Proceedings of the 2005 Winter Simulation Conference
M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.

ESTIMATION OF PERCENTILES OF CYCLE TIME IN MANUFACTURING SIMULATION

Feng Yang
Bruce E. Ankenman
Barry L. Nelson

Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208-3119, U.S.A.

ABSTRACT

Cycle time-throughput (CT-TH) percentile curves quantify the relationship between percentiles of cycle time and factory throughput, and they can play an important role in strategic planning for manufacturing systems. In this paper, a highly flexible distribution, the generalized gamma, is used to represent the underlying distribution of cycle time. To obtain CT-TH percentile curves, we fit metamodels for the first three CT-TH moment curves throughout the throughput range of interest, determine the parameters of the generalized gamma by matching moments, and obtain percentiles by inverting the distribution. To insure efficiency and control estimation error, simulation experiments are built up sequentially using a multistage procedure. Numerical results are presented to demonstrate the effectiveness of the approach.

1 INTRODUCTION

Computer simulation is an essential tool for the design and analysis of complex manufacturing systems. Nevertheless, it is difficult to use simulation for strategic planning because, by its nature, a simulation can only evaluate one particular scenario at a time. A cycle time-throughput (CT-TH) curve, on the other hand, shows the projected cycle time plotted against a range of throughput. Using this curve, a company can control cycle time by controlling the rate at which products are released into the system, or assess the impact on cycle time of increasing or decreasing the release rate. Since long-run average cycle time and long-run average Work in Process (WIP) are proportional, a CT-TH mean curve can also be used to control WIP.

Our focus in this paper is providing CT-TH curves that can be used to assess lead-time commitments with a high level of confidence. For this application, a CT-TH percentile curve is more relevant than a mean curve. A CT-TH percentile curve is simply a given percentile of the cycle time distribution as a function of the throughput desired. For example, if the CT-TH 95^{th} percentile curve is used to set the throughput level, then 95% of the time the actual lead time of any given product will meet the promised delivery time.

In Yang et al. (2004), a procedure was developed to efficiently generate CT-TH mean curves. In this paper, we will significantly extend this methodology to the generation of simulation-based CT-TH percentile curves. Since most manufacturing systems operate in a throughput range where the CT-TH curve is monotonically increasing, we restrict ourselves to this case. Once a CT-TH curve is constructed, the sensitivity of cycle time to throughput can be appraised by examining the curvature or steepness, and different operating policies can be quantitatively evaluated by comparing the curves generated for different scenarios of product mix, production targets and capital expansion.

The goal is to provide a methodology that generates CT-TH percentile curves given the following inputs

- the simulation model,
- a throughput range of interest, say $[x_L, x_U]$,
- a percentile range of interest, say $[\alpha_L, \alpha_U]$, and
- a measure of the required precision for the estimated curves.

Simulation is often used to provide percentile estimates, and substantial research effort has been devoted to the estimation of cycle time percentiles via simulation. However, efficiently generating cycle time percentile estimates remains a challenging topic for at least two reasons: Standard estimators based on order statistics may require excessive data storage unless all of the percentiles of interest are known in advance, and even then it is difficult to do sequential estimation until a fixed precision is reached (Chen and Kelton 1999). On the other hand, approximations based on only the first two moments of cycle time and assuming a normal distribution can be grossly inaccurate (McNeill et al. 2003). A technique based on the Cornish-Fisher expansion has been proposed by McNeill et al. (2003) to estimate percentiles of cycle times; it takes into account the first four moments of the cycle time distribution and allows accurate and precise percentile estimates to be generated for moderately non-normal distributions. However, this method can only give percentiles at fixed, prespecified throughputs where simulation experiments have been performed. The methodology proposed in this paper aims at providing a more comprehensive profile of the system by generating CT-TH percentile curves throughout a throughput range.

2 METHODOLOGY

In this section, we describe the methodology for generating the CT-TH percentile curves for a given simulation model.

