at time t

R_p: number of jobs release into system in period p

$R(t)$: number of jobs release into system in time interval $[t, t + \Delta t)$

s_p: starting time of planning period p

u: service rate w_p: WIP holding cost per time unit in period p

W_p: cumulative WIP in period p

$x(t)$: the input variable of a queueing system. $x(t) = a(t)$

$X(t)$: simulation recorded input variable of a queueing system. $X(t) = \hat{a}(t)$

$\mathbf{y}(t)$: the output variable of a queueing system. $\mathbf{y}(t) = (m_1(t), m_2(t), d(t))$

$\mathbf{Y}(t)$: simulation recorded output variable of a queueing system. $\mathbf{Y}(t) = \hat{\mathbf{y}}(t)$

Z_p: number of product produced in period p

$Z(t)$: number of product produced in time interval $[t, t + \Delta t)$

Chapter 1

Introduction

The goal of any manufacturing facility is to utilize the available resources (materials, labor, machine, etc.) to satisfy demand in the most efficient manner. Production planning is an important part of production management [73] and is concerned with finding the best release plan of jobs so that the actual outputs over time satisfy the predetermined requirements [66] with the least cost. Production planning is usually performed by dividing the planning horizon into a number of discrete time buckets and determining the quantity of jobs to be released within each time bucket with the purpose of minimizing the total cost involved. This is clearly an optimization problem with the decision variables being the quantities of jobs released within the time bucket, and the objective being the total cost which usually includes inventory,

work in process (WIP) and penalty cost.

1.1 Research Challenges

Evaluating the cost associated with a production plan is the basis to minimize the total cost with respect to a plan, and is very challenging due to the variability (or uncertainty) involved in the manufacturing planning environment. As pointed out by Chen [12], two sources of variability are inherent in any production planning: fluctuation in product demand, and variation in operation work contents such as random processing times and machine failures. The presence of such variability leads to two major difficulties in production planning. (i) First, for any release plan, the total cost incurred over the planning horizon is a random variable, and hence, a thorough evaluation of a plan needs to be made based on the distribution of the associated cost, or the mean and variance of that distribution which are considered as the most important distribution characteristics. (ii) Second, it is difficult to quantify the dependence of the mean/variance of the total cost of the release plan. The cost depends on the output performance of the manufacturing system over the planning horizon: the quantity of products produced and the WIP (i.e., the number of jobs in the system). Due to the queueing effects resulting from the system variability, both these

outputs are stochastic processes whose distributions evolve over time depending on the input release of jobs and the initial status of the system (e.g., the WIP and inventory levels at the beginning of the planning period). So far, no analytical methods can accurately characterize the relationships between the release plan and these stochastic outputs, which is not only nonlinear but also time-dependent. Simulation appears to be the only approach that is able to provide good estimations based on numerical instances, but is computationally intensive and thus does not allow for real-time "what-if" analysis.

The existing production planning work has not adequately addressed these two difficulties, i.e., the risk assessment of a production plan and the quantification of the input-output relationships for manufacturing systems. To the best of our knowledge, no work in the literature has considered the risk factor (or variance of total cost) in the production plan optimization. The majority of the work in production planning, including the widely used Material Requirements Planning (MRP) procedure and most mathematical programming models (see 2.1.1), completely disregards the stochastic nature of the problem and does not even consider the dependence of a system's outputs upon the input release (or the system workload). In these work, the output performances (e.g., the WIP and cycle time) are treated as exogenous constant parameters independent of the release decision. Ignoring this fundamental interde-

pendency may well lead to inferior production plans [10,71]. Recently, research effort has been devoted to addressing this input-output relationship in production planning models , and the basic idea of this new stream of work is to integrate deterministic mathematical programming with computer simulation. Some researchers (see 2.1.2 for detail) embedded iterative simulation in a mathematical programming model to evaluate the effects of the release decisions upon the performance of a manufacturing system. Other researchers developed the clearing function (CF) methods, in which the CFs are first estimated from simulation and then incorporated in the optimization model as constraints quantifying the dependence of expected WIP upon release rates of jobs. In the linear/nonlinear programming models of these more recent work, simulation is used to refine or estimate static constraints, which only provides a snapshot of the time-dependent input-output relationships. In addition, there is no guarantee of convergence in the optimization search of these hybrid methods.

1.2 Research Objectives

In light of this, this research aims at addressing the difficulties involved in responsive production planning.

The first objective of this research is to develop a statistical methodology to de-

quately characterize a general queueing system's input-output dynamic, which will be represented by a number of transfer function models (TFMs). For a given system of interest, this method assumes the availability of its simulation model, and fully utilizes offline simulation time, which is typically plentiful in practice, to estimate a number of TFMs. The TFMs are difference equations quantifying the nonlinear time-dependent relationship between the release rate of jobs and the system output performances, including the throughput of products and the first two moments of work in process (WIP). This simulation-based transfer function modeling approach combines the advantages of both existing transient analysis methods, i.e., computer simulation and pure analytical methods, while avoiding their shortcomings. (i) The TFMs embody the high fidelity of simulation to real systems since they are estimated from detailed simulation data. (ii) The TFMs are difference equations, the discrete-time counterpart of the ODEs provided by an analytical approach; supposing that a certain input is fed to the system under given initial conditions, the TFMs can be used to recursively compute the system's future output performance in a *timely* manner. Hence, the TFMs resulting from the proposed method are able to accurately describe the transient behavior of realistic systems and, at the same time, they allow for prompt "what-if" analysis.

The second objective of this work is to integrate the proposed TFMs into the

optimization of production planning. The TFMs allow for the evaluation of the mean and variance of the total cost associated with a release plan, which serves as the solid basis for cost minimization with respect to the production plan. A multi-objective genetic algorithm (GA) was adopted to search for the Pareto optimum [18] plans that are the best in the sense of both the mean and variance of the total cost. The proposed work is the first production planning method that is able to take into account both the mean and variance of the total cost. This allows decision makers to evaluate the expected cost as well as the risk involved in each production plan.

1.3 Overview of the TFMs Methodology

The proposed TFM method is three fold.

- *Queueing analysis* (Section 3.2): Queueing analysis is performed under fairly general assumptions. Such a theoretical analysis, although inadequate to address the time-dependent behavior of realistic systems, sheds lights on the functional forms for the TFMs.

- *Data collection via offline simulation* (Section 3.3): Under selected input processes, simulations will be run to obtain paired input-output time series. The author would like to emphasize that the simulation is carried out offline in

advance of the need to make a decision.

- *Transfer function modeling* (Section 3.4): From the simulation data, statistical methods to obtain the parsimonious TFMs (3.1) will be developed, which are adequate to capture the system's dynamic behavior.

1.4 Contribution of the Research

It is worth mentioning that the simulation-based transfer function modeling falls into the category of metamodeling (Chapter 18 of [34]), which refers to the techniques that utilize simulation to generate mathematical approximations quantifying the relationships implied by the simulation. This work, to the best of our knowledge, is the first attempt to develop a metamodel that takes the form of difference equations, and applies it to the context where metamodeling can realize the maximum potential. Such a context is articulated in [1] as follows: the time to exercise the simulation model in advance of the decision making is relatively plentiful, whereas the decision-making or decision-maker time is relatively scarce or expensive. The responsive production planning mentioned above represents one of such contexts: simulation models for the manufacturing system can be developed and kept running for weeks (or even months) as soon as the system configuration has been established; while in case of production

disruptions, a decision needs to be made quickly–as soon as possible–regarding how to adjust the production plan for that system. The metamodel, i.e., the TFMs, fully utilizes the plentiful offline simulation time and allows for responsive decision making in time of urgency.

Utilizing the ability of the TFMs to timely and accurately predict a system's future evolution, our production planning model is able to find the Pareto optimum plans that have the best performance in terms of not only the mean but also the variance of the total cost. This research work provides the first production planning method that is able to take into account both the mean and variance of the total cost in real-time plan optimization.

1.5 Organization of the Dissertation

The remainder of this dissertation is organized as follows. Chapter 2 provides a literature review of the existing transient analysis and production planning methods. Chapter 3 is devoted to the simulation-based transfer function modeling approach for the transient analysis of general queueing systems. In Chapter 4, the input-output dynamics described by the TFMs are utilized to perform the optimization of production planning. Chapter 5 finishes this thesis with conclusions and discussions.

Chapter 2

Literature Review

Production planning refers to the problem of determining the proper rates of job release into a manufacturing system to achieve the best overall performance. The input decision here is the release rate (RR) of jobs, and the performance metrics of interest include the throughput (TH), work in process (WIP), and cycle time (CT) of jobs. Loosely speaking, WIP and CT play equivalent roles in terms of characterizing the system's output performance. This is implied by the classic Little's law E[WIP] =TH×E[CT] assuming steady state: given TH, the expected CT can be calculated from the expected WIP, and vice versa. The transient Little's law developed by [5] gives the same indication for transient systems. In the proposed work, TH and WIP are chosen as the outputs to be modeled by the TFMs, and our major task is

to quantify this input-output (RR vs. multiple performance metrics) relationship.

For real manufacturing systems, both the RR input and the output indicators tend to vary with time over a finite planning horizon. Hence, a manufacturing system may rarely be operated in its steady state, and long-run stationary analysis may only offer a poor approximation for system behavior. Green et al. [26] examined the effects of non-stationarity on the performance of Markovian queueing systems, and concluded that steady-state results deviate substantially from non-stationary behavior in the cases they investigated. The limitations of the widely-used stationary models in manufacturing are emphasized in [66, 67, 78]. Nevertheless, time-dependent behavior is rarely considered in the context of production planning since transient analysis of general queueing systems is extremely difficult. In the remainder of this chapter, review of past literatures will be given. The literature review will be divided into two parts: existing production planning models and transient analysis methods.

2.1 Production Planning

The production planning models in the literature can be divided into two categories: load-independent and load-dependent models.

2.1.1 Load Independent Production Planning Models

The most widely used Material Requirements Planning (MRP) [70], and the later introduced Capacitated Material Requirements Planning (MRP-C) [84] and Manufacturing Resources Planning (MRP-II) all belong to the load-independent models, which treat the system performance (CT and WIP) as independent of the input release. These models have been criticized ever since their introduction. The performance measures (e.g., CT) were considered as exogenous constants which are obtained from steady stage simulation or historical data. Hence, they suffer from the serious problem of planning circularity [71]: that is, to obtain a production plan the CT of the products needs to be known, whereas the CT of a manufacturing system in turn depends on the input plan which is to be determined. Some later works [62, 30, 31, 32] also adapted the MRP system to uncertain demand environments. However, none of them resolves the fundamental issue of treating the CT as independent of the manufacturing system. For detailed review of MRP, MRP-C, MRP-II please refer to [86].

Deterministic linear and integer optimization models have been used for production planning since 1950's [35, 36, 41, 87]. Hackman and Leachman [45] present a generic linear programming (LP) formulation for production planning in a multi-stage system. In their model, available machine hours within each planning bucket is

treated as a capacity and output time lags between stages need not be integer multiples of the planning bucket time. However, these time lags are still independent of the system workload, and only represent delays such as transportation or curing time, rather than congestion due to the queueing effects. Billington [6] proposed a general LP model for multi-production, multi-period production planning, whose objective is to minimize the inventory and labor costs. As pointed out by Hackman et al. [45], LP models assume that production is instantaneous and uniformly distributed across the planning bucket. Another anomaly for such LP models is that the dual price of capacity is zero until the capacity constraint is saturated, implying 100% resource utilization. However, queueing models suggest that the system performance degrade nonlinearly as utilization reaches 100%, which implies a positive dual price for capacity at lower utilization levels [4, 63]. For a complete review of LP models in production planning, refer to the work by Thomas and McClain [85]. Other works, which try to integrate this non-linear relationship into LP models, include [2, 22, 74, 75, 89].

2.1.2 Load Dependent Production Planning Model

Hackman and Leachman [45] were among the first researchers to incorporate the load-dependent relationship into the linear programming (LP) models. They used a

framework to model the non-integer lead times in a LP model, which allows the material release in one planning bucket to departure from the system at different planning buckets. They use pre-established weights to associate the cumulative output in a planning bucket to the input release. Later, Learchman [57] improved this model by obtaining the weights from historical data of CT. This model provides the basis for IMPReSS, a successful industrial application at Harris Corporation [58], the winner of the INFORMS Franz Edelman Award. Implementation of IMPReSS improved the on-time delivery from 75% to 95%. Caramanis et al. [11] developed an LP model, in which the lead time is modeled as a nonlinear function of the workload, which is built upon general queueing analysis estimates. Their lead time (cycle time)-work load nonlinear function is based on steady state, assuming the planning bucket is long enough for system to reach steady state. However, this nonlinear function is not able to describe the system's input-output behavior, but rather serves as a upper bound to limit the output that can be produced within one planning bucket.

Hung and Leachman [47] combines the simulation and LP model and developed an iterative approach for production planning. Their approach switches between simulation, which estimates the weight of projected output, and LP model, which takes the simulation results as certain kinds of constrains. However, research work [48] has shown that convergence of such iterative approach is not guarantee, and it

depends on the structure of the underlying production system.

Graves in 1986 [24], Karmarkar [42, 43, 44] and Srinivasan et al. [80] developed the approach of integrating the clearing function (CF) idea into the mathematical programming for production planning. Clearing functions are generally expressed in the following form: $Capacity = \alpha(WIP) \times WIP = f(WIP)$. Where α is the so-call clearing factor. The clearing function models the system capacity as a function of the workload (WIP). In [42], the clearing factor specifies the fraction of WIP that can be completed (cleared) by a resource in a give period of time [71]. The CF functions are stationary models and thus are not able to capture the dynamic evolution of the system.

In light of the discussions above, all the existing production planning models fall short in fully characterizing the nonlinear and time-dependent behaviors of manufacturing systems, which is a notoriously difficult task.

2.2 Transient Analysis of Queueing System

In the literature, both analytical methods and computer simulation have been used to address the time-dependent behavior of queueing systems. For Markov queueing models, time-dependent ordinary differential equations (ODEs) can be developed

to represent their input-output dynamics. However, analytical solutions to these ODEs are rare. A few exceptions include the known solutions for the M/M/1 and M(t)/G/∞ systems [28, 53], and the Ph(t)/Ph(t)/∞ systems investigated by Nelson and Taaffe [68, 69]. The mainstay of the analytical work on transient analysis has been the development of numerical solutions of the time-dependent ODEs characterizing the transient behavior of the Markov models. Ingolfsson et al. [39] provides a fairly complete review of these methods including Rothkopf and Oren [76], Clark [14], Gross and Miller [29], Taaffe and Ong [82], Green and Kolesar [25, 26], Eick et al. [20, 21], Jennings et al. [40], and Massey and Whitt [61]. Other techniques for approximating the transient behavior of queues include fluid approximations, which are accurate when there is little variability, and diffusion models, which are good for heavily loaded systems [13, 52, 60]. The analytical method developed in Riano [75] can be considered as a parallel to the fluid and diffusion approximations. Green et al. [27] also reviews various queueing methods for approximating the transient performance of service systems such as call centers. All these methods can be roughly divided into two categories: those that are highly accurate but computationally intensive (comparable to detailed simulation), and those that are fast but inaccurate. Nevertheless, a common limitation of these methods is that they rely on analytical assumptions of one sort or another, and thus are inadequate to capture many features

of realistic manufacturing systems such as non-Markovian interarrival/service times, machine failures, reentrant product flows, etc.

Computer simulation is an alternative approach to address the transient behavior of queueing systems because of its high fidelity and flexibility, and also because of its ease of use and wide acceptance among practitioners. The shortcoming of simulation is that many replication runs are required to obtain good estimates of time-dependent performance measures, and thus simulation is frequently too computationally demanding for real-time "what-if" analysis.

The objective of this work is to develop a TFMs-based approach, which is able to overcome the drawbacks of the existing transient analysis methods, and to utilize the system dynamics described by the TFMs to support responsive production planning.

Chapter 3

Evaluating the Transient Behavior of Queueing Systems via Transfer Function Modeling

3.1 TFMs Methodology

The system of interest will be considered as a queuing system that involves three major time-dependent processes:

$Q(t)$: the state process representing the number of jobs in the system at time t, with $t \in (-\infty, \infty)$, the whole time axis.

$A(u, v)$: the random variable counting the number of arrivals in the system within the time interval $(u, v]$, $u < v \in (-\infty, \infty)$.

$D(u, v)$: the random variable counting the number of departures in the system within the time interval $(u, v]$, $u < v \in (-\infty, \infty)$.

Both $A(u, v)$ and $D(u, v)$ are general point processes [15] that count the number of event occurrences over a time interval, special examples for which include Poisson, renewal, self-exciting processes, and marked point processes [16]. In this work, it is assumed that neither the arrival pattern nor the service times depend on the state of the system. Hereby (until Section 3.5.2), discussions will be restricted to a single-station system. The extension to multi-station environment will be discussed in Section 3.5.3.

Let $\mathcal{H}_0 = \{Q(t), A(-\infty, t), D(-\infty, t), t \in (-\infty, 0]\}$ denote the history of the system evolution up to time 0. The question intended to address here is: at time 0, how to predict the system's behavior from time 0 onward given the history $\mathcal{H}_0$? Obviously, the following flow-balance equation holds for the system

$$Q(t) = Q(0) + A(0, t) - D(0, t) \qquad t > 0,$$

where $\{A(0, t), t > 0\}$ is considered as the independent variable (the input flow imposed on the system), and $\{Q(t), D(0, t), t > 0\}$ the dependent variables representing

the output performance of the system. Note that $Q(t)$, $A(0,t)$, and $D(0,t)$ are all time-varying random variables, and the objective of this work is to establish the time-dependent relationship between these three processes. Defining the following notations,

$m_1(t) = \mathrm{E}[Q(t)]$, the expectation (first moment) of the number of jobs in the system at time t.

$m_2(t) = \mathrm{E}[Q(t)^2]$, the second moment of the number of jobs in the system at time t.

$a(t) = \lim_{\delta \to 0^+} \delta^{-1} \mathrm{E}\{A(t, t+\delta) > 0\}$, the expected arrival rate to the system at time t.

$d(t) = \lim_{\delta \to 0^+} \delta^{-1} \mathrm{E}\{D(t, t+\delta) > 0\}$, the expected departure rate from the system at time t.

It is assumed that $a(t)$ and $d(t)$ exist and are finite. Usually, $a(t)$ and $d(t)$ are also referred to as the intensity or rate of the corresponding point process. Denoting $\mathbf{y}(t) = (m_1(t), m_2(t), d(t))$ as the 3×1 vector including the output performance variables, and $x(t) = a(t)$ the input variable, the goal is to characterize the input-output dynamics of a queueing system by a number of TFMs:

$$\mathbf{y}(t) = \mathbf{F}(\mathbf{x}(t-1), \mathbf{x}(t-2), \ldots, \mathbf{y}(t-1), \mathbf{y}(t-2), \ldots), \tag{3.1}$$

which is a discrete-time functional approximation that describes the dynamics of the queueing system. The time t in equation (3.1) denotes discrete time points. In the rest of this paper, t will be used to represent both continuous and discrete time index, and any possible confusion is avoidable at the price of a negligible amount of mental energy.

The vector function $\mathbf{F}$ in the TFMs (3.1) includes two equations, and is of the same dimension as $\mathbf{y}(t)$. Each component of $\mathbf{F}$ is a difference equation relating an output performance at time t to the input and output history of the system. Suppose that given the current time 0 and the future time horizon is $(0, T]$. Further given the seed values of $\{x(t), \mathbf{y}(t)\}$, which can be derived from $\mathcal{H}_0$, we can use the TFMs to compute recursively the system's future performance $\{\mathbf{y}(t), t \in (0, T]\}$ under any input $\{x(t), t \in (0, T]\}$.

To estimate the TFMs, which can accurately characterize the transient dynamics of a general queueing system, the three steps as mentioned in 1.3 will be discussed respectively in the following sections.

3.2 Non-Stationary Queueing Analysis

In this section, analytical analysis on some simple queueing systems is performed to gain insights to their non-stationary behavior.

3.2.1 An M(t)/M/∞ Example

For the purpose of intuition and motivation, let's consider the input-output dynamics of the simple queueing model M(t)/M/∞, which is one of the very few models whose transient behavior can be characterized analytically. Suppose that the service rate for each job is μ. From the Kolmogorov forward equations for the state probabilities [77], the following equations for the M(t)/M/∞ can be easily derived:

$$\begin{aligned} m_1'(t) &= dm_1(t)/dt = x(t) - \mu \cdot m_1(t) \\ d(t) &= \mu \cdot m_1(t) \end{aligned} \tag{3.2}$$

These equations characterize the system evolution in terms of $m_1(t)$ and $d(t)$. Given the initial state of the system at time 0, the numeric solution of $\mathbf{y}(t) = (m_1(t), d(t))$, $t > 0$ can be obtained for any input $x(t)$, $t > 0$.

Unfortunately, the situation becomes much more complicated as a finite number of servers is introduced or the Markovian assumption is relaxed. The objective of the proposed work is to obtain a discrete-time approximation of equations like (3.2) for

a general queueing system so that its dynamic behavior can be characterized.

3.2.2 A General Queue

Let's consider a single-station queueing process $Q(t)$ with arrivals $A(t)$ and departures $D(t)$, as described in Section 1.3. The arrival and departure rates are denoted as $a(t)$ and $d(t)$ respectively. The additional assumptions made solely for the analytical analysis of this section are:

$$\Pr\{A(t, t+\delta) > 1\} = o(\delta); \quad \Pr\{D(t, t+\delta) > 1\} = o(\delta) \tag{3.3}$$

where $o(\delta)$ denotes any function that goes to zero with δ faster than δ itself.

Conditions (3.3) imply that there are no multiple simultaneous arrivals or departures, i.e., both $A(t)$ and $D(t)$ are orderly point processes [16].

Following the notation given in Section 1.3, let

$$\begin{aligned} a_n(t) &= \lim_{\delta \to 0^+} \delta^{-1} \Pr\{A(t, t+\delta) = 1, Q(t) = n\} \\ d_n(t) &= \lim_{\delta \to 0^+} \delta^{-1} \Pr\{D(t, t+\delta) = 1, Q(t) = n\} \end{aligned} \tag{3.4}$$

Here, $a_n(t)$ denotes the arrival rate at time t while there are n jobs (not including the one that is about to enter) in the system, and $d_n(t)$ represents the departure rate at time t with n jobs (including the one that is about to leave) in the system.

Apparently, the following equation can be established:

$$a(t) = \sum_{n=0}^{\infty} a_n(t) \text{ and } d(t) = \sum_{n=1}^{\infty} d_n(t).$$

Suppose that the system consists of a single server with service time following a general distribution, say $G(\tau)$, where $\tau \in (\tau_L, \tau_U)$, the feasible time range for the service time. Jobs are served on a first come first serve basis. For this general queue, the dynamic equations for the $x(t)$-$\mathbf{y}(t)$ relationship with $x(t) = a(t)$ and $\mathbf{y}(t) = (m_1(t), d(t))$ have been derived in Appendix A and listed as following:

$$\begin{aligned} m_1'(t) &= a(t) - d(t) \\ d(t) &= \int_{\tau_L}^{\tau_U} a_0(t-\tau) dG(\tau) + \int_{\tau_L}^{\tau_U} (d(t-\tau) - d_1(t-\tau)) dG(\tau) \end{aligned} \tag{3.5}$$

Unlike equations (3.2) for the M(t)/M/∞ system, equations (3.5) for the general queue are not closed, and thus not solvable: Aside from the input process $a(t)$ and the output processes of interest $m_1(t)$ and $d(t)$, (3.5) also involves unknown time-dependent functions $a_0(t)$ and $d_1(t)$. However, for relatively heavily loaded queues (utilization > 0.5), it is reasonable to assume that $a_0(t)$, the arrival rate when no job is in the system, and $d_1(t)$, the departure rate when no job is in the waiting queue, are relatively small and can be approximated by:

$$a_0(t) \approx p_0(t) \times a(t) \approx e_1 a(t) \text{ and } d_1(t) \approx p_1(t) \times d(t) \approx e_2 d(t).$$

Both e_1 and e_2 are small fractional constants. Further, taking the finite-difference approximation of the derivative and integrals in (3.5), it is clear that the discrete approximations of equations (3.5) fall into the category of TFMs (3.1). Similar dynamic equations have also been obtained for single-station systems with multiple servers in Appendix A.

The analytical results (3.5) serve three purposes here. First, it shows that even for a single-server queue with general arrivals and services (a simplest queue), its non-stationary behavior is analytically intractable. Hence, the approach of TFMs-based discrete approximation may be appropriate for investigating the time-dependent behavior of general queueing systems. Second, as will become clear in Section 3.5 , the basis of describing the dynamics of a multi-station system lies in the use of TFMs (3.1) to approximate the transient behavior of a single station (or a group of stations that can be considered as a whole), for which the single-station queue considered above is fairly general and representative. Therefore, equations (3.5) strongly suggest that the TFMs as (3.1) are likely to be successful in terms of capturing the system dynamics. As a matter of fact, it was these analytical results that motivated the author to adopt the TFMs (3.1) in the first place. Third, equations (3.5) provide some valuable insights as to the specific functional forms of the target TFMs, which is very useful in the statistical fitting of the parametric models.

As already noted, the additional assumptions (3.3) were made here solely for the analytical analysis. Whereas the TFM modeling, as evident in Section 3.5, is expected to be able to describe the dynamic behavior for general queueing systems with failures and re-entrant flows. Next, in Sections 3.3 and 3.4, the issues associated with the simulation-based TFM modeling will be discussed in detail.

3.3 Data Collection via Offline Simulation

In this part, how to obtain the simultaneous pairs of input-output observations $\{(X(t), \mathbf{Y}(t)), t = 1, 2, \ldots, T\}$ by running simulation is discussed. Note that the capital letters here are used to represent the estimated time series obtained from simulation.

3.3.1 Simulation

In this work, discrete event simulation models are constructed both in Matlab and C++ to represent the queueing systems of interest. The simulation models are verified and validated following the approaches recommended in Law and Kelton [56], such as running the model with simplified assumptions to detect logical mistakes and testing the model outputs under a variety of input settings, comparing simulation results

from simple queueing system with known results. The input flow of entities $A(t)$ is modeled as a point process which is characterized by its first moment measure, i.e., the input rate $a(t)$ (Section 1.3). As will be seen in Section 3.5, in our experiments two types of input processes are fed to the system: Poisson and equilibrium renewal processes with $a(t)$ being a piece-wise constant function over time t (e.g., Figure 3.1 in section 3.5.1).

For a given queueing system, a number of, say I, simulation replications are performed with the input flow being a stochastic process characterized by a time-varying rate. For replication i ($i = 1, 2, \ldots, I$), the arrival, departure and state processes $\{A_i(t), D_i(t), Q_i(t); t = 1, 2, \ldots, T\}$ are recorded. It is assumed that the system is observed at discrete, equal-spaced intervals of time, and that the basic sampling interval Δt serves as the unit of time. The paired time series $\{(X(t), \mathbf{Y}(t)), t = 1, 2, \ldots, T\}$ are estimated as follows.

$$
\begin{aligned}
X(t) &= \widehat{a}(t) = \frac{I^{-1} \sum_{i=1}^{I} A_i(t - \Delta t/2, t + \Delta t/2)}{\Delta t} \\
Y_1(t) &= \widehat{m_1}(t) = I^{-1} \sum_{i=1}^{I} Q_i(t) \qquad (3.6) \\
Y_2(t) &= \widehat{m_2}(t) = I^{-1} \sum_{i=1}^{I} Q_i^2(t) \qquad (3.7) \\
Y_3(t) &= \widehat{d}(t) = \frac{I^{-1} \sum_{i=1}^{I} D_i(t - \Delta t/2, t + \Delta t/2)}{\Delta t}
\end{aligned}
$$

It can be seen from (3.6) that both the arrival rate $X(t)$ and the departure rate $Y_3(t)$ are defined in terms of the average number of occurrences per Δt. The sampling interval should be sufficiently small to allow all the systematic variation which occurred in the inputs/outputs to be taken account of. In our experiments, Δt is set to be one tenth of the expected processing time of the server, which is typically smaller than the average interarrival time of entities.

3.3.2 Design of Simulation Experiments

Although the simulation involved in the proposed work is performed offline, it remains important to design simulation experiments so that accurate TFMs can be obtained at high computational efficiency.

Input Range of Interest

To collect data for investigating the system's transient behavior, stochastic arrival processes with piece-wise constant rate $x(t) = a(t)$ are fed to the simulation model. The different levels for $x(t)$ are denoted as $\{x_1, x_2, \ldots, x_M\}$, and here the range of interest for those input levels is established.

This work aims at characterizing the performance of a queueing system when it is relatively heavily loaded, which is the typical situation for transfer line manufacturing.

The capacity of a system $\mu(\Pi)$, defined as the upper limit on the arrival rate for long-term stability, depends on the system configuration Π (e.g., number of servers, service time, machine failures) in a single-job environment considered in this paper. Given Π, existing queueing models [38, 55, 64] can be used to accurately or even exactly calculate the capacity $\mu(\Pi)$ of real systems that involve batching, re-entrant flows, machine setups, etc. Thus, in this paper the author use the analytically obtained capacity $\mu(\Pi)$ to specify the range $[x_L, x_U]$ for the arrival rates $\{x_i; x_i = 1, 2, \ldots, M\}$ so that the facility is reasonably utilized. Denote the steady-state system utilization by $\rho = x/\mu(\Pi)$, and the range for ρ is given as $[\rho_L, \rho_U]$. Then, the input range for arrival rates is given as:

$$[x_L, x_U] = \mu(\Pi)[\rho_L, \rho_U] \tag{3.8}$$

In the experiments, the system utilization range $[\rho_L, \rho_U] = [0.5, 0.94]$ is set, which covers relatively heavily loaded utilization range.

Experiment Design Strategies

In the context of this work, the design of experiments (DOE) needs to address the following questions regarding the selection of input $x(t)$ and the number of simulation replications required.

First, how should $\{x_1, x_2, \ldots, x_M\}$ be selected? This includes determining the size

M and the values of x_m $(m = 1, 2, \ldots, M)$ with $x_m \in [x_L, x_U]$. In the proposed methods, such selection is guided by the system's steady-state behavior, which has been thoroughly investigated in the literature and a recent sequence of papers by Yang and co-authors [90, 91, 92, 93]. Based on their empirical experience, five different arrival rates, approximately evenly spread over relatively heavily utilized range $[x_L, x_U]$, are sufficient to characterize the steady-state input-output relationships. That is, the system input rate is a step function with five levels, representing the five arrival rates. Here, transient experiments inherit this five evenly-spaced arrival rates selection from steady-state study:

$$(x_L, x_L + (x_U - x_L)/4, x_L + 2(x_U - x_L)/4, x_L + 3(x_U - x_L)/4, x_U). \tag{3.9}$$

Second, how to determine the sequence of these five arrival rates? It is desirable to be able to examine the interaction effects of $x(t)$ and $\mathbf{y}(t) = (m_1(t), m_2(t), d(t))$. For instance, does the same $x(t)$ cause $m_1(t)$ to respond in a different manner when $m_1(t)$ is at different values? In light of this, the author determine the sequence of $\{x_1, x_2, \ldots, x_5\}$ in such a way that the minimum jumps (up or down) between the successive levels of arrival rates is maximized. Initially at time 0^-, the system is empty and the arrival rate is $x_0 = 0$, and the values of $\{x_1, x_2, \ldots, x_5\}$ is determined

by

$$\max_{x_1,x_2,\ldots,x_5} \min\{|x_m - x_{m-1}|, i = 1, 2, \ldots, 5\} \tag{3.10}$$

The solution to (3.10) can be easily obtained by permutating the five different levels.

Third, how to determine the simulation length ℓ_m at each input level x_m ($m = 1, 2, \ldots, M$)? Preliminary simulation experiments are performed to determine the length of transient period $\ell_m^{(0)}$ of initially empty system under input x_m. The author set $\ell_m = 2\ell_m^{(0)}$ to ensure that sufficient data is collected in steady state, the reason of which will become clearer in Section 3.4.1.

Fourth, how many replications should be performed at the selected time-varying input process? The author use $\gamma\%$, the desired precision of estimate $Y_1(t) = \widehat{m_1}(t)$ to determine the number of replications in a two-step sequential process. In the first step, I_0 replications are performed feeding the input process $x(t)$ to the simulation. Denoting $\widehat{m}_{1,0}(t)$ as the estimate from the I_0 replications, and the maximum sample standard deviation of $\widehat{m}_{1,0}(t)$ over the simulation period $[1, T]$ is given as:

$$\widehat{\sigma}^{(0)}_{\max} = \max_{t=1,2,\ldots,T} \widehat{\sigma}(\widehat{m}_{1,0}(t)) \tag{3.11}$$

Let $t_{\max}$ be the time index that achieves $\widehat{\sigma}^{(0)}_{\max}$. The number of replications I that is likely to provide the desired precision $\gamma\%$ is estimated as: $n = \lceil (\widehat{\sigma}^{(0)}_{\max})^2 / (\widehat{m}_{1,0}(t_{\max}) \times \gamma\%)^2 \rceil$

In the second step, $I - I_0$ simulation replications are performed, and based on all the simulation carried out, the time series estimates are obtained using equations (3.6) and will be used for the TFM fitting discussed in Section 3.4.