2.1 Overview of the Approach

A highly flexible distribution type, the generalized gamma distribution (GGD), is assumed for the underlying random variable, cycle time. The first three moments of cycle time are utilized to generate a fit of the GGD distribution. The percentile estimates are then obtained by taking the inverse of the fitted GGD. More specifically, the strategy we propose for estimating $\mathcal{C}_\alpha(x)$, the $\alpha \in [\alpha_L, \alpha_U]$ percentile of cycle time at any throughput rate $x \in [x_L, x_U]$, is as follows:

1. Use an extended version of the methodology of Yang, et al. (2004) to estimate not only the CT-TH mean (1^{st} moment) curve, but also the CT-TH 2^{nd} and 3^{rd} moment curves over the throughput range of interest. This allows for the prediction of the first three moments of cycle time at any throughput x, say $\mu_1(x)$, $\mu_2(x)$, and $\mu_3(x)$. The form of the models for the relationship between throughput and the moments of cycle time are derived from heavy traffic approximations that hold for many queueing system at levels of throughput that are near the capacity of the system (see Whitt 1989).
2. Use method of moments to fit a GGD distribution $G(t; a(x), b(x), k(x))$ as an approximation for the cycle time distribution ($a(x)$,$b(x)$, and $k(x)$ are distribution parameters that depend on x). We write the resulting fitted GGD as $G(t; \widehat{a}(x), \widehat{b}(x), \widehat{k}(x))$.
3. Estimate the percentile $\mathcal{C}_\alpha(x)$ by taking the inverse of the c.d.f. of the cycle time: $\widehat{\mathcal{C}_\alpha}(x) = G^{-1}(\alpha; \widehat{a}(x), \widehat{b}(x), \widehat{k}(x))$.

2.2 Technical Details

In this subsection, we discuss the technical details of the approach.

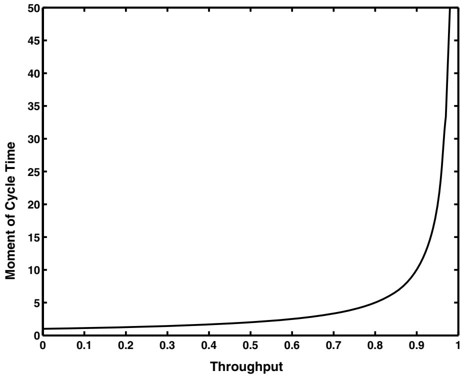

Figure 1: A Generic CT-TH Moment Curve

2.2.1 Estimation of CT-TH Moment Curves

As indicated in Section 2.1, providing the first three CT-TH moment curves over the throughput range of interest is the primary step in the estimation of $\mathcal{C}_\alpha(x)$.

In Yang, et al. (2004), a metamodeling-based methodology was developed for estimating CT-TH mean (1^{st} moment $\mu_1(x)$) curves via simulation. The same method can be directly applied to the simultaneous estimation of higher moment curves based on a single set of simulation experiments.

In manufacturing systems, CT-TH moment curves typically follow the shape in Figure 1. We suppose that the experiment is made up of a number of independent simulation runs performed at m distinct levels of throughput $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ with $x_i \in [x_L, x_U]$ for $i = 1, 2, \ldots, m$. From the j^{th} run performed at throughput x, an output response $\{Y_j^{(\nu)}(x), \nu = 1, 2, 3\}$, can be obtained for the purpose of estimating the ν^{th} moment curve:

$$Y_j^{(\nu)}(x) = \frac{1}{H(x)} \sum_{h=1}^{H(x)} (CT_{jh}(x))^\nu \quad j = 1, 2, \ldots, n(x) \quad (1)$$

where $n(x)$ is the number of replications placed at throughput x; $CT_{jh}(x)$ represents the individual cycle time of the h^{th} product simulated in the j^{th} replication at x; and $H(x)$ is the selected number of products simulated in steady state for simulations at x. For a given experiment consisting of a number of simulation replications carried out at m design points, the data sets

$$\mathbf{Y}_\nu = (Y_1^{(\nu)}(x_i), \ldots, Y_{n(x_i)}^{(\nu)}(x_i), i = 1, 2, \ldots, m)$$

can be extracted for $\nu = 1, 2, 3$.