3.4 Statistical Modeling Issues of the TFMs

The modeling of the system dynamic behavior is based on the pair estimates $\{X(t), \mathbf{Y}(t), t = 1, 2, \ldots, T\}$ obtained from simulation experiments (Section 3.3.1). These estimates are subject to random errors, and the author use the following parametric model to represent the stochastic correspondent of the TFMs (3.1):

$$\mathbf{Y}(t) = \mathbf{F}(\boldsymbol{\theta}; X(t-1), X(t-2), \ldots, \mathbf{Y}(t-1), \mathbf{Y}(t-2), \ldots) + \mathbf{e}(t), \tag{3.12}$$

where $X(t) = \widehat{a}(t)$ and $\mathbf{Y}(t) = (\widehat{m_1}(t), \widehat{m_2}(t), \widehat{d}(t))$ as given in (3.6). The term $\mathbf{e}(t) = (e_1(t), e_2(t), e_3(t))$ denotes the disturbance. The parameter vector $\boldsymbol{\theta}$ includes all the unknown parameters involved in the vector function $\mathbf{F}$. For convenience of

discussion, model (3.12) are written as:

$$\begin{aligned} Y_1(t) &= F_1(\boldsymbol{\theta}_1; X(t-1), X(t-2), \ldots, \mathbf{Y}(t-1), \mathbf{Y}(t-2), \mathbf{Y}(t-3), \ldots) + e_1(t) \\ Y_2(t) &= F_2(\boldsymbol{\theta}_2; X(t-1), X(t-2), \ldots, \mathbf{Y}(t-1), \mathbf{Y}(t-2), \mathbf{Y}(t-3), \ldots) + e_2(t) \\ Y_3(t) &= F_3(\boldsymbol{\theta}_3; X(t-1), X(t-2), \ldots, \mathbf{Y}(t-1), \mathbf{Y}(t-2), \mathbf{Y}(t-3), \ldots) + e_3(t) \end{aligned} \tag{3.13}$$

with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$. Our task here is to obtain the TFMs that are of the simplest functional form and adequate to describe the system's dynamic evolution based on the paired simulation data $(X(t), \mathbf{Y}(t))$.

3.4.1 Estimation of the TFMs

In this section, the fitting of the TFMs assuming that a specific functional form (model structure) has been selected will be discussed.

Error Term

In this work, it is assumed that each disturbance $e_i(t)$ $(i = 1, 2, 3)$ can be approximated by a stationary autoregressive moving average (ARMA) process [8], which can be expressed as follows

$$e_i(t) = \frac{C_i(q)}{D_i(q)} w_i(t) \qquad i = 1, 2, 3. \tag{3.14}$$

The white noise $w_i(t)$ is normally distributed with a mean of zero and a variance of σ_i^2. The backward shift operator q^{-1} is defined by $q^{-1}z(t) = z(t-1)$, and $q^{-m}z(t) = z(t-m)$. The operators $C_i(q)$ and $D_i(q)$ are defined as:

$$C_i(q) = \sum_{k=0}^{\infty} c_i(k)q^{-k} \tag{3.15}$$

$$D_i(q) = \sum_{k=0}^{\infty} d_i(k)q^{-k} \quad i = 1, 2, 3. \tag{3.16}$$

For an ARMA process, the number of non-zero coefficients $c_i(k)$ (or $d_i(k)$), $k = 0, \ldots, \infty$ is finite, and typically does not exceed two [8].

Two-Step Fitting Process

Since the error terms $\{e_i(t), i = 1, 2, 3\}$ are not i.i.d (identically and independently distributed) normal, a two-step fitting process to estimate the TFMs are proposed.

First, the TFMs are fitted to the $\{X(t), \mathbf{Y}(t), t = 1, 2, \ldots, T\}$ data using least-square methods [59] as if the errors were i.i.d normal. The residuals $\{\widetilde{e}_i(t), i = 1, 2, 3\}$ will be computed based on the resulting TFMs obtained in this step, and the ARMA process that can best approximates $\widetilde{e}_i(t)$ $(i = 1, 2, 3)$ will be identified and estimated following the approaches in [8]. That is, for $\widetilde{e}_i(t)$ $(i = 1, 2, 3)$, $C_i(q)$ as in (3.15) and $D_i(q)$ as in (3.16) will be completely specified with the estimated coefficients $\{\widehat{c}_i(k), k = 0, \ldots, \infty\}$ and $\{\widehat{d}_i(k), k = 0, \ldots, \infty\}$; and the variance of the white noise $w_i(t)$ can also be obtained as $\widehat{\sigma}_i^2$.

Second, least square method has been used to refit the TFMs assuming that the errors are given as the ARMA processes estimated from the previous step. Specifically, the TFMs (3.13) will be transformed as follows to achieve additive white noise $w_i(t)/\sigma_i$ $(i = 1, 2, 3)$ with constant variance 1.

$$\begin{aligned}
\frac{D_1(q)}{\sigma_1 C_1(q)} Y_1(t) &= \frac{D_1(q)}{\sigma_1 C_1(q)} F_1(\boldsymbol{\theta}_1; X(t-1), X(t-2), \ldots, \mathbf{Y}(t-1), \mathbf{Y}(t-2), \ldots) + \frac{w_1(t)}{\sigma_1} \\
\frac{D_2(q)}{\sigma_2 C_2(q)} Y_2(t) &= \frac{D_2(q)}{\sigma_2 C_2(q)} F_2(\boldsymbol{\theta}_2; X(t-1), X(t-2), \ldots, \mathbf{Y}(t-1), \mathbf{Y}(t-2), \ldots) + \frac{w_2(t)}{\sigma_2} \\
\frac{D_3(q)}{\sigma_3 C_3(q)} Y_3(t) &= \frac{D_3(q)}{\sigma_3 C_3(q)} F_3(\boldsymbol{\theta}_3; X(t-1), X(t-2), \ldots, \mathbf{Y}(t-1), \mathbf{Y}(t-2), \ldots) + \frac{w_3(t)}{\sigma_3}
\end{aligned}$$

The least-square fitted parameters $\widehat{\boldsymbol{\theta}}$ is the solution to the following optimization problem:

$$\begin{aligned}
&\min_{\boldsymbol{\theta}} \mathrm{SSE}(\boldsymbol{\theta}) \\
&= \sum_{t=1}^{T} \sum_{i=1}^{3} [\frac{D_i(q)}{\sigma_i C_i(q)} Y_i(t) - \frac{D_i(q)}{\sigma_i C_i(q)} F_i(\boldsymbol{\theta}_i; X(t-1), X(t-2), \ldots, \mathbf{Y}(t-1), \mathbf{Y}(t-2), \ldots)]^2,
\end{aligned}$$

given time-series data $\{X(t), \mathbf{Y}(t), t = 1, 2, \ldots, T\}$. The process of the two-step fitting process can be repeated until there is no significant changes in the fitted TFMs. However, in our experience, it suffices to perform one round of the fitting process to achieve well-estimated TFMs.

Stability

The dynamic TFMs are required to be stable: the TFMs should be able to converge

to the system's steady-state input-output relationships. Suppose that the input $x(t)$ is held constant at the level x. the outputs are written as $\mathbf{y}(x,t)$ to emphasizing the dependence of the outputs on time t as well as on the input level x. With the input held fixed at x, the dynamic outputs described by the TFMs should eventually converge to $\mathbf{y}(x,\infty)$, the steady-state equilibrium. In the least square estimation of the TFMs described above, no additional constraints were imposed on the model fitting to ensure stability. Rather, in our method, the simulation data are collected in such a way that a substantial amount of steady-state time series (Section 3.3.2) are included. Hence, the TFMs fitted from the data also well reflect the steady-state behavior of the system.

Statistical Inference

In [59], the asymptotic normality of the least-square parameters $\widehat{\boldsymbol{\theta}}$ has been proved, and the statistical inference on the estimated TFMs is discussed in Appendix B.

3.4.2 Model Selection

The estimation of the TFMs in Section 3.4.1 is based on a given functional form. In this section, the selection of the most appropriate structure for the target model are discussed. Achieving the parsimonious TFMs that can accurately describe the system's transient performance is difficult, and a number of venues in search of the

best TFMs are resorted.

Identification of the Model Family

Transient Queueing Analysis

According to the transient queueing analysis in Section 3.5.2, the simplest possible form for the TFMs is likely to be:

$$y_1(t) = b_0 + b_1 x(t-1) + b_2 y_1(t-1) + b_3 y_2(t-1) \tag{3.17}$$

$$y_2(t) = c_0 + c_1 x(t-1) + c_2 y_1(t-1) + c_3 y_2(t-1). \tag{3.18}$$

$$y_3(t) = d_0 + d_1 x(t-1) + c_2 y_3(t-1). \tag{3.19}$$

Steady-State Behavior

The previous study performed in [90] suggests that models of the following functional form can be used to approximate the steady-state behavior of real manufacturing systems.

$$y_1(\infty) \approx \frac{P(x/\mu(\Pi))}{1 - x/\mu(\Pi)} \approx P(x/\mu(\Pi))[1 + x/\mu(\Pi) + (x/\mu(\Pi))^2 + \cdots] \tag{3.20}$$

$$y_3(\infty) = x \tag{3.21}$$

Here, x denotes the constant rate of arrivals into the system, and $\mu(\Pi)$ denotes the system capacity (Section 3.3.2). The stability condition requires: $x/\mu(\Pi) < 1$. The term $P(x/\mu(\Pi))$ represents a polynomial function of $x/\mu(\Pi)$. As mentioned earlier,

the TFMs are supposed to converge to the stationary equations (3.20) and (3.21) in steady state. While equation (3.21) simply means that the departure rate is equal to the arrival rate in the long run, Equation (3.20) indicates the possible existence of higher-order polynomial terms such as $x \cdot d$, x^3, and d^3 in the TFMs. Note that y_2 share the similar shape with y_1, because they are the first and second moment of the number of jobs in process. Therefore, the similar model form will be used for y_2.

Empirical Experience

The above analysis assists to identify the potential TFMs as a model family that may contain main effects, higher-order polynomials or interactions of historical inputs/outputs. In our work, TFMs of such a model family are used to explore the transient behavior of a wide range of queueing systems that involve non-Markov interarrival and processing times, machine failures, re-entrant flows, etc. Based on our experience, a model including up to third-order polynomials/interactions is sufficient to provide an adequate description of the dynamic behavior of a general queueing system:

$$y_i(t) = \sum_{j=0}^{3} \sum_{k=0}^{3-j} \sum_{l=0}^{3-j-k} b_{jkl} \cdot y_1^j(t-1) \cdot y_2^k(t-1) \cdot x^l(t-1) \quad for\ i = 1, 2, 3 \qquad (3.22)$$

In (3.22), no higher time order term is included. The reason is that in our empirical experience, the transient behavior of $y_i(t)$ $(i = 1, 2, 3)$ mainly follows an exponential

trend (Figures 3.2, 3.4, and 3.8), which can be adequately approximated by difference equations of first time order.

Stepwise Model Selection

Stepwise model-building techniques for regression designs with a single depend variable are performed to select the parsimonious TFMs. The basic procedures for stepwise regression include : identifying an initial model, iteratively "stepping" and terminating the search. Good references on the discussion of stepwise regression method can be found in [46, 19].

Given the model family identified in equation (3.22), the selection approach starts with the most complicated model (3.22) as initial model. Backward elimination is followed. Each regressor is tested for statistical significance. Then a Forward selection approach will examine all the deleted regressors for significance, and any regressor which shows significant contribution in explaining the explanatory variable will be added back to the TFMs. The backward and forward selections are repeated until no regressors can eliminated/added to the model or a specified maximum number of steps has been reached.

3.5 Empirical Examples

For a given simulation representing a general queueing system, the job is to describe its transient behavior by generating a number of TFMs from simulation data. Using such TFMs, a system's future dynamics can be predicted in a timely manner without running additional simulation, which could be very time consuming. In this part, three examples are presented to illustrate the effectiveness of the proposed methods: an $E_k(t)/G/1$ system (Section 3.5.1), an $M(t)/G/3$ system with server failures (Section 3.5.2), and a system with six different stations and re-entrant flows (Section 3.5.3). The first two single-station cases are selected from a number of queueing models (Table 3.1) for which the TFM modeling methods have been successfully applied, and these queueing models are intended to show that the proposed methods can handle a wide range of input flows and different types of service time distribution. Note that in Table 3.1, ACV and SCV denote the coefficient of variation (CV) for the distribution of interarrival times and service times respectively. Accurately characterizing the transient behavior of a single station (or a group of stations that can be well approximated as a single one) serves as the basis for capturing the dynamics of a multi-station system. In Section 3.5.3, the six-station example also involves re-entrant flows, one of the main features of real semiconductor fabrication systems, and

the specifics of extending the TFM modeling to multi-station systems are detailed through this example.

Table 3.1: Single station system configurations

# Servers	Interarrival Time	ACV	Service Time	SCV	Failures
$1 \sim 3$	exponential,Erlang,deterministic	$0 \sim 1$	gamma	$0.1 \sim 1$	Yes/No

For each queueing system, the proposed methods were applied for the generation of the TFMs describing the system dynamics. Simulation experiments were performed following the design strategies in Section 3.3, and the resulting data set, which will be referred to as the estimation data set (EDS), will be used to estimate the TFMs following the statistical modeling approaches in Section 3.4. With the fitted TFMs, the future evolution of the system can be predicted under any input flow given the history of the system. To evaluate the prediction provided by the TFMs, a validation data set (VDS), which contains simulation data different than and independent of those in the EDS, was collected and the system dynamics estimated from the VDS was compared to that predicted by the TFMs. For all the numeric examples that have been investigated, the resulting TFMs are able to accurately predict the future evolution of the system, judging from the VDS-based cross validation.

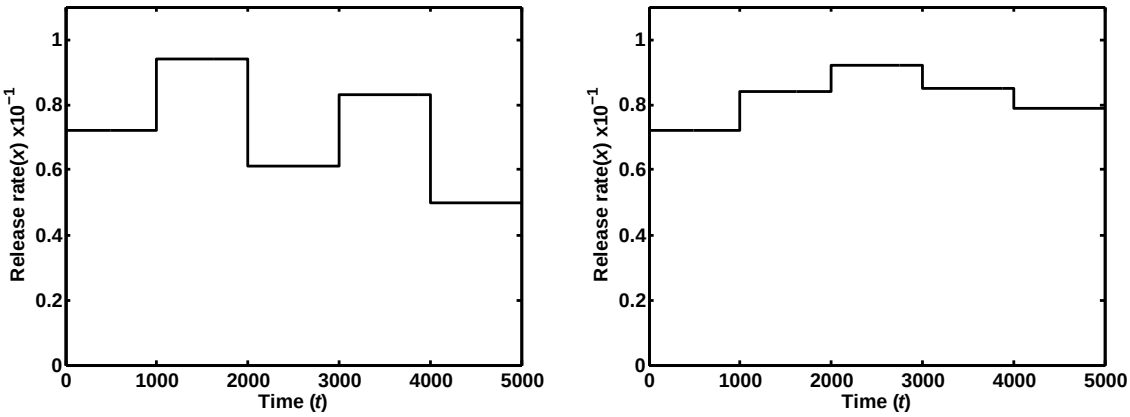
Before discussing the results, it is worth mentioning that in our discrete TFMs,

one time unit represents the sampling interval Δt, which is set as about one tenth of the expected service time of the most heavily utilized server (Section 3.3). To avoid possible confusion, in the examples below, all the time periods (interarrival time, service time, simulation length, and future horizon) are defined in terms of the time unit Δt.

3.5.1 An $E_k(t)/G/1$ System

Next a single-server system is test by the TFMs, whose service time follows a gamma distribution with a mean of 10 time units (i.e., $10\Delta t$) and standard deviation of 5 time unit. The interarrival time of entities follow an Erlang distribution with $k = 25$ stages (denoted as E_{25}), corresponding to a coefficient of variation of 0.2.

To collect the time series data $\{X(t), \mathbf{Y}(t)\}$ for the TFM modeling, simulation experiments were carried out by feeding to the system the arrivals with the piecewise arrival rate shown in Figure 3.1(a). Each piece in Figure 3.1(a) corresponds to a stationary renewal process with a certain first moment measure (rate), the simulation methods of which are discussed in [16] and implemented in our simulation model. The five selected arrival rates are evenly-spaced to cover the system utilization range of $[0.5, 0.94]$, and they are sequenced in such a way that (3.10) is achieved. The simulation length of each constant-rate period is selected to ensure that sufficient



(a) Estimation data (b) Validation data

Figure 3.1: Release rate for estimation and validation data set for Ek/G/1 system

.

data is obtained in steady state (Section 3.3.2). The number of simulation replications performed in this case is 10000, which is determined following the two-step process in Section 3.3.2 to achieve a relative precision level of $\gamma = 5\%$ for the estimate $Y_1(t) = \widehat{m_1}(t)$. From the multiple replications, the paired estimates $\{X(t), \mathbf{Y}(t)\}$ were calculated using equations (3.6). With the collected EDS, the statistical modeling methods (Section 3.4) were applied and the resulting TFMs for this $E_k(t)/G/1$ system

are given as follows:

$$\begin{aligned}
\widehat{m_1}(t) &= 1.0045m_1(t-1) - 0.0798d(t-1) + 0.0902x(t-1) \\
&\quad -0.0512m_1(t-1)x(t-1) + 0.0452m_1(t-1)d(t-1)x(t-1) \\
\widehat{m_2}(t) &= 0.0567m_1(t-1) - 0.0097d(t-1) + 0.9469m_2(t-1) - 0.0107m_1^2(t-1) \\
&\quad -0.2830m_1(t-1)d(t-1) + 0.2494m_1(t-1)a(t-1) \\
&\quad +0.0705d(t-1)m_2(t-1) - 0.0121m_2(t-1)a(t-1) \\
\widehat{d}(t) &= 0.0032m_1(t-1) + 0.9474d(t-1) + 0.0431x(t-1) \\
&\quad +0.0286m_1(t-1)x(t-1) - 0.0304m_1(t-1)d(t-1)x(t-1)
\end{aligned} \tag{3.23}$$

Apparently, given the history $(x(t), \mathbf{y}(t) = (m_1(t), m_2(t), d(t)), t \leq 0\}$, the fitted TFMs (3.23) can be used to recursively compute the future performance for any input $x(t)$ over $(0, T]$, and the computational effort required is negligible.

To evaluate the accuracy of the TFMs (3.23), the VDS were collected by running simulation with the interarrival time following E_{25} and the time-varying arrival rate given in Figure 3.1(b). To avoid TFMs-based extrapolation, the arrival rates in the VDS are set within the rate range $[x_L, x_U]$ used in the EDS. For the VDS, 20000 simulation replications were performed, and highly accurate time series $\mathbf{y}(t) = (m_1(t), m_2(t), d(t))$ were obtained and considered as the "true" dynamic outputs with "zero" variance under the specified input flow. In Figure 3.2, the "true"

outputs $m_1(t)$, $m_2(t)$ and $d(t)$ are plotted as the dotted curves in Figure 3.2(a) and (b) respectively. The solid curves in Figure 3.2 represent the predicted dynamic outputs resulting from the fitted TFMs (3.23). To obtain the predicted curves, the TFMs-based recursive computation was initiated by using the first pair of time-series points in the VDS as the seed values, and iteratively it leads to the prediction of the system evolution over the entire period given that the arrival rate follows Figure 3.1(b). Figure 3.2 shows that the predicted dynamics from the TFMs almost coincide with the "true" system evolution.

3.5.2 An $M(t)/G/3$ System with Failures

The system consists of three identical servers, and the service time follows Gamma distribution with a mean of 10 time units and standard deviation of 2 time units. The service time distribution of low coefficient of variation is again chosen here because in our experience, the cases with high coefficient of variation (e.g., Exponential service time) can be much more easily handled by the TFM modeling. In addition, for this system, each server is subject to exponential failures. The mean time to failure is 36 time units and mean time to repair is 4 time units.

For the EDS, the Markov input flow is characterized by the arrival rates depicted in Figure 3.3(a), while in the VDS, the rate of the Poisson arrivals fed to the system

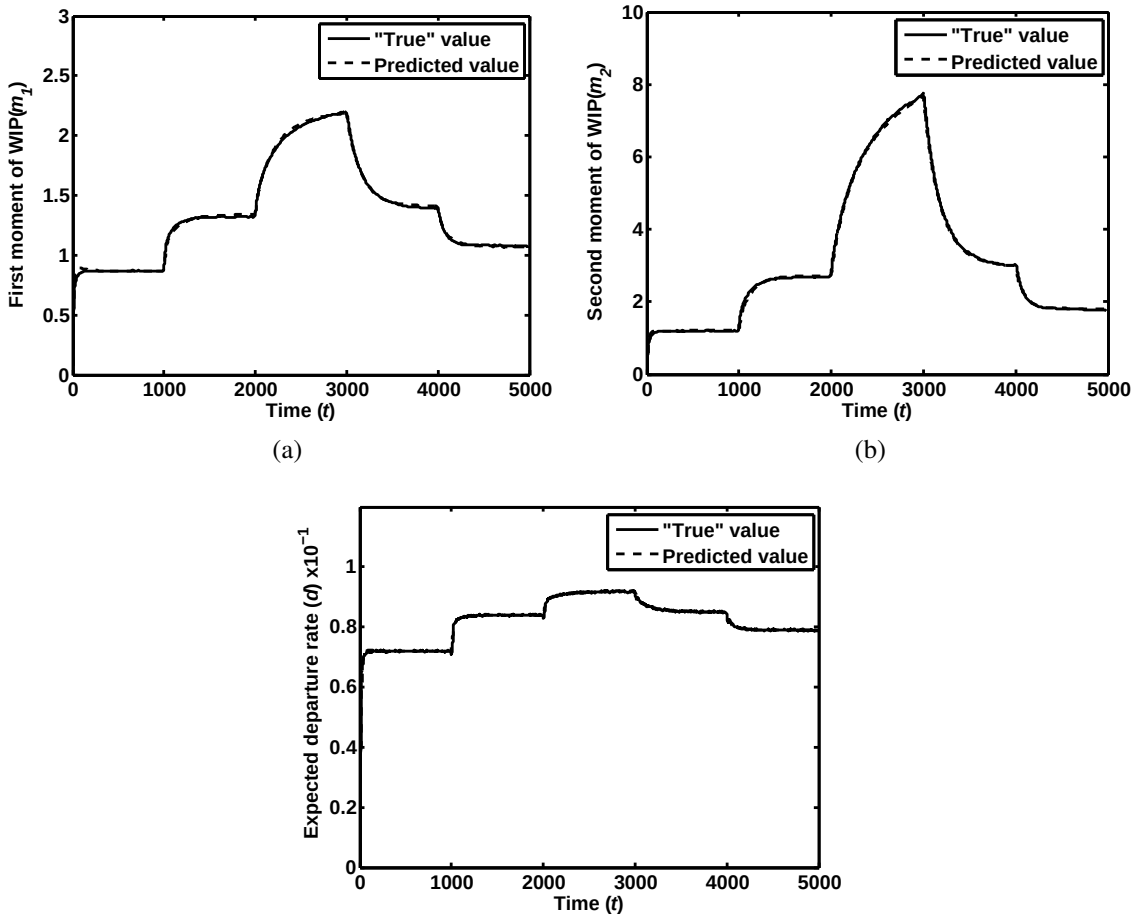
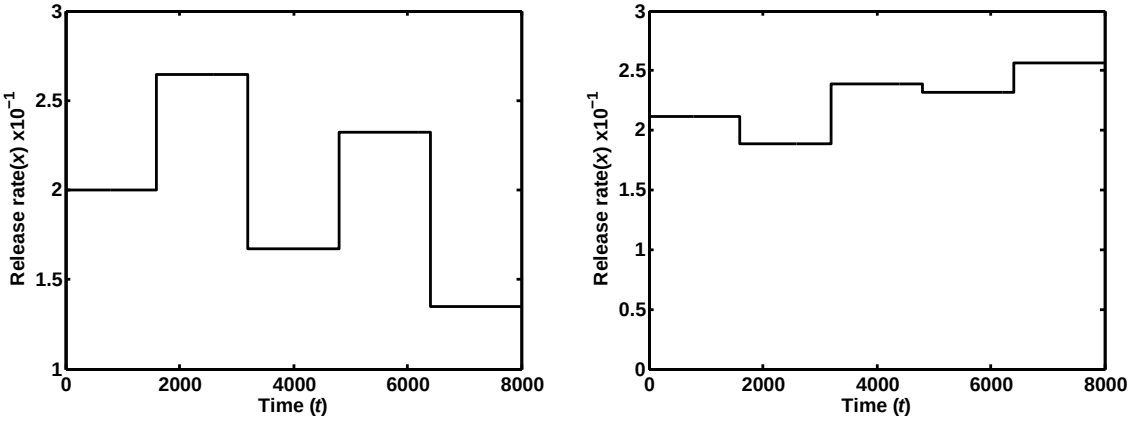


Figure 3.2: Evaluation of the fitted TFMs for the Ek/G/1 system

.



(a) Estimation data (b) Validation data

Figure 3.3: Release rate of estimation and validation data set for M/G/3 system

.

follow Figure 3.3(b). A total of 10000 simulation replications were performed for the EDS, and 20000 carried out for the VDS to obtain the "true" dynamic outputs. As Figure 3.2 for the $E_k(t)/G/1$ case, Figure 3.4 compares the "true" dynamic behavior, represented by the dotted curves, with the system evolution predicted by the TFMs, depicted by the solid curves, under the Markov input flow with the rate given in Figure 3.3(b). Evidently, the TFMs-base prediction is highly accurate.

3.5.3 A Multi-Station System with Re-entrant Flows

The TFMs that well characterize the transient behavior of a single station (or a group of stations) provide the building blocks for describing the dynamics of multi-station

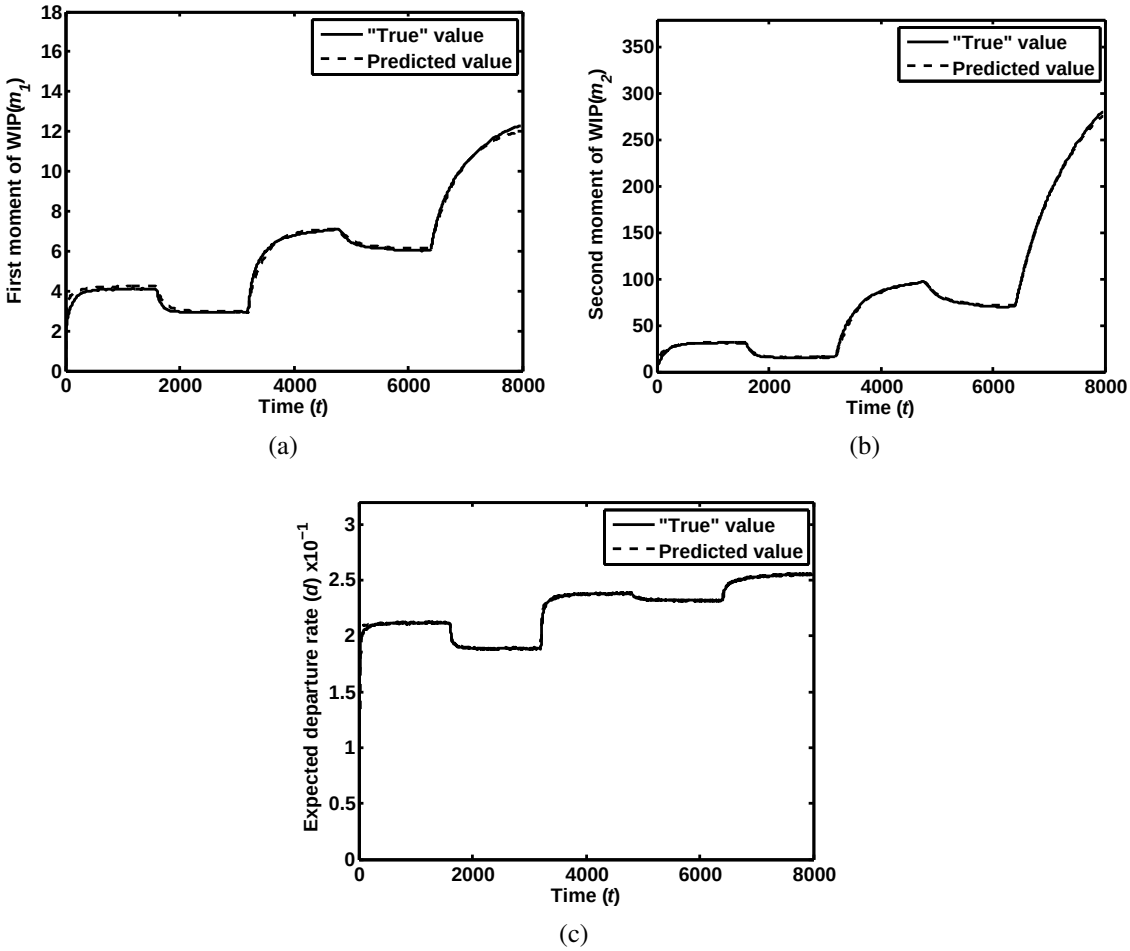


Figure 3.4: Evaluation of the fitted TFMs for the M/G/3 system.

systems, as will become clear in this subsection. The TFMs-based modeling methods is illustrated through the system depicted in Figure 3.5, which includes re-entrant flows, one of the main features of real semiconductor fabrication systems. The system consists of six stations with two re-entrant cycles: $2 \rightarrow 3$, and $4 \rightarrow 5$. Each entity has to visit the first cycle twice before it enters the second cycle, which also needs to be repeated by an entity for two times. Each station consists of three identical servers, and all the service times follow Gamma distribution with a coefficient of variation of 0.5. The mean service time at each of the six stations is given in Table 3.2.

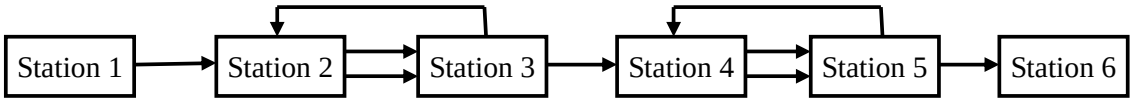


Figure 3.5: Flow chart for six stations in tandem system

.

Table 3.2: Six stations in tandem system configurations

	Station 1	Station 2	Station 3	Station 4	Station 5	Station 6
Mean Service Time	10	10	7	10	7.8	10

Extension to Multi-Station Systems

The basic idea to analyze a multi-station system is to decompose the system into a number of subgroups, treat each subgroup as a single station, and characterize each of them by its TFMs, like those in equations (3.23). The dynamic behavior of the entire system can be described by the multiple sets of TFMs with each set corresponding to a subgroup. The specifics are discussed as follows.

The decomposition of a target system is based on the identification of the most heavily utilized stations (HUSs). A bottleneck station (BNS) is defined as a station that has the maximum utilization in the system. A station whose utilization is above 80% of that of the BNS(s) is considered as a HUS. The HUSs are the stations that restrict the entity flows and thus play a key role in determining the overall performance of the system. For a given system, analytical queueing models in the literature [38, 55, 64] are available to perform utilization analysis for even the most complicated manufacturing systems (i.e., semiconductor manufacturing systems), and thus the HUSs can be identified analytically prior to the simulation-based transfer function modeling. Denoting G as the number of HUSs in a system, the author suggest formulating G subgroups: each subgroup includes one HUS, which dominates the queueing behavior of the group, and some upstream/downstream non-HUSs of that HUS.

The system decomposition has to be made on a case-by-case basis. Here, a simple illustration is provided through the example in Figure 3.5.

The six-station system was decomposed into two subgroups, mainly based on the utilization analysis discussed above. Stations 3 and 5 are considered as HUSs, and the rest stations are non-HUSs. Hence, Subgroup 1 contains Stations 1, 2, and 3; and Subgroup 2 includes Stations 4, 5, and 6. As illustrated in Figure 3.6, in our transient analysis, Subgroup i is characterized by the TFMs$^{[i]}$, a set of TFMs like the one in (3.23), with the superscript $^{[i]}$ denoting the group i ($i = 1, 2$). The input rate to the first group $a^{[1]}(t)$ is the input rate to the entire system $a(t)$, and the input rate to the second group $a^{[2]}(t)$ is the departure rate from the first group $d^{[1]}(t)$. The two sets of TFMs$^{[i]}$ ($i = 1, 2$), will be used to characterize the transient behavior of the system and to predict the system dynamics under any input $x(t)$.

The approach of decomposing a system into subgroups and characterizing each group by a set of TFMs is obviously approximate. The rationale behind this approximation is two fold. First, the transient effects at non-HUSs are negligible, that is, the time it takes for a non-HUS to reach steady state is negligible. Thus a subgroup can be considered as a whole with its behavior dominated by the sole HUS. Second, the implicit assumption made in modeling a subsequent group is that the departures from the previous group (i.e., the arrivals to this subsequent group) are approximately

completely characterized by the first moment measure, the departure rate. The practical validity of this assumption is assessed both theoretically and empirically. In [9], the CV of the interdeparture time, denoted as c_d, from a steady-state $G/G/1$ queue is derived analytically. When the queue is heavily loaded, c_d^2 can be approximated as $a \cdot c_s^2 + b$ with c_s being the CV of the service time for the $G/G/1$, and a and b are constants. Thus, the CV of the interdeparture time c_d is considered as fixed for a given station. Assume that the first two moments of the interdeparture time, i.e., departure rate plus c_d, are adequate to describe the departure process, then the departure rate alone gives us a relatively complete picture with c_d determined by the station parameter. Aside from the analytical approximation, large amount of empirical data for the interdeparture time from $G/G/s$ queues have been collected, which have shown that the first moment measure carries sufficient information regarding the departures.

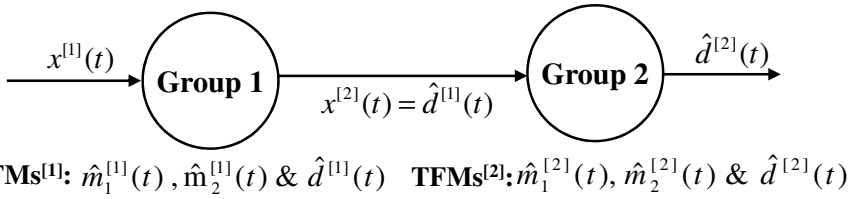


Figure 3.6: Decomposition of six stations in tandem system

.

Modeling Results for the Multi-Station System

The author presents the modeling results of the six-station system which is decomposed into two subgroups as shown in Figure 3.6. The EDS was obtained by simulating the system with Poisson arrivals at a piecewise-constant rate $x(t)$ similar to that in Figure 3.3. A total of 20000 simulation replications were performed, and the time series data $\{X^{[i]}(t), \mathbf{Y}^{[i]}(t), t = 1, 2, \ldots\}$ were collected for the fitting of TFMs$^{[i]}$ $(i = 1, 2)$. The resulting two sets of TFMs$^{[i]}$ $(i = 1, 2)$ can be used to predict the system performance under any input $X^*(t)$ $(t \in (0, T])$, and the prediction consists of two steps corresponding to two subgroups.