To these data sets $\{\mathbf{Y}_\nu, \nu = 1, 2, 3\}$ we fit the three moment curves based on the regression metamodel (2) below, which is assumed to represent the CT-TH ν^{th} moment curve. For the sake of simplicity, since all of the metamodels have the same form, we omit the superscript ν representing the

ν^{th} moment in the model:

$$Y_j(x) = \mu(x, \mathbf{c}, p) + \varepsilon_j(x) \qquad j = 1, 2, \ldots, n(x) \tag{2}$$

where

$$\mu(x, \mathbf{c}, p) = \frac{\sum_{\ell=0}^{t} c_\ell x^\ell}{(1-x)^p}. \tag{3}$$

The exponent p, the polynomial order t, and the coefficient vector $\mathbf{c} = (c_0, c_1, \ldots, c_t)$ are unknown parameters in each model. Thus, a total of $3(t+3)$ parameter estimates are needed to fit all three moment models.

We write the resulting fitted curves for the first three moments as $\widehat{\mu}_\nu(x)$ $(\nu = 1, 2, 3)$ with throughput $x \in [x_L, x_U]$.

2.2.2 The Generalized Gamma Distribution

The distribution family chosen to fit the individual cycle times for manufacturing settings should be able to provide a good fit for a variety of cycle time distributions. It has been pointed out by Rose (1999) that for complicated systems, cycle times tend to be close to normally distributed. However, as the system is more and more heavily loaded, even for complicated systems, the cycle time distribution becomes more and more skewed (McNeill et al. 2003). Therefore, we decided to adopt the generalized gamma distribution because of its flexibility compared to other commonly used distributions. The GGD can cover a wide range of skewness as well as kurtosis.

The three-parameter GGD, first presented in Stacy (1962), has the following p.d.f.

$$\begin{aligned} g(t; a, b, k) &= \frac{|k|}{\Gamma(a)} \cdot \frac{t^{ak-1}}{b^{ak}} \cdot \exp[-(t/b)^k], \quad t > 0 \\ & \quad a > 0, b > 0, k \neq 0 \end{aligned} \tag{4}$$

where a and k are the shape parameters, and b the scale parameter. The GGD includes a variety of distributions as special cases, such as exponential ($a = k = 1$), gamma ($k = 1$), and Weibull ($a = 1$) distributions. The lognormal and normal distributions also arise as limiting cases.

Noncentral moments of GGD are given by:

$$m_\nu = \frac{b^\nu \Gamma(a + \nu/k)}{\Gamma(a)} \qquad \nu = 1, 2, 3, \ldots \tag{5}$$

where ν is the order of the moment. Choosing any three distinct values for ν will provide the equations required by the method of moments to obtain the three distribution parameters, a, b, and k.

2.2.3 Estimation of Percentiles

In this section, we discuss how to fit the generalized gamma distribution at a given throughput level $x \in [x_L, x_U]$ based on the first three moment estimators, how the percentiles are estimated once $G(t, \widehat{a}(x), \widehat{b}(x), \widehat{k}(x))$ is obtained, and how an approximate variance can be provided for the percentile estimators.

Point Estimation: As explained in Section 2.2.1, the first three moment curves can be fitted simultaneously based on a single set of simulation experiments performed at different levels of throughput. Therefore, for any $x \in [x_L, x_U]$, the first three moments $\{\widehat{\mu}_\nu(x), \nu = 1, 2, 3\}$ can be predicted. Substituting the moment estimates into (5) results in the following equations:

$$\begin{aligned} \widehat{\mu}_1(x) &= \frac{\widehat{a}(x)\Gamma(\widehat{k}(x) + 1/\widehat{b}(x))}{\Gamma(\widehat{k}(x))} \\ \widehat{\mu}_2(x) &= \frac{\widehat{a}(x)^2\Gamma(\widehat{k}(x) + 2/\widehat{b}(x))}{\Gamma(\widehat{k}(x))} \\ \widehat{\mu}_3(x) &= \frac{\widehat{a}(x)^3\Gamma(\widehat{k}(x) + 3/\widehat{b}(x))}{\Gamma(\widehat{k}(x))}. \end{aligned} \tag{6}$$

Numerically solving the three Equations (6) gives the three estimated distribution parameters $(\widehat{a}(x), \widehat{b}(x), \widehat{k}(x))$ for the fitted GGD distribution at throughput x.

With the estimated distribution of cycle time at throughput x, $G(t; \widehat{a}(x), \widehat{b}(x), \widehat{k}(x))$, the percentile $\mathcal{C}_\alpha(x)$ can be estimated for any $\alpha \in [\alpha_L, \alpha_U]$ by taking the inverse of the distribution.

Statistical Inference for the Percentile Estimator: Drawing inference about a parameter obtained indirectly is in general difficult. In this paper, the delta method is applied to make inferences concerning the estimated percentiles.

The percentile $\mathcal{C}_\alpha(x)$ is estimated based on the fitted GGD distribution, and is obviously a function of the distribution parameters, $a(x)$, $b(x)$ and $k(x)$. The delta method provides the following approximation for calculating the variance of percentile estimators, where we supress the dependence on x for convenience:

$$\begin{aligned} \mathrm{Var}[\widehat{\mathcal{C}}_\alpha(x)] &\doteq \left(\frac{\partial \mathcal{C}_\alpha(x)}{\partial a}\right)^2 \mathrm{Var}[\widehat{a}] + \left(\frac{\partial \mathcal{C}_\alpha(x)}{\partial b}\right)^2 \mathrm{Var}[\widehat{b}] \\ & \left(\frac{\partial \mathcal{C}_\alpha(x)}{\partial k}\right)^2 \mathrm{Var}[\widehat{k}] + 2\left(\frac{\partial \mathcal{C}_\alpha(x)}{\partial a}\right)\left(\frac{\partial \mathcal{C}_\alpha(x)}{\partial b}\right) \mathrm{Cov}[\widehat{a}, \widehat{b}] \\ & +2\left(\frac{\partial \mathcal{C}_\alpha(x)}{\partial b}\right)\left(\frac{\partial \mathcal{C}_\alpha(x)}{\partial k}\right) \mathrm{Cov}[\widehat{b}, \widehat{k}] \\ & +2\left(\frac{\partial \mathcal{C}_\alpha(x)}{\partial k}\right)\left(\frac{\partial \mathcal{C}_\alpha(x)}{\partial a}\right) \mathrm{Cov}[\widehat{k}, \widehat{a}]. \end{aligned} \tag{7}$$

In (7), the partial derivatives of the percentile $\mathcal{C}_\alpha(x)$ with respect to the GGD parameters can be approximately calculated numerically (for details, see Yang et al. 2005).

Since the GGD parameters are estimated by matching the first three moments of the GGD distribution to the moment estimates $\{\widehat{\mu}_\nu(x), \nu = 1, 2, 3\}$, the variances and covariances in (7) are functions of the variances and covariances of $\widehat{\mu}_1(x)$, $\widehat{\mu}_2(x)$ and $\widehat{\mu}_3(x)$. This is where the delta method is applied for a second time. Using matrix notation, we have the following relationship as derived in Ashkar et al. (1988):

$$\begin{pmatrix} \mathrm{Var}[\widehat{a}] \\ \mathrm{Var}[\widehat{b}] \\ \mathrm{Var}[\widehat{k}] \\ \mathrm{Cov}[\widehat{a},\widehat{b}] \\ \mathrm{Cov}[\widehat{a},\widehat{k}] \\ \mathrm{Cov}[\widehat{b},\widehat{k}] \end{pmatrix} \doteq \begin{pmatrix} C_{11} & C_{12} & \cdots & C_{16} \\ C_{21} & C_{22} & \cdots & C_{26} \\ C_{31} & C_{32} & \cdots & C_{36} \\ C_{41} & C_{42} & \cdots & C_{46} \\ C_{51} & C_{52} & \cdots & C_{56} \\ C_{61} & C_{62} & \cdots & C_{66} \end{pmatrix}^{-1} \times \begin{pmatrix} \mathrm{Var}[\widehat{\mu}_1(x)] \\ \mathrm{Var}[\widehat{\mu}_2(x)] \\ \mathrm{Var}[\widehat{\mu}_3(x)] \\ \mathrm{Cov}[\widehat{\mu}_1(x),\widehat{\mu}_2(x)] \\ \mathrm{Cov}[\widehat{\mu}_1(x),\widehat{\mu}_3(x)] \\ \mathrm{Cov}[\widehat{\mu}_2(x),\widehat{\mu}_3(x)] \end{pmatrix} \tag{8}$$

For calculation of the matrix **C**, please refer to Ashkar et al. (1988). As can be seen from the derivation above, estimating the variance of percentile estimators comes down to the estimation of $\{\mathrm{Var}[\widehat{\mu}_\nu(x)], \nu = 1,2,3\}$ and $\{\mathrm{Cov}[\widehat{\mu}_{\nu_1}(x), \widehat{\mu}_{\nu_2}(x)]; \nu_1, \nu_2 = 1,2,3, \nu_1 \neq \nu_2\}$ involved in (8). These estimators can be obtained from the standard linear approximation to the covariance matrix of the parameters when fitting the three moment curves with nonlinear regression (details are given in Bates and Watts 1988 and Yang et al. 2005).