1. With $X^{[1]}(t) = X^*(t)$ and the identified history for Group 1, the TFMs$^{[1]}$ are used to recursively compute $\widehat{\mathbf{Y}}^{[1]}(t) = (\widehat{m}_1^{[1]}(t), \widehat{m}_2^{[1]}(t), \widehat{d}^{[1]}(t))$ for $t \in (0, T]$.

2. Given $X^{[2]}(t) = \widehat{d}^{[1]}(t)$ and the identified history for Group 2, the TFMs$^{[2]}$ are then used to recursively compute $\widehat{\mathbf{Y}}^{[2]}(t) = (\widehat{m}_1^{[2]}(t), \widehat{m}_2^{[2]}(t), \widehat{d}^{[2]}(t))$ for $t \in (0, T]$.

The goodness of the fitted TFMs$^{[i]}$ $(i = 1, 2)$ is evaluated based on the VDS, which is obtained by simulating the system with Poisson arrivals following the piece-wise constant rate given in Figure 3.7. From the VDS, time series $\mathbf{y}^{[i]}(t) = (m_1^{[i]}(t), m_2^{[i]}(t), d^{[i]}(t))$ $(i = 1, 2)$ were obtained, and considered as the "true" dynamic outputs. In Figure 3.8, comparing $\widehat{\mathbf{y}}^{[i]}(t)$, the predicted outputs from the TFMs which are represented by the

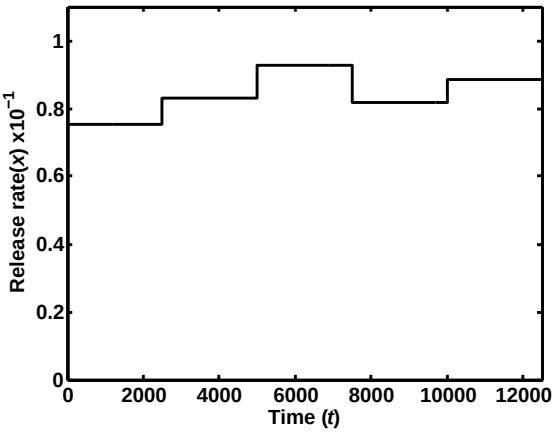


Figure 3.7: Release rate of validation data for six stations in tandem system

.

solid curves, and the "true" system evolution $\mathbf{y}^{[i]}(t)$ $(i = 1, 2)$ which are denoted as the dotted curves. Evidently, the TFMs$^{[i]}$ $(i = 1, 2)$ can accurately predict the dynamic outputs of this six-station system.

3.5.4 A Jackson Network System with Failures

The Jackson network models proposed by James Jackson in 1963 [49] have been recognized as one of the "Ten Most Influential Titles of Management Sciences First Fifty Years" by *Management Science* [50]. The Jackson network is a job-shop like queueing system and has been one of the most widely studied systems in queueing network theory. The Jackson network model can be described as: a network consisting

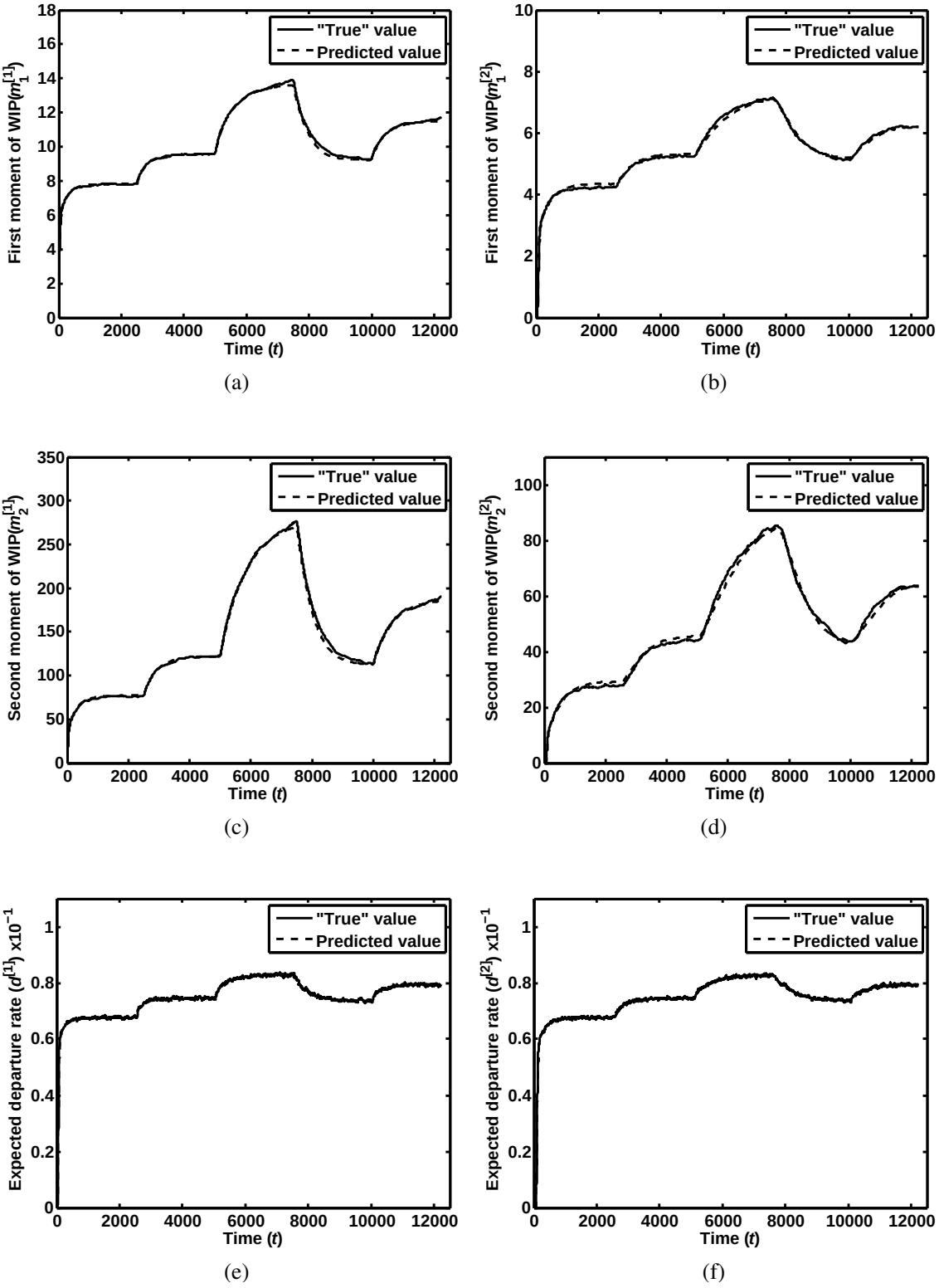


Figure 3.8: Evaluation of the fitted TFMs for the tandem system.

of J nodes; Each node $j, 1 \leq j \leq J$, has an infinite buffer and s_j identical servers, and the service time of server j follows exponential with a rate of μ_j; External jobs arrive at node j as a Poisson process of rate λ_j. And when a job completes service at node j, it immediately joins the queue at node k with probability p_{jk}, $1 \leq k \leq J$, and leaves the network with probability $p_j^* = 1 - \sum_{k=1}^{J} p_{jk}$.

According to Jackson theory [49], when the utilization at each node is less than 1, the equilibrium state probability distribution exists and for state $(k_1, k_2, \ldots, k_J)$ is given by the product of the individual queue equilibrium distribution:

$$\pi(k_1, k_2, \ldots, k_J) = \Pi_{j=1}^{J} \pi(k_j) \tag{3.24}$$

However, the Jackson theory (3.24) only applies to steady-state systems. In this subsection, the proposed method is applied on a general Jackson network model with failures. The Jackson network considered includes 4 stations. And each station has multiple machines which are subject to random failures. The following table shows the system parameters and the probability matrix of the system. Figure 3.9 illustrates the product flows with the expected flow rates. All jobs enters system by station 1 and leave system from station 3 or 4.

Following the same experimental design strategies (section 3.3.2), the EDS was obtained from simulation. The entire system were treated as a whole, and a single

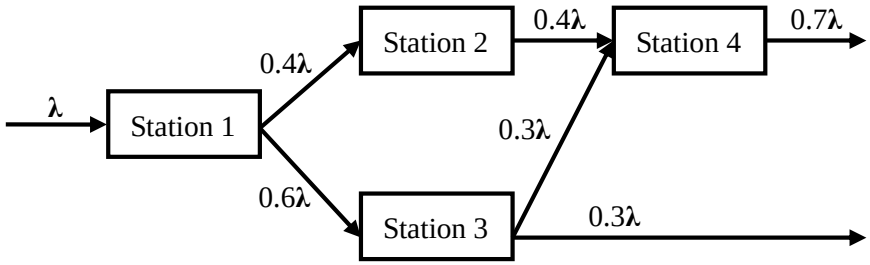


Figure 3.9: The flow diagram of the Jackson network

.

set of TFMs as those in (3.23) were fitted to describe the system evolution. Note that for the Jackson network, the output performance includes four components: $m_1(t)$ and $m_2(t)$, the first and second moments of the WIP in the system; $d_3(t)$ and $d_4(t)$, the departure rate from station 3 and 4 respectively. An independent VDS was obtained for the evaluation of the fitted TFMs. The results predicted by the TFMs are compared with the "true" system evolution. The comparison is shown in Figure 3.10. Evidently, the fitted TFMs are able to capture the transient dynamics of such a Jackson network.

It is worth mentioning that the decomposition described in Section 3.5.3 can also be applied to the Jackson network. For instance, the network can be decomposed into 4 subgroups (or workstations), and each subgroup can be described by a set of TFMs. The departures from an upstream subgroup serve as the arrivals to the downstream

group.

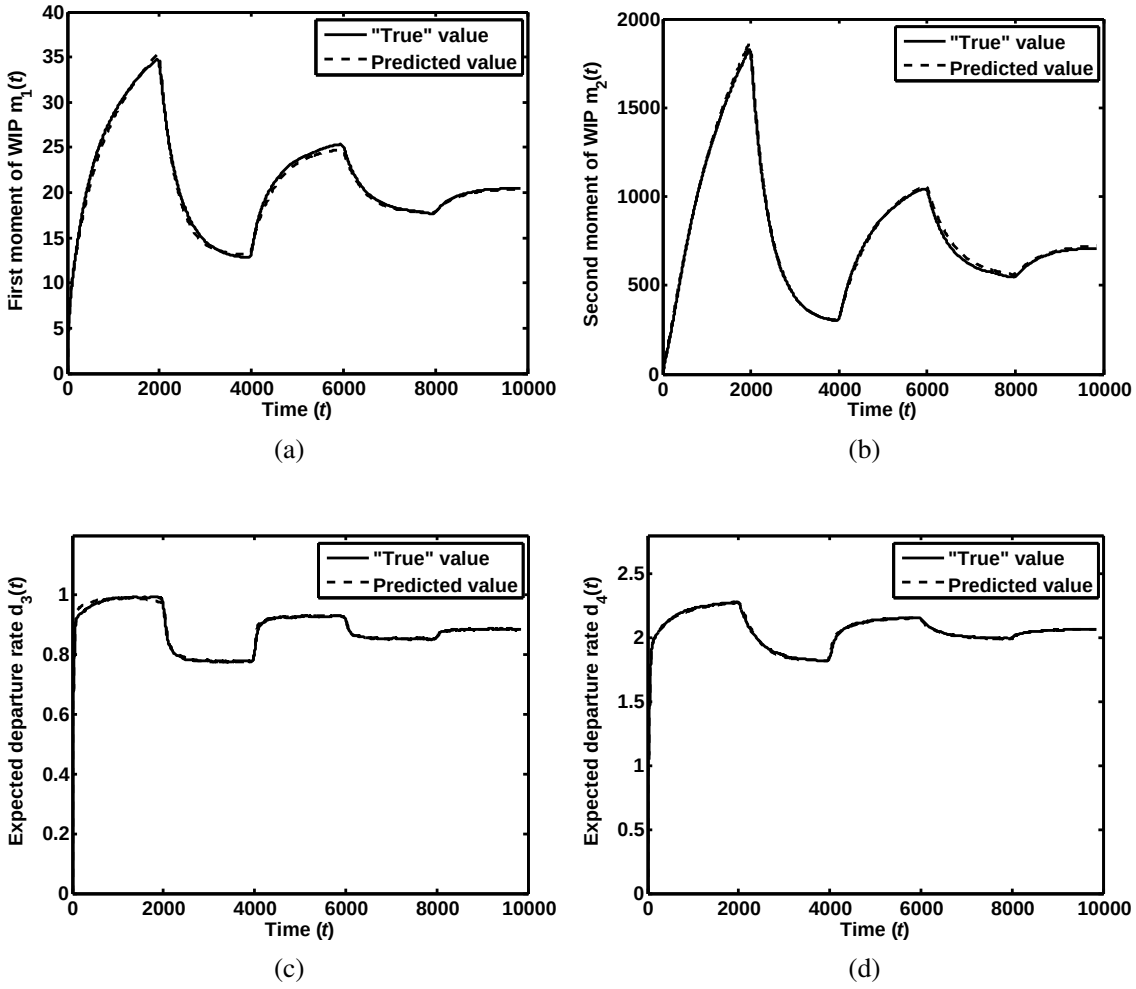


Figure 3.10: Evaluation of the fitted TFMs for the Jackson network

Table 3.3: The configuration of the Jackson network.

	Station 1	Station 2	Station 3	Station 4
# of Machines	3	3	3	3
Service Rate	0.3	0.1	0.15	0.14
MTTF	900	900	900	900
MTTR	100	100	100	100
Failure time	Exponential	Exponential	Exponential	Exponential
Repair time distribution	Exponential	Exponential	Exponential	Exponential

Chapter 4

TFMs-Based Production Planning

Chapter 3 provided simulation-based statistical methods to characterize the transient behavior of a manufacturing system by TFMs. In this part, the proposed TMFs will be integrated into the production planning optimization, which can be loosely defined as the problem of finding the best release plan of jobs so that the actual outputs overtime satisfy the predetermined requirements [66].

The remainder of this chapter is organized as follows. In Section 4.1, the production planning optimization problem is defined and formulated. Section 4.2 utilizes the input-output dynamics described by the TFMs to relate the optimization objective as a function of the system's input release. A numerical example is provided in Section 4.4.

4.1 Formulation of Production Planning Problem

Following the existing production planning models, the optimization model for production planning was formulated. Again, suppose that the time interval Δt determined in the transfer function modeling method (section 3.4) serves as the time unit. All the times involved in production planning are measured in such a time unit (i.e., Δt). The planning horizon is denoted as $(0, T]$ with the length being T time units. As in the existing production planning models, the planning horizon is divided into a number of, say P, time period. The p^{th} time period is represented by $(s_p, e_p]$, where s_p denotes the starting time, and e_p the ending time of planning period p; both are measured in terms of the time unit Δt.

Parameters:

w_p**:** WIP holding cost per time unit in period p.

h_p**:** Finished goods (FG) inventory holding cost per time unit in period p.

b_p**:** Backlogging cost per time unit in period p.

D_p**:** Demand quantity in period p; random variable that follows a pre-specified distribution obtained from forecasting models which are outside the scope of this work. It is assumed that all demands are realized at the end of each planning

period.

Independent decision variables:

The decision problem is to find the best or near optimum production plan represented by the release rates over the planning horizon: $\{x(t);\ t = 0, 1, \ldots, T\}$. It is assumed that for $t \in (s_p, e_p]$, $x(t) = x_p$; that is, the release rate within each planning period is constant. Thus, $\mathbf{x} = \{x_p; p = 1, 2, \ldots, P\}$ are the decision variables to be determined in the plan optimization. Denote R_p as the number of jobs to be released into the system for processing in period p; R_p is closely associated with the decision variable:

$$R_p = \sum_{t=s_p}^{e_p} x_t = (e_p - s_p) \times x_p \tag{4.1}$$

Dependent variables:

Z_p: Quantity of products produced in period p, which is the sum of $\{Z(t); t \in (s_p, e_p]\}$, with $Z(t)$ being the products produced within a time unit t. Therefore,

$$Z_p = \sum_{t=s_p}^{e_p} Z(t). \tag{4.2}$$

W_p: Cumulative WIP in period p, which depends on $\{Q(t); t \in (s_p, e_p]\}$, the number of WIPs in the system over production planning period p:

$$W_p = \sum_{t=s_p}^{e_p} Q(t) \tag{4.3}$$

I_p: Cumulative inventory level of FG at the end of planning period p. Let I_0 represents the initial FG inventory, then the inventory level at the beginning of each period can be written as follows:

$$I(s_1) = I_0 \tag{4.4}$$

$$I(s_p) = \max\{0, I(s_{p-1}) + \sum_{t=s_{p-1}}^{e_{p-1}} Z(t) - B_{p-1} - D_{p-1}\} \tag{4.5}$$

where B_{p-1} is defined as following.