2.3 Procedure for Estimating Percentiles of Cycle Time

We now describe a multistage procedure to collect simulation data for estimating percentiles of cycle time. A high-level description of the procedure is provided in Figure 2.

Simulation experiments are carried out sequentially until the prespecified stopping criterion is satisfied. The experimentation is initiated with a starting design which allocates some replications to the two end points of the throughput range $[x_L, x_U]$. As the procedure progresses, new design points are added and additional replications are added in batches. Each batch of replications is allocated to the design points to minimize PM, an experiment design criterion that is related to the variance of the percentile estimators. Since the design criterion depends on the unknown parameters of the moment curve, the current best estimates of the parameters are used in the allocation of each batch of replications. As more simulation data are collected, increasingly precise estimators are obtained until the precision of the estimators matches the stopping criterion.

In the remainder of this subsection we will discuss the issue of designing experiments and the stopping criterion used in our procedure.

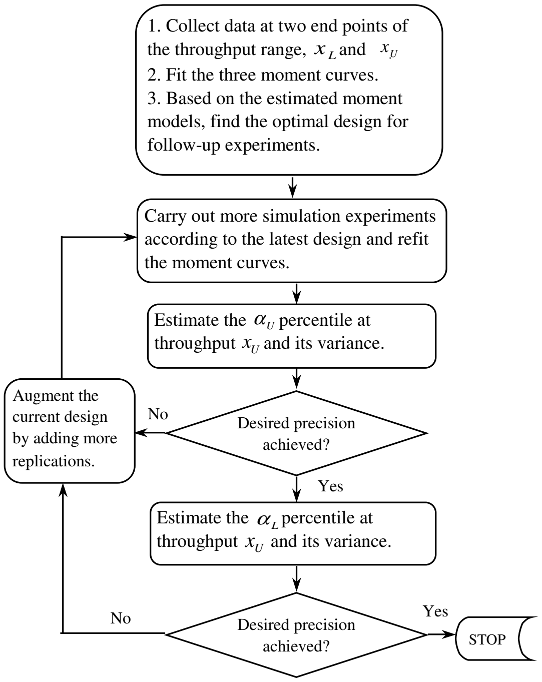
1. Collect data at two end points of the throughput range, x_L and x_U
2. Fit the three moment curves.
3. Based on the estimated moment models, find the optimal design for follow-up experiments.
Carry out more simulation experiments according to the latest design and refit the moment curves.
Estimate the α_U percentile at throughput x_U and its variance.
Desired precision achieved?
No
Augment the current design by adding more replications.
Yes
Estimate the α_L percentile at throughput x_U and its variance.
Desired precision achieved?
No
Yes
STOP

Figure 2: Flow Chart for the Multistage Procedure

2.3.1 Experiment Design

The experiment design consists of selecting the design points **x**, the throughput levels at which simulations will be executed, and the allocation $\boldsymbol{\pi}$, a vector of proportions that allocate a fraction of the total replications to each design point. Recall that our goal is to estimate the percentile $\mathcal{C}_\alpha(x)$ for $x \in [x_L, x_U]$ and $\alpha \in [\alpha_L, \alpha_U]$. Therefore, the experiment design will seek to minimize some measure of the variance of $\widehat{\mathcal{C}_\alpha}(x)$. Suppose that N is the number of replications available for allocation in the next step (see Yang et al. 2005 for the determination of N). A natural performance measure, which is inherited from Cheng and Kleijnen (1999), is the weighted average variance over the throughput range of interest normalized by N:

$$PM_0 = N \frac{\int_{x_L}^{x_U} w(x) \mathrm{Var}[\widehat{\mathcal{C}}_{\alpha_U}(x)] dx}{\int_{x_L}^{x_U} w(x) dx} \tag{9}$$

where $w(x)$ is the weight function which in the simplest case is $w(x) = 1$. We chose to base (9) on the variance of the largest percentile α_U because $\widehat{\mathcal{C}}_{\alpha_U}(x)$ is typically much more variable than other percentile estimators. Unfortunately, it is not practical to determine $[\mathbf{x}, \boldsymbol{\pi}]$ by minimizing PM_0, because $\mathrm{Var}[\widehat{\mathcal{C}}_\alpha(x)]$ can only be numerically estimated for given values of x and α, as can be seen from Section 2.2.3.

Hence, we use the finite difference approximation of (9):

$$PM = N \sum_{\kappa \in \boldsymbol{C}_x} \mathrm{Var}[\widehat{\mathcal{C}}_{\alpha_U}(\kappa)] \cdot \Delta\kappa \qquad (10)$$

where $\boldsymbol{C}_x$ is the chosen set of evenly spaced grid points in the range $[x_L, x_U]$, and $\Delta\kappa$ is the interval between two neighboring points. Since $\Delta\kappa$ is a constant, it can be dropped from (10), and we define our design criterion as:

$$PM = N \sum_{\kappa \in \boldsymbol{C}_x} \mathrm{Var}[\widehat{\mathcal{C}}_{\alpha_U}(\kappa)] \qquad (11)$$

Clearly, PM is a function of the design $[\mathbf{x}, \boldsymbol{\pi}]$, and by minimizing PM we determine how to allocate simulation replications in our experiments.

2.3.2 Stopping Criterion

The proposed procedure collects simulation data to allow for estimation of $\mathcal{C}_\alpha(x)$ for $x \in [x_L, x_U]$ and $\alpha \in [\alpha_L, \alpha_U]$ with both ranges of interest being specified by the user. Obviously, the upper end of throughput is where the variability of cycle time is most pronounced, and it is known that estimators of larger percentiles are more variable than their lower counterparts. Consequently, $\widehat{\mathcal{C}}_{x_U}(\alpha_U)$ is considered to possess the highest variability among all the estimable percentiles, which motivates us to use the relative error on $\widehat{\mathcal{C}}_{\alpha_U}(x_U)$ as the stopping criterion for our procedure. By controlling the precision of the most variable estimator $\widehat{\mathcal{C}}_{\alpha_U}(x_U)$, the other percentiles should be well estimated.

Specifically, we let the user specify a precision level, say $\gamma\%$, and the procedure terminates only when the condition

$$\frac{2\mathrm{SE}[\widehat{\mathcal{C}}_{\alpha_U}(x_U)]}{\widehat{\mathcal{C}}_{\alpha_U}(x_U)} \leq \gamma\%$$

is satisfied. We define $\mathrm{SE}[\cdot] = \sqrt{\mathrm{Var}[\cdot]}$. Moreover, a safe fall-back strategy is adopted. As illustrated in Figure 2, a check is also performed on the precision of $\widehat{\mathcal{C}}_{\alpha_L}(x_U)$, and simulation data will be collected until

$$\frac{2\mathrm{SE}[\widehat{\mathcal{C}}_{\alpha_L}(x_U)]}{\widehat{\mathcal{C}}_{\alpha_L}(x_U)} \leq \gamma\%$$

also.

Controlling the precision of all percentile estimators $\mathcal{C}_\alpha(x)$ to within a certain prespecified level is difficult. In the next section, we will show that, for some well known queueing systems, controlling the relative precision of only the two estimators $\widehat{\mathcal{C}}_{\alpha_U}(x_U)$ and $\widehat{\mathcal{C}}_{\alpha_L}(x_U)$ effectively controls the error in percentile estimators across the range of interest.

3 NUMERICAL EVALUATION

In this section, we evaluate the performance of the proposed procedure based on queueing models. Simple queueing systems, M/M/1, M/E$_2$/1, D/E$_2$/1 and D/M/1, have been considered in our experiments because they represent a range of cycle time distributions while still being analytically tractable. Not surprisingly, our procedure performs best on M/M/1, where the assumptions concerning the form of moment models and the distribution of cycle times are known to be true. Among these three systems, our procedure has the worst performance on the D/M/1 system. Due to space constraints, we only present the results for M/M/1 and D/M/1.