B_p: The quantity of FG that cannot be satisfied on time at the end of planning period p. The unmet demand is considered as backlog, and will be fulfilled at the end of nearest planning period when enough FG is available. B_p can be written as:

$$B_p = -\min\{0, I(s_{p-1}) + \sum_{t=s_p}^{e_p} Z(t) - B_{p-1} - D_p\} \tag{4.6}$$

Hence, given the initial value I_0 and B_0, $\{I_p, B_p; p = 1, 2, \ldots, P\}$ can be obtained recursively depending on the products produced and customer demand during the planning horizon.

It is assumed that after the planning horizon, demands for each product type will continue forever following the same rate as in the last planning period [47]. An extra post-planning period is added, during which the demands are satisfied by the

products that are released but not finished during the planning horizon. This extra period lasts until all the release during the planning horizon have been completed.

Thus, the cost objective function associated with production plan $\mathbf{x}$ is:

$$L(\mathbf{x}) = \sum_{p=1}^{P+1} w_p W_p + \sum_{p=1}^{P+1} h_p I_p + \sum_{p=1}^{P+1} b_p B_p \tag{4.7}$$

which is a random variable whose distribution depends on the decision $\mathbf{x}$. The total cost consists of three parts: the WIP holding cost, the inventory (FG) holding cost and the backlog cost. The purpose of production planning is to minimize the mean and variance of the total cost with respect to the release plan $\mathbf{x}$.

4.2 Evaluation of the Total Cost Objective

The basis of the planning optimization lies in the ability to evaluate the total cost objective for any release plan $\mathbf{x}$. Next how to use the TFMs to obtain the three parts of the total cost $L(\mathbf{x})$ will be discussed respectively.

WIP holding cost

As can be seen from (4.3), the WIP holding cost $\sum_{p=1}^{P+1} w_p W_p$ depends on $Q(t)$. For a plan $\mathbf{x}$, the TFMs provide $m_1(t) = \mathrm{E}[Q(t)]$, and thus can be directly used to estimate the first moment of the cumulative WIP holding cost within a planning period.

The variance of cumulative WIP holding cost within a planning period, which can

be approximated as:

$$\text{var}[\sum_{t=s_p}^{e_p} Q(t)] \approx \sum_{t=s_p}^{e_p} \text{var}[Q^2(t)] + 2\sum_{t=s_p}^{e_p-1} \text{cov}[Q(t), Q(t+1)]. \tag{4.8}$$

In (4.8), $\text{var}[Q^2(t)] = m_2(t) - (m_1(t))^2$ can be obtained from the outputs of the TFMs (3.1). The component in the covariance $\text{cov}[Q(t), Q(t+1)] = \text{E}[Q(t)Q(t+1)] - m_1(t) \cdot m_1(t+1)$ yet to be specified is $cv(t) = \text{E}[Q(t)Q(t+1)]$. By nature, $cv(t)$ is similar to the second moment $m_2(t)$, and a transfer function model like that for $m_2(t)$ can be estimated from simulation data to describe the evolution of $cv(t)$ as well using the methods in Chapter 3. Note that in 4.8, only the covariances between successive WIPs are considered, which is a good approximation based on the author's empirical experience.

FG holding and backlog cost

The FG holding cost and backlog cost are closely related to the output $\{Z_p; p = 1, 2, \ldots, P\}$, as shown in (4.5) and (4.6). The key to estimate the first two moments of the FG holding and backlog cost is to characterize the random variables $\{Z_p; p = 1, 2, \ldots, P\}$. In this work, the following assumptions are made in this regard. (i) Z_p is approximately normally distributed; since it is the sum of a relatively large number of random variables (4.2), the normal assumption can be justified by the large sample theory [56]. (ii) Z_i and Z_j are uncorrelated with each other for $i \neq j$, which is

a practically reasonable assumption for non-overlapping time buckets. Appendix C provides the details on how to derive the distribution (i.e., mean and variance) of Z_p from the TFMs.

With $\{Z_p; p = 1, 2, \ldots, P\}$ specified by P non-identical and independent normal distributions, the FG holding and backlog cost can be evaluated by performing Monte Carlo simulation. More specifically, for a demand scenario $\{D_p; p = 1, 2, \ldots, P\}$, a realization of $\{Z_p; p = 1, 2, \ldots, P\}$ can be generated from their distributions, and hence a realization of the cost can be calculated. From the realizations of the cost, the mean and variance of the cost can be estimated.

Therefore, using the TFMs obtained from the previous chapter, the total cost involved for any release plan $\mathbf{x}$ can be numerically evaluated.

4.3 Multi-Objective Optimization for Production Planning

The optimization problem for production planning aims at minimizing two objective functions, $\mathrm{E}[L(\mathbf{x})]$ and $\mathrm{Var}[L(\mathbf{x})]$, with respect to $\mathbf{x}$, and thus is a multi-objective problem (MOP). For an MOP, there generally does not exist any single solution which can provide the optimal value on all the objectives, and thus it is of interest

to generate a set of non-dominated solutions, where no objective can be improved without worsening at least one other objective. The set of all non-dominated solutions is referred to as the Pareto optimal front [18], and our goal is to obtain a set of solutions for $\mathbf{x}$ as close as possible to the Pareto optimal front.

In this work, the Elitism Non-Dominated Sorting GA (NSGA-II) [18] provided by the Matlab Optimization toolbox was adopted to solve the MOP in search of the optimum production plan. The NSGA-II has the following features, which makes it a standard and most widely used algorithm in solving an MOP [50].

- An efficient sorting procedure is embedded in NSGA-II to rank the candidate solutions based on multiple criteria.
- In the multi-objective search of NSGA-II, an elitism-preserving approach is developed which enhances the convergence toward the true Pareto optimal set.
- A parameterless diversity preservation mechanism is adopted to ensure the diversity and spread of solutions and to guide the search toward a uniformly spread Pareto frontier.
- The constraint handling method does not rely on the use of penalty parameters. The algorithm implements a modified definition of dominance in order to solve constrained multi-objective problems efficiently.

- NSGA-II can handle both continuous and discrete design variables.

For NSGA-II, the user-specified parameters are: population size, mutation probability, elite number, maximum generation performed, and termination tolerance.

4.4 Numerical Results

The empirical system studies is the same as 3.5.3. The system composed of 6 stations. Each station has multiple identical machines and each machine subject to random failures. Both the machine failure time and repair time follow Exponential distribution with known mean value. The Production planning horizon is 15000 time units (Δt), with five planning periods plus an extra planning periods as discussed in 4.1. The decision variables are $\mathbf{x} = \{x_1, x_2, \ldots, x_5\}$. Three demand profile scenarios are considered: stable demand, increasing-decreasing demand and fluctuating demand pattern. For each case, the Pareto frontier is found and the associated decision variable are tabled. The WIP holding cost, FG holding cost and backlog cost per unit time used are 1, 2 and 5. This cost ratio represents the typical cost ratio for semiconductor industrial as recommended by an anonymous reviewer.

To evaluate the accuracy of the estimated total cost from the TFMs based approach (plus Monte Carlo simulation), extensive discrete-event simulation was per-

formed which is considered to be able to provide the "true" cost incurred by a production plan. Specifically, the TFMs-based optimization problem was solved by the multi-objective GA. The resulting release plans (i.e., the solutions) were fed to the detailed discrete-event simulation model for the evaluation of the total cost. The mean and standard deviation of the total cost obtained from the TFMs approach were compared with those from the discrete-event simulation.

4.4.1 Stable Demand Case

For this case, the mean demands for five planning periods stay constant and are given as $\mathbf{d} = \{180, 180, 180, 180, 180\}$. The demand within each planning period follows a Normal distribution, with standard deviation equals one-tenth of the mean. Applying the multi-objective GA algorithm with 1000 population size and 15 generation, the solution set obtained is plotted in figure 4.1. The detailed optimization results are shown in table 4.1. Each row represents one production plan, which includes five decision variables, $\mathbf{x} = \{x_1, \ldots, x_5\}$. The mean and standard deviation of the total cost estimated from the TFMs based approach and simulation are listed following the decision variables. Comparing with the simulation results, the TFMs estimated results are very close to the simulations. For all the release plans, the difference are within 5% for mean total cost and 10% for the standard deviations.

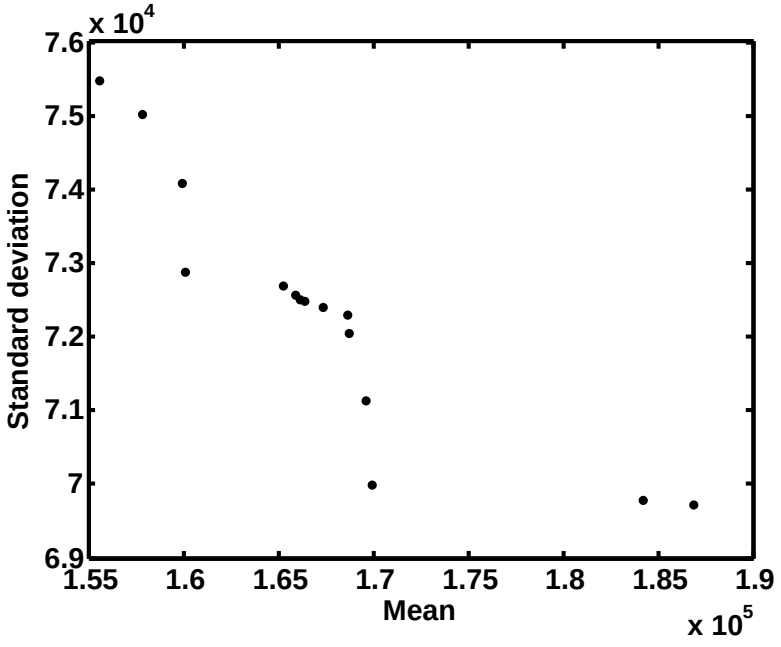


Figure 4.1: The solution set obtained from the multi-objective optimization for production planning (constant demand case).

Table 4.1: Multi-objective optimization solutions for the constant demand case.

Decision variables					TFMs		Simulation	
$x1$	$x2$	$x3$	$x4$	$x5$	Mean	Stdev	Mean	Stdev
0.7651	0.7618	0.7583	0.8348	0.7888	184209	69781	185813	73481
0.8257	0.7560	0.7501	0.7596	0.7089	157824	75006	167809	84214
0.7646	0.7588	0.7600	0.7530	0.7472	155568	75471	170239	92093
0.8340	0.7549	0.7508	0.7531	0.7146	159921	74087	170061	84366
0.7644	0.7602	0.7588	0.8003	0.7574	166205	72496	178433	86260
0.8266	0.7561	0.7489	0.7567	0.7402	160160	72881	170436	81046
0.7651	0.7597	0.7590	0.8263	0.7570	169984	69979	185753	80200
0.7646	0.7594	0.7599	0.8211	0.7368	165892	72560	178598	84356
0.7642	0.7592	0.7591	0.8025	0.7662	169636	71127	181406	84396
0.7641	0.7594	0.7574	0.7992	0.7625	168711	72041	178368	79014
0.7650	0.7571	0.7552	0.8286	0.8028	186915	69724	182548	72024
0.7640	0.7591	0.7584	0.7929	0.7565	165320	72683	170177	78057
0.7646	0.7586	0.7568	0.8076	0.7526	166378	72480	177440	84518
0.7655	0.7600	0.7555	0.8215	0.7451	168648	72300	178415	82010
0.7652	0.7585	0.7574	0.8177	0.7524	167387	72387	175845	77437

4.4.2 Increasing - Decreasing Demand Case

Now, assuming the demand pattern first increases, and reaches the peak value at the middle planning period. Then, the demand fall back. The mean demands are $\mathbf{d} = \{170, 190, 210, 190, 170\}$. The parameters for the GA algorithm as set as before. The similar Pareto frontier can be found as well (Figure 4.2). The detailed optimization results are listed in table 4.2.

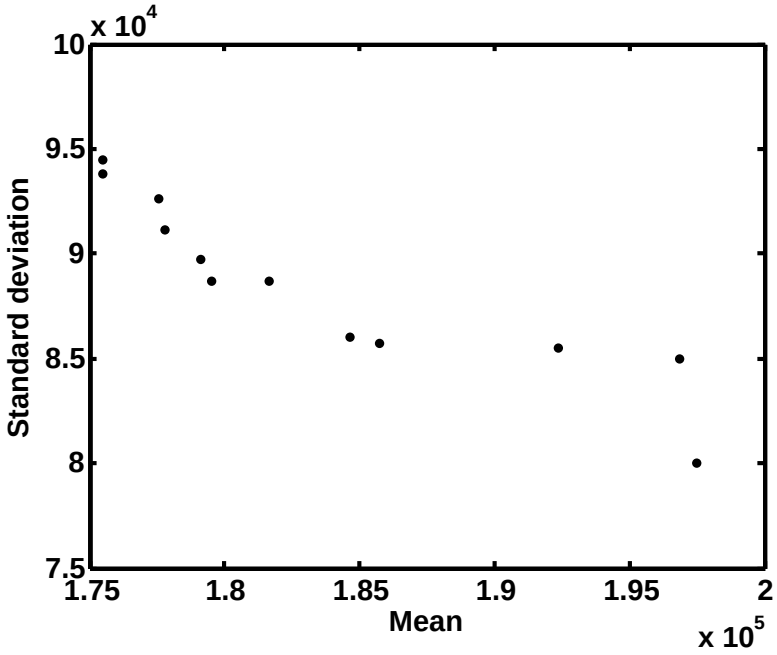


Figure 4.2: The solution set obtained from the multi-objective optimization for production planning (increasing-decreasing demand case).

.

Table 4.2: Multi-objective optimization solutions for the increasing-decreasing demand case.

Decision variables					TFMs		Simulation	
$x1$	$x2$	$x3$	$x4$	$x5$	Mean	Stdev	Mean	Stdev
0.8432	0.7371	0.8522	0.7376	0.7108	175499	94483	173368	90775
0.8428	0.6925	0.8605	0.7851	0.7857	192387	85475	182625	78895
0.8386	0.7252	0.8529	0.7543	0.8144	196843	84959	190719	80889
0.8445	0.7331	0.8618	0.7565	0.8178	197472	79992	194554	82992
0.8436	0.7397	0.8690	0.7569	0.7298	179546	88710	171461	76076
0.8443	0.7237	0.8521	0.7961	0.7211	179106	89757	174890	84695
0.8380	0.7276	0.8699	0.7270	0.7225	175513	93774	176843	90623
0.8292	0.7507	0.8536	0.7694	0.7399	181678	88695	180600	85389
0.8377	0.7427	0.8651	0.7551	0.7125	177569	92588	173930	86351
0.8396	0.7134	0.8577	0.7695	0.7654	184677	86053	175878	82763
0.8391	0.7268	0.8559	0.7868	0.7585	185750	85715	182222	81407
0.8434	0.7335	0.8510	0.7471	0.7403	177802	91140	174837	81695

4.4.3 Fluctuating Demand Case

The last demand pattern consider is a fluctuating scenario. The demand repeats at low and high two levels. The mean demands are $\mathbf{d} = \{160, 210, 160, 210, 160\}$. The parameters for the GA algorithm as set as before. The Pareto frontier is shown in Figure 4.3. Table 4.3 shows the detailed optimization results.

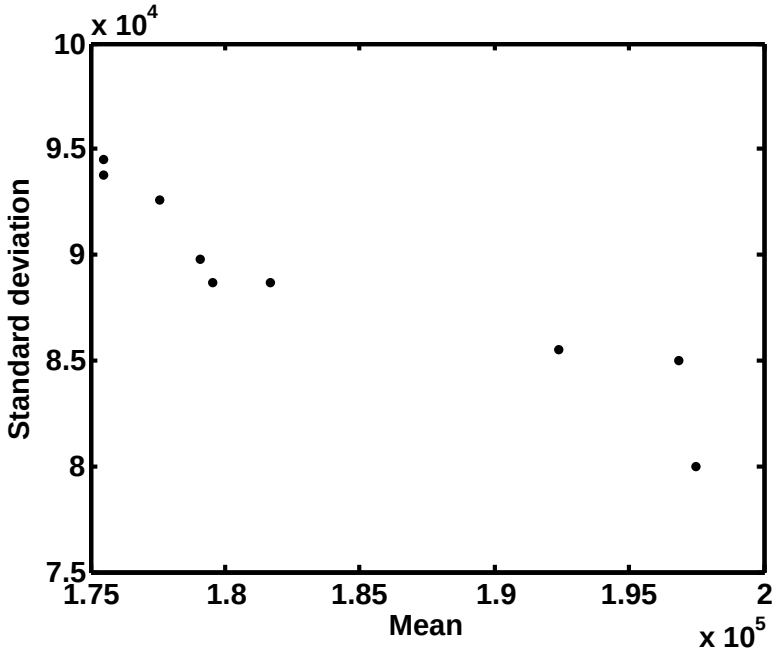


Figure 4.3: The solution set obtained from the multi-objective optimization for production planning (fluctuating demand case)

.

Table 4.3: Multi-objective optimization solutions for the fluctuating demand case

Decision variables					TFMs		Simulation	
$x1$	$x2$	$x3$	$x4$	$x5$	Mean	Stdev	Mean	Stdev
0.7129	0.8518	0.7142	0.8523	0.5974	168100	97110	177763	102296
0.7270	0.8497	0.7402	0.8579	0.6407	173671	86749	171094	81097
0.7212	0.8622	0.7153	0.7929	0.7842	194470	78080	184421	75951
0.8304	0.7106	0.8426	0.6281	0.7920	196368	76796	189949	78271
0.8165	0.7045	0.8411	0.7120	0.7120	181710	84438	178100	79958
0.8249	0.7172	0.8646	0.7096	0.6957	183819	82957	179122	73793
0.7156	0.8535	0.7549	0.8548	0.6404	174695	85397	173409	76375
0.8391	0.7055	0.8346	0.6634	0.8102	203650	75863	193339	73212
0.8397	0.7354	0.8735	0.7176	0.5950	180967	85374	180768	79258
0.7507	0.8663	0.6559	0.8201	0.6719	168875	90031	171892	93884

Chapter 5

Conclusions and Future Work

5.1 Conclusions

The originality of this work lies in the integration of statistical methods, computer simulation, and queueing theory to tackle the ever-difficult yet critical research problem of characterizing the transient behavior of general queueing systems. Such an approach is expected to overcome the computational burden of simulation and the intractability of analytical methods for general queues.

The resulting TFMs from the proposed method are able to describe system dynamics and have two advantages. First, the TFMs embody the high fidelity of simulation since they are estimated from detailed simulation data. Second, the TFMs

are difference equations, like the discrete approximations of the ordinary differential equations provided by an analytical approach; supposing that a certain input is fed to the system under given initial conditions, the TFMs can be used to recursively compute the system's future performance in a timely manner. The TFMs are able to predict not only the first moment but also the second moment measure of the system performance. To efficiently generate such TFMs for queueing systems, analytical queueing analysis were performed to suggest appropriate functional forms of the TFMs; experimental design strategies were developed to efficiently collect data via offline simulation; and statistical TFM fitting methods were developed to obtain well-estimated TFMs from simulation data.

The proposed TFMs have been tested on a variety of transfer line queueing systems, ranging from single station with single machine, to multi-station with multi-machine, systems with machine failures, reentrant flows and non-Markov arrival and service time scenario. The TFMs can accurately quantify the dynamic relationships between release rate and output performances for the tested transfer line type queueing system. The proposed methodology has also been successfully tested on Jackson network type queueing system, which is parallel to the job shop type manufacturing system.

Utilizing the dynamics captured by the TFMs, a new production planning method

has been developed which is able to minimize the mean as well as the variance of the total cost with respect to the production plan. The variance of the total cost is critical to evaluate the risk of the production plan, and our method is the first one that allows the consideration of the cost variance in a timely manner.

5.2 Future Work

5.2.1 Transient Analysis

The major limitation of the transfer function modeling approach for transient analysis lies in the assumption that the time-varying arrival process to the system can be fully characterized by its first moment measure, that is, the arrival rate. The TFMs take the time-varying arrival rate as input, and predict the evolution of system outputs. This assumption obviously holds for arrivals with exponential or constant interarrival times, which cover a wide range of applications in manufacturing and service. However, for more general arrivals, using only the rate to characterize the arrival processes is certainly a restrictive approximation. Hence, the future work will focus on extending the current approach to incorporate into the TFMs the second moment information of the arrival process.

The proposed transient analysis method has been evaluated by applying it to the

transfer lines and job shops. In the future, application to other real manufacturing systems will be performed. Systems involving labor factor will be considered. Including the labor factor into simulation model will require more deliberated simulation modeling of the manufacturing system. This more detailed simulation model for such systems is expected to be developed and the TFMs will be applied to demonstrate its applicability. The future work will be to apply the TFMs approach to handle queueing system with more realistic manufacturing features, such as batching and rework. Extending the TFMs to multi-product queueing systems will be also be studied. For a multi-product queueing system, there will be more input variables and output performances. Model fitting and parsimonious regressor selection for such high dimension TFMs will be studied as well.

5.2.2 Production Planning

The transient analysis method will be generalized to take into account the second moment information of the system arrivals and to analyze real manufacturing systems of higher complexity. The generalized TFMs will be incorporated into the production planning method, which will be extended to handle systems that involve multiple products, operators, batch processing, etc.

Appendix

Analytical Transient Analysis of a General Single-Server Queue

Following the notations in Sections 1.3 and 3.5.2, the equations (3.5) can be derived for a general single-server queue with orderly arrivals and departures. The service time of jobs follows distribution $G(\tau)$, $\tau \in (\tau_L, \tau_U)$.

The Kolmogorov forward equations for the state probabilities $p_n(t) = \Pr\{Q(t) = n, n = 0, 1, 2, ...\}$ are given as follows:

$$p'_n(t) = a_{n-1}(t) + d_{n+1}(t) - a_n(t) - d_n(t); \quad n \geq 1 \tag{a.1}$$

$$p'_n(t) = d_{n+1}(t) - a_n(t); \quad n = 0 \tag{a.2}$$

Multiplying both sides of equations (a.1) and (a.2) by n and taking the sum across

all values of n, will have

$$m_1'(t) = \frac{d\mathrm{E}[m_1(t)]}{dt} = \sum_{n=0}^{\infty} n \cdot p_n'(t) = a(t) - d(t) \tag{a.3}$$

which is the first equation in (3.5). Recall that $a(t)$ is the independent input variable representing the arrival rate of jobs, and $m(t)$ and $d(t)$ characterize the output processes of interest. Next, we proceed to derive the dynamic evolution of $d(t) = \sum_{n=1}^{\infty} d_n(t)$.

We first consider $d_n(t) = \lim_{\delta \to 0^+} \delta^{-1} \Pr\{D(t, t+\delta) = 1, Q(t) = n\}$. A departure will occur during the interval $(t, t+\delta]$ with n jobs in the system at time t if one of the two following conditions holds:

(i) The system was empty at time $t - \tau$, and onejob entered during the instant $(t-\tau, t-\tau+\delta]$. During the service of this job, which lasted for a period of τ, there were $n-1$ new arrivals to the system.

(ii) A departure occurred during the instant $(t-\tau, t-\tau+\delta]$ while there are $k \geq 2$ jobs in the system at time $t-\tau$. Immediately after the departure, the service for the first job in the queue was initiated and lasted for a period of τ. During the service of this job, $n-k+1$ new jobs entered the system.

Thus,

$$
\begin{aligned}
&\Pr\{D(t,t+\delta)=1, Q(t)=n\} \\
&= \int_{\tau_L}^{\tau_U} \Pr\{A(t-\tau,t-\tau+\delta)=1, Q(t-\tau)=0, A(t-x,t)=n-1\} dG(\tau) \\
&+ \int_{\tau_L}^{\tau_U} \sum_{k=2}^{n+1} \Pr\{D(t-\tau,t-\tau+\delta)=1, Q(t-\tau)=k, A(t-\tau,t)=n-k+1\} dG(\tau)
\end{aligned}
$$

and thus,

$$
\begin{aligned}
&\Pr\{D(t,t+\delta)=1\} \\
&= \int_{\tau_L}^{\tau_U} \sum_{n=1}^{\infty} \Pr\{A(t-\tau,t-\tau+\delta)=1, Q(t-\tau)=0, A(t-x,t)=n-1\} dG(\tau) \\
&+ \int_{\tau_L}^{\tau_U} \sum_{n=1}^{\infty} \sum_{k=2}^{n+1} \Pr\{D(t-\tau,t-\tau+\delta)=1, Q(t-\tau)=k, A(t-\tau,t)=n-k+1\} dG(\tau) \\
&= \int_{\tau_L}^{\tau_U} \Pr\{A(t-\tau,t-\tau+\delta)=1, Q(t-\tau)=0\} dG(\tau) \\
&+ \int_{\tau_L}^{\tau_U} (\Pr\{D(t-\tau,t-\tau+\delta)=1\} - \Pr\{D(t-\tau,t-\tau+\delta)=1, Q(t-\tau)=1\}) dG(\tau)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
d(t) &= \lim_{\delta \to 0^+} \delta^{-1} \Pr\{D(t,t+\delta)>0\} \\
&= \int_{\tau_L}^{\tau_U} a_0(t-\tau) dG(\tau) + \int_{\tau_L}^{\tau_U} (d(t-\tau) - d_1(t-\tau)) dG(\tau),
\end{aligned}
$$

which is the second equation in (3.5).

Statistical Inference on the TFMs

The notation used in Section 3.4 is inherited here. For convenience of the discussion, the transformed models (3.17) and (3.17) are rewritten as follows:

$$\widetilde{Y}_1(t) = \widetilde{F}_1(\boldsymbol{\theta}_1, \widetilde{\mathcal{H}}_t) + w(t) \tag{a.4}$$

$$\widetilde{Y}_2(t) = \widetilde{F}_2(\boldsymbol{\theta}_2, \widetilde{\mathcal{H}}_t) + w(t) \tag{a.5}$$

where $\widetilde{Y}_i(t) = \frac{D_i(q)}{\sigma_i C_i(q)} Y_i(t)$ $(i = 1, 2)$, the transformed function $\widetilde{F}_i = \frac{D_i(q)}{\sigma_i C_i(q)} F_i$ $(i = 1, 2)$, and $w(t)$ is the white noise with variance 1. The history of the system prior to time t is denoted as $\widetilde{\mathcal{H}}_t = \{X(\tau), \mathbf{Y}(\tau), \tau < t\}$.

Suppose that T "time series pairs" $\{X(t), \mathbf{Y}(t), t = 1, \ldots, T\}$ have been obtained for the estimation of the models (a.4 and a.5). Defining the additional notation as follows:

- $\mathbf{f}_1(\boldsymbol{\theta}_1) = (\widetilde{F}_1(\boldsymbol{\theta}_1, \widetilde{\mathcal{H}}_1), \widetilde{F}_1(\boldsymbol{\theta}_1, \widetilde{\mathcal{H}}_2), \ldots, \widetilde{F}_1(\boldsymbol{\theta}_1, \widetilde{\mathcal{H}}_T))'$ is a $T \times 1$ vector function of $\boldsymbol{\theta}_1$.

- $\mathbf{f}_2(\boldsymbol{\theta}_2) = (\widetilde{F}_2(\boldsymbol{\theta}_2, \widetilde{\mathcal{H}}_1), \widetilde{F}_2(\boldsymbol{\theta}_2, \widetilde{\mathcal{H}}_2), \ldots, \widetilde{F}_2(\boldsymbol{\theta}_2, \widetilde{\mathcal{H}}_T))'$ is a $T \times 1$ vector function of $\boldsymbol{\theta}_2$.

- $\mathbf{f}_3(\boldsymbol{\theta}_3) = (\widetilde{F}_3(\boldsymbol{\theta}_3, \widetilde{\mathcal{H}}_1), \widetilde{F}_3(\boldsymbol{\theta}_3, \widetilde{\mathcal{H}}_3), \ldots, \widetilde{F}_3(\boldsymbol{\theta}_3, \widetilde{\mathcal{H}}_T))'$ is a $T \times 1$ vector function of $\boldsymbol{\theta}_3$.

- $\mathbf{D}_1(\widehat{\boldsymbol{\theta}_1}) = \partial \mathbf{f}_1(\widehat{\boldsymbol{\theta}_1})/\partial \boldsymbol{\theta}_1'$ is a $T \times N_1$ first-derivative matrix, where N_1 is the dimension of $\boldsymbol{\theta}_1$.

- $\mathbf{D}_2(\widehat{\boldsymbol{\theta}_2}) = \partial \mathbf{f}_2(\widehat{\boldsymbol{\theta}_2})/\partial \boldsymbol{\theta}_2'$ is a $T \times N_2$ first-derivative matrix, where N_2 is the dimension of $\boldsymbol{\theta}_2$.

- $\mathbf{D}_3(\widehat{\boldsymbol{\theta}_3}) = \partial \mathbf{f}_3(\widehat{\boldsymbol{\theta}_3})/\partial \boldsymbol{\theta}_3'$ is a $T \times N_3$ first-derivative matrix, where N_3 is the dimension of $\boldsymbol{\theta}_3$.

- The design matrix $\mathbf{D}$ is defined as:

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_1(\widehat{\boldsymbol{\theta}_1}) & 0 & 0 \\ 0 & \mathbf{D}_2(\widehat{\boldsymbol{\theta}_2}) & 0 \\ 0 & 0 & \mathbf{D}_3(\widehat{\boldsymbol{\theta}_3}) \end{pmatrix} \tag{a.6}$$

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$, then the estimated parameters $\widehat{\boldsymbol{\theta}}$ is approximately normally distributed with the variance-covariance matrix:

$$\widehat{\mathrm{Var}}[\widehat{\boldsymbol{\theta}}] = \sigma^2(\mathbf{D}'\mathbf{D})^{-1} \tag{a.7}$$

where $\sigma^2 = \mathrm{Var}[w(t)] = 1$ for models (a.4 and a.5).

Second Moment of Cumulative Output

All symbols are the same as defined in section 4.1. The quantity of products produced in a single planning period, say period p, can be expressed as $Z_p = \sum_{t=s_p}^{e_p} Z(t)$. Both sides taking variance and expanding the right hand side will have the following:

$$\begin{aligned}
\text{var}[Z_p] &= \text{var}[\sum_{t=s_p}^{e_p} Z(t)] \\
&= \text{var}[\sum_{t=s_p}^{e_p} R(t) + Q(s_p) - Q(e_p) \\
&= \text{var}[\sum_{t=s_p}^{e_p} R(t)] + \text{var}[Q(s_p)] + \text{var}[Q(e_p)] \\
&\quad +2\text{cov}[\sum_{t=s_p}^{e_p} R(t), Q(s_p)] - 2\text{cov}[\sum_{t=s_p}^{e_p} R(t), Q(e_p)] - 2\text{cov}[Q(s_p), Q(e_p)]
\end{aligned}$$

Where $\text{cov}[\sum_{t=s_p}^{e_p} R(t), Q(s_p)] = 0$. This is because the quantity of raw material released within the p^{th} planning period can not affect the WIP level at the beginning of that planning period. When the planning period length is fairly long, comparing to the pure processing time, the last term in equation 5.1 will be approximately equals zeros as well. Since, the WIP level at the starting of the planning period has no or very limited effect on the WIP level at the end of the same planning period. If the planning period is long enough, the system will reach stead state at the end of that planning period. Since, no matter what's the initial WIP level, the steady state expected WIP will be fixed for a specific release plan.

The term, $\text{cov}[\sum_{t=s_p}^{e_p} R(t), Q(e_p)]$, reflects the relation between release within the p^{th} planning period and the WIP level at the end of that planning period. Since, the expected release rate is constant during the time interval $(s_p, e_p]$ (section 4.1). For a fairly long planning period, the system will reach steady state by the end of the planning period e_p. The $\text{E}[Q(e_p)]$ has a nonlinear relationship with the expected release rate or workload of the system. So, the author believe $\text{cov}[\sum_{t=s_p}^{e_p} R(t), Q(e_p)]$ only depends on release rate and $e_p - s_p$. Further broke down this covariance will get:

$$\text{cov}[\sum_{t=s_p}^{e_p} R(t), Q(e_p)] = \text{E}[\sum_{t=s_p}^{e_p} R(t) * Q(e_p)] - \text{E}[\sum_{t=s_p}^{e_p} R(t)] * \text{E}[Q(e_p)]$$

$\text{E}[\sum_{t=s_p}^{e_p} R(t) \cdot Q(e_p)]$ can be estimated from offline simulation. $\text{E}[\sum_{t=s_p}^{e_p} R(t)]$ is known and $\text{E}[Q(e_p)]$ can be estimated from the TFMs directly. Therefore, the cumulative output variance within one planning period can be estimated. Figure a.1 shows the simulation estimated function of $\text{E}[\sum_{t=s_p}^{e_p} R(t) \cdot Q(e_p)]$. Therefore, the expectation can be read from the function by given any expected release rate.

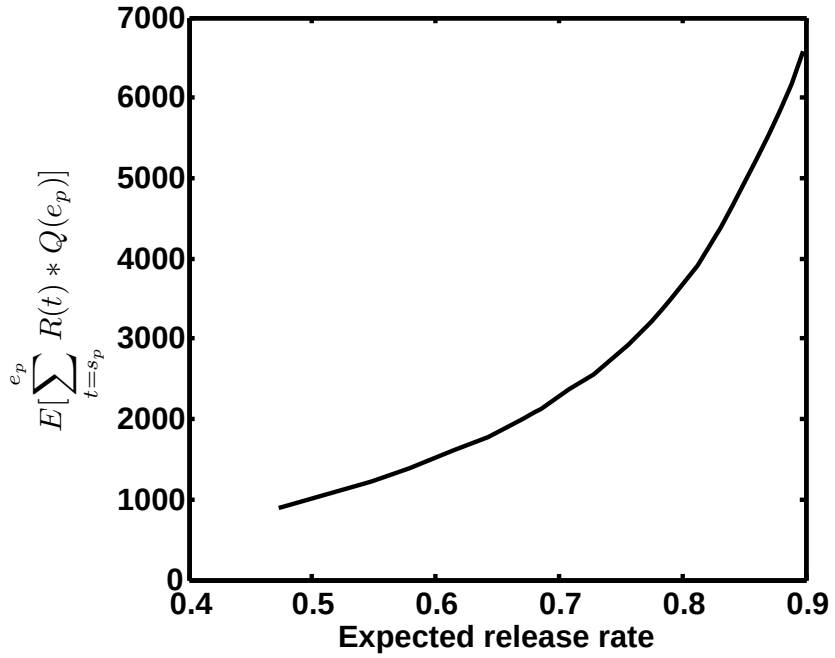


Figure a.1: The expectation term function for second moment estimation

.

References

1. Ankenman, B. E., J. M. Bekki, J. W. Fowler, G. T. Mackulak, B. L. Nelson and F. Yang. 2008. Simulation in Production Planning. Chapter 6 in: *Planning in the Extended Enterprise: A State of the Art Handbook* (eds. Kempf, K. G., P. Keskinocak and R. Uzsoy), to be published by Springer, New York.

2. Asmundsson, J. M., R. L. Rardin and R. Uzsoy. 2006. Tractable Nonlinear Production Planning Models for Semiconductor Wafer Fabrication Facilities. *IEEE Transactions on Semiconductor Manufacturing* **19**: 95–111.

3. Asmundsson, J. M., R. L. Rardin, R. Uzsoy, C.H. Turkseven. 2009. Production Planning Models with Resources Subject to Congestion. *Naval Research Logistics* **56**: 142–157.

4. Banker, R. D., S. Datar and S. Kekre. 1986. Relevant Costs, Congestion and Stochasticity in Production Environments. *Journal of Accounting and Eco-*

nomics **10**: 171–197. .

5. Bertsimas, D. and G. Mourtzinou. 1997. Transient Laws of Non-Stationary Queueing Systems and Their Applications. *Queueing Systems* **25** (1–4): 115–155.

6. Billington, P., J. O. McClain and L. J. Thomas. 1986. Mathematical Programming Approaches to Capacity-constrained MRP Systems: Review, Formulations and Problem Reduction. *Management Science* **32** (8): 989–1006.

7. Blackhurst, J., C. W. Craighead, D. Elkins and R. B. Handfield. 2005. An Empirically Derived Agenda of Critical Research Issues for Managing Supply–Chain Disruptions. *International Journal of Production Research* **43** (19): 4067–4081.

8. Box, E. P. G., G. M. Jenkins and G. C. Reinsel. 1994. *Time Series Analysis: Forecasting and Control*, 3rd Edition. Prentice Hall.

9. Buzacott, J. A. and J. G. Shanthikumar. 1993. *Stochastic Models of Manufacturing Systems*. New Jersey: Prentice–Hall.

10. Byrne, M. D. and M. A. Bakir. 1999. Production Planning Using a Hybrid Simulation–Analytical Approach. *International Journal of Production Eco-*

nomics **59**: 305–311.

11. Caramanis, M. C., I. C. Paschalidis, and O. Anli. 1999. A Framework for the Decentralized Control of Manufacturing Enterprises. *Proceedings of the 1999 DARPA-JFACC Symposium on Advances in Enterprise Control.* San Diego, CA, November 15-16: 99–109.

12. Chen, H., J. M. Harrison, A. Mandelbaum, A. V. Ackere and L. M. Wein. 1988. Empirical Evaluation of a Queueing Network Model for Semiconductor Wafer Fabrication. *Operations Research.* **36** (2): 202–215.

13. Chen, H. B. and A. Mandelbaum. 1994. Hierarchical Modelling of Stochastic Networks Part I: Fluid Models. *Stochastic Modeling and Analysis of Manufacturing Systems* (eds. Yao, D. D.). New York: Springer.

14. Clark, G. M. 1981. Use of Polya Distributions in Approximate Solutions to Nonstationary M/M/s Queues. *Communications of the ACM* **24**: 206–217.

15. Cox, D. R. and V. Isham. 1980. Point Processes. Chapman&Hall.

16. Daley, D. J. and D. Vere-Jones. 2002. An Introduction to the Theory of Point Processes, Vol I: Elementary Theory and Methods. 2nd Ed. Springer.

17. Datta, P. P., M. Christopher and P. Allen. 2007. Agent–based Modeling of Complex Production/Distribution Systems to Improve Resilience. *International Journal of Logistics* **10** (3): 187–203

18. Deb, K. 2001. *Multi-Objective Optimization using Evolutionary Algorithms*. Chichester, UK: Wiley.

19. Draper, N. and H. Smith. 1981. *Applied Regression Analysis*, 2nd Edition. New York: John Wiley & Sons, Inc.

20. Eick, S. G., W. A. Massey and W. Whitt. 1993a. The physics of the Mt/G/infty Queue. *Operations Research* **41** (4): 731–742.

21. Eick, S. G., W. A. Massey and W. Whitt. 1993b. Mt/G/infty Queues with Sinusoidal Arrival Rates. *Management Science* **39** (2): 241–252.

22. Ettl, M, G. E. Feigin, G. Y. Lin and D. D. Yao. 2000. A Supply Network Model with Base–stock Control and Service Requirements. *Operations Research*. **48** (2): 216–232.

23. Golub, G. H. and V. L. F. Charles. 1996. *Matrix Computations*, 3rd edition. Johns Hopkins.

24. Graves, S. C. 1986. A Tactical Planning Model for a Job Shop. *Operations Research* **34**: 552–533.

25. Green, L. V. and P. J. Kolesar. 1991a. The Pointwise Stationary Approximation with for Queues with Nonstationary Arrivals. *Management Science* **37** (1): 84–97.

26. Green, L. V., P. J. Kolesar and A. Svornos. 1991b. Some Effects of Nonstationarity on Multiserver Markovian Queueing Systems. *Operations Research* **39** (3): 502–511.

27. Green, L. V., P. J. Kolesar and W. Whitt. 2007. Coping with Time-varying Demand when Setting Staffing Requirements for a Service System. *Production and Operations Management* **16** (1): 13–39.

28. Gross, D. and C. Harris. 1985. *Fundamentals of Queueing Theory.* New Jersey: John Wiley & Sons.

29. Gross, D. and D. Miller. 1984. The Randomization Technique as a Modeling Tool and Solution Procedure for Transient Markov Processes. *Operations Research* **32** (6): 926–944.

30. Grubbstrom, R. W. 1998, A Net Present Value Approach to Safety Stocks in

Planned Production. *International Journal of Production Economics* **56-57**: 213–229.

31. Grubbstrom, R. W., and A. Molinder, 1996. Safety Production Plans in MRP-Systems using Transform Methodology. *International Journal of Production Economics* **46**: 297–309

32. Grubbstrom, R. W., Z. P. Wang, 2002, A Stochastic Model of Multi-level/multi-stage Capacity-constrained Production-inventory Systems, *International Journal of Production Economics* **81-82**: 483-494.

33. Hackman, S. T. and R. C. Leachman (1989b). A General Framework for Modeling Production. *Management Science* **35**: 478-C495.

34. Henderson, S. G. and B. L. Nelson. 2006. *Handbooks in Operations Research and Management Science: Simulation.* Amsterdam, Netherlands: Elsevier Science.

35. Holt, C. C., F. Modigliani, J. F. Muth. 1956. Derivation of a Linear Rule for Production and Employment. *Management Science* **2** (2): 159–177.

36. Holt, C. C., F. Modigliani, J. F. Muth, H. A. Simon. 1960. Planning Production, Inventories and Work Force. Englewood Cliffs, New Jersey: Prentice Hall.

37. Hopp, W. and M. Spearman. 2000. *Factory Physics*, Second Edition. New York: Irwin/McGraw–Hill.

38. Hopp, W. J., M. L. Spearman, S. Chayet, K. L. Donohue, and E. S. Gel. 2002. Using an Optimized Queueing Network Model to Support Wafer Fab Design *IIE Transactions* **34**: 119–130.

39. Ingolfsson, A., E. Akhmetshina, S. Budge, Y. Li and X. Wu. 2007. A Survey and Experimental Comparison of Service Level Approximation Methods for Non–Stationary M/M/s Queueing Systems. *INFORMS Journal on Computing* **19** (2): 201–214.

40. Jennings, O. B., A. Mandelbaum, W. A. Massey and W. Whitt. 1996. Server Staffing to Meet Time–Varying Demand. *Management Science* **42** (10): 1383–1394.

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

42. Karmarkar, U. S. 1987. Lot Sizes, Lead Times and In-process Inventories. *Management Science* **33**: 409–418

43. Karmarkar, U. S. 1989 Capacity Loading and Release Planning with Work-in-process (WIP) and LeadTimes. *Journal of Manufacturing and Operations Management* **2**: 105–123.

44. Karmarkar, U. S. 1993. Manufacturing Lead Tmes, Order Release and Capacity Loading. In: Graves SC, Rinnooy Kan A, Zipkin P (eds) *Logistics of Production and Inventory* vol **4** of *Handbooks in Operations Research and Management Science*. 287–329. Amsterdam: North-Holland.

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

46. Hocking, R. R. 1976. The Analysis and Selection of Variables in Linear Regression. *Biometrics* **32**.

47. 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.

48. Irdem, D. F., N. B. Kacar and R. Uzsoy. 2008. An Experimental Study of an Iterative Simulation-optimization Algorithm for Production Planning. *Proceed-*

ings of the 2008 Winter Simulation Conference. Baltimore, MA.

49. Jackson, J. R. 1963. Jobshop-like Queueing Systems. *Management Science* **10** (1): 131–142.

50. Jackson, J. R. 2004. Jobshop-like Queueing Systems -Ten Most Influential Titles of Management Sciences First Fifty Years. *Management Science* **50** (12).

51. Kacar, N. B. and R. Uzsoy. 2010. Estimation Clearing Functions from Simulation Data. *Proceedings of the 2010 Winter Simulation Conference*.

52. Kelly, F. P., S. Zachary and I. Ziedins. 1996. *Stochastic Networks: Theory and Applications Royal Statistical Society Lecture Note Series*. New York: Oxford University Press.

53. Kleinrock, L. 1975. *Queueing Systems*. New York: John Wiley & Sons.

54. Koh, S. C. L. 2004. MRP-Controlled Batch–Manufacturing Environment under Uncertainty. *The Journal of the Operational Research Society* **55** (3): 219–232.

55. Kumar, S. and P. R. Kumar. 2001. Queueing Network Models in the Design and Analysis of Semiconductor Wafer Fabs. *IEEE Transactions on Robotics and Automation* **17** (5): 548–561.

56. Law, A. M. and W. D. Kelton, *Simulation Modeling and Analysis*, 3rd edition. New York: McGraw–Hill.

57. Leachman, R. C. 1994. Modeling Techniques for Automated Production Planning in the Semiconductor Industry. *Optimization in Industry* (eds. Ciriani, T. and R. C. Leachman), New York: John Wiley & Sons.

58. Leachman, R., R. Benson, C. Liu, and D. J. Raar. 1996. IMPReSS: an Automated Production Planning and Delivery Quotation Syatem at Harris Corporation C Semiconductor Sector. *Interfaces* **26** (1): 6–37.

59. Ljung, L. 1999. *System Identification: Theory for the User*, 2nd edition. Englcwood Cliffs, NJ: Prentice Hall.

60. Mandelbaum, A. and W. A. Massey. 1995. Strong Approximations for Time–Dependent Queues. *Mathematics of Operations Research* **20** (11): 33–64.

61. Massey, W. A. and W. Whitt. 1997. Peak Congestion in Multi-Server Service Systems with Slowly Varying Arrival Rates. *Queueing Systems* **25**: 157–172.

62. Meal, H. C. 1979, Safety Stocks in MRP system, *Technical report: no. 166. Work Performed under Office of Naval Research Contract N00014-75-C-0556, Project no. NR 345-027.* Cambridge: Massachusetts Institute of Technology,

Operations Research Center.

63. Morton, T. E. and M. R. Singh, 1988. Implicit Costs and Prices for Resources with Busy Periods. *Journal of Manufacturing and Operations Management* **1**: 305-332.

64. Meng, G. and S. Heragu. 2004. Batch Size Modeling in a Multi-item, Discrete Manufacturing System via an Open Queueing Network. *IIE transactions* **36**: 743–753.

65. Missbauer, H. 2002. Aggregate Order Release Planning for Time-Varying Demand. *International Journal of Production Research* **40**: 688–718.

66. Missbauer, H and R. Uzsoy, 2010. *Optimization Models of Production Planning Problems*. New York: Springer.

67. Narahari, Y., N. Hemachandra and M. S. Gaur. 1997. Transient Analysis of Multiclass Manufacturing Systems with Priority Scheduling. *Computers & Operations Research* **24** (5): 387–398.

68. Nelson, B. L. and M. R. Taaffe. 2004a. The $Ph_t/Ph_t/\infty$ Queueing System: Part I – the Single Node. *INFORMS Journal on Computing* **16** (3): 266–274.

69. Nelson, B. L. and M. R. Taaffe. 2004b. The $[Ph_t/Ph_t/\infty\,]^K$ Queueing System: Part II – the Multiclass Network. *INFORMS Journal on Computing* **16** (3): 275–283.

70. Orlicky, J. 1975. *Material Requirements Planning: the New Way of Life in Production and Inventory Management*. New York: McGraw-Hill.

71. Pahl, J. S. Vob and D. L. Woodruff. 2005. Production Planning with Load Dependent Lead Times. *4OR* **3**: 257–302.

72. Pinedo, M. L. 2007. *Planning and Scheduling in Manufacturing and Services*. New York: Springer

73. Ravindran, A., 2008, Aggregate Capacitated Production Planning in a Stochastic Demand Environment, Dissertation, Purdue University, West Laffeyte, IN.

74. Riano, G. 2002. Transient Behavior of Stochastic Networks: Application to Production Planning with Load-Dependent Lead Times. Dissertation, School of Industrial and Systems Engineering, Georgia Institute of Technology. Atlanta, GA.

75. Riano, G., R. Serfozo, S. Hackman, S. H. Ng, L. P. Chan and P. Lendermann.

2003. Benchmarking of a Stochastic Production Planning Model in a Simulation Testbed. *Proceedings of 2003 Winter Simulation Conference*. New Orleans, LA. 1183-1191.

76. Rothkopf, M. H. and S. S. Oren. 1979. A Closure Approximation for the Nonstationary M/M/s Queue. *Management Science* **25**: 522–534.

77. Ross, S. M. 1995. *Stochastic Process*, 2nd edition. New York: John Wiley & Sons .

78. Shanthikumar, J. G., S. Ding and M. T. Zhang. 2007. Queueing Theory for Semiconductor Manufacturing Systems: A Survey and Open Problems. *IEEE Transactions on Automation Science and Engineeing* **4** (4): 513–522.

79. Sheffi, Y. and J. B. Rice. 2005. A Supply Chain View of the Resilient Enterprise. *Sloan Management Review* **47** (1): 41–48.

80. Srinivasan, A., M. Carey, T. E. Morton. 1988. *Resource Pricing and Aggregate Scheduling in Manufacturing Systems. Graduate School of Industrial Administration*, Carnegie-Mellon University. Pittsburgh, PA.

81. Stadtler, H. and C. Kilger. 2007. *Supply Chain Management and Advanced Planning: Concepts, Models, Software, and Case Studies*, 4th edition. New

York: Springer

82. Taaffe, M. R. and K. L. Ong. 1987. Approximating Nonstationary Ph(t)/M(t)/S/C Queueing Systems. *Annals of Operations Research* **8**: 103–116.

83. Tang, C. S. 2006. Robust Strategies for Mitigating Supply Chain Disruptions. *International Journal of Logistics: Research and Applications* **9** (1): 33–45.

84. Tardif, V., and M. L., Spearman. 1997. Diagnostic Scheduling in Finite-capacity Production Environments. *Computers and Industrial Engineering* **32** (4): 867–878.

85. Thomas, L. J. and J. McClain. 1993. Overview of Production Planning. In S. Graves, A. H. G. Rinnooy Kan and P. Zipkin (Eds.) *Handbooks in OR and MS, Vol 4: Logistics of Production and Inventory*, Chapter 7. North-Holland.

86. Vollmann, T. E., W. L. Berry, Whybark D.C, F. R. Jacobs. 2005. *Manufacturing Planning and Control for Supply Chain Management.* New York: McGraw-Hill.

87. Voss, S. and D. L. Woodruff. 2003. *Introduction to Computational Optimization Models for Production Planning in a Supply Chain.* Berlin, New York: Springer.

88. Winter, E. 2006. Optimal Incentives for Sequential Production Processes. *RAND Journal of Economics*. **37** (2): 376-390.

89. Woodruff, D. L, and S. Voss. 2004. A Model for Multi-stage Production Planning with Load Dependent Lead Times. *Proceedings of the International Conference on System Sciences*. Hawaii: 1425-1434.

90. Yang, F., B. E. Ankenman and B. L. Nelson. 2007. Efficient Generation of Cycle Time–Throughput Curves through Simulation and Metamodeling. *Naval Research Logistics* **54**: 78–93.

91. Yang, Y., B. E. Ankenman and B. L. Nelson. 2008a. Cycle Time Percentile Curves for Manufacturing Systems. *INFORMS Journal on Computing* to appear.

92. Yang, F., J. Liu, B. L. Nelson, B. E. Ankenman and M. Tongarlak 2008b. Metamodeling for Cycle Time–Throughput–Product Mix Surfaces using Progressive Model Fitting, submitted to *Production Planning and Control* (under minor revision).

93. Yang, F. 2008. Neural Network Metamodeling for Cycle–Time Based Performance Profiles in Manufacturing, submitted to *Naval Research Logistics*.