3.1 Results for Queueing Systems

For both M/M/1 and D/M/1, the true percentiles of cycle time at different throughputs can be analytically computed, and hence the quality of percentile estimation can be easily evaluated. For each model, the proposed procedure was applied 100 times, and from each of the 100 macro-replications the following outputs were recorded:

- The estimated CT-TH curves for the first three moments, $\{\widehat{\mu}_\nu(x), \nu = 1, 2, 3\}$, which allows for the estimation of percentile $\mathcal{C}_\alpha(x)$ for any $x \in [x_L, x_U]$ and $\alpha \in [\alpha_L, \alpha_U]$; and
- the inferred variance information from fitting the moment models, which allows for estimating both the $\{\mathrm{Var}[\widehat{\mu}_\nu(x)], \nu = 1, 2, 3\}$ and the $\{\mathrm{Cov}[\widehat{\mu}_{\nu_1}(x), \widehat{\mu}_{\nu_2}(x)]; \nu_1, \nu_2 = 1, 2, 3, \nu_1 \neq \nu_2\}$ for $x \in [x_L, x_U]$, and hence the variance of the percentile estimator $\widehat{\mathcal{C}}_\alpha(x)$ throughout the given throughput and percentile ranges of interest.

In our experiments, the throughput range of interest was chosen to be $[x_L, x_U] = [0.7, 0.95]$, and the percentile range $[\alpha_L, \alpha_U] = [85\%, 95\%]$, where we have normalized the throughput so that the maximum system capacity is 1. The precision level of the relative error used as the stopping criterion was set at $\gamma = 5\%$ (see Section 2.3.2). As already noted, our procedure is able to give percentile estimates $\mathcal{C}_\alpha(x)$ for any point in the two-dimensional region defined by the percentile $\alpha \in [\alpha_L, \alpha_U]$ and throughput $x \in [x_L, x_U]$. We call this region the feasible region. To evaluate the accuracy and precision of the percentile estimation, check points were selected inside this feasible region, as shown in Figure 3. At each of these points, the estimates were compared to the true percentiles of the queueing system.

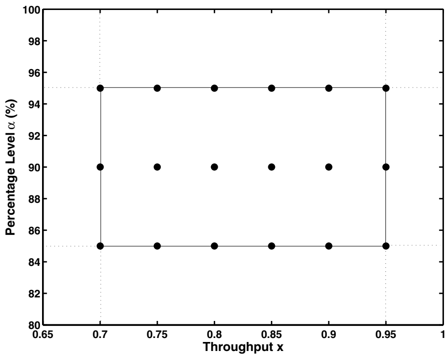

Figure 3: Check Points Selected in the Feasible Region

3.1.1 Point Estimators

All the point estimators for percentiles performed similarly well in terms of deviation from the true value for both M/M/1 and D/M/1. Two types of plots were made to display graphically the 100 realizations of each percentile estimator made at the check points: (i) relative error plots where the y-axis is defined as

$$\frac{\text{Percentile Estimate} - \text{True Percentile}}{\text{True Percentile}} \times 100\% \qquad (12)$$

and (ii) absolute error plots in which percentile estimates are plotted around their true values.

Figure 4 shows the percentile estimation results for M/M/1. Figure 4a, 4b, and 4c are relative error plots with the percentile α being 85%, 90%, and 95%, respectively. For these graphs, the x-axis represents throughput rate x, and every plot in the graph represents the relative deviation at corresponding check point (α, x) calculated by (12) from one of the 100 macro-replications. Notice that a very high proportion of the relative deviations of the percentile estimates at the selected check points are within 5% (the precision level $\gamma\%$ imposed prior to experimentation). Figures 4a$'$, 4b$'$, and 4c$'$ are the absolute error plots, in which the solid curve represents a piecewise linear version of the true percentile curve across the throughput range and the percentile estimates are plotted in absolute units. From these plots, it is evident that the variability of the percentile estimators at the highest throughput $x_U = 0.95$ is the most pronounced, and as explained in Section 2.3.2, it has been well controlled in our procedure.

Figure 5 shows an analogous plot for the D/M/1 system, and similar conclusions can be drawn, although the performance not as good as the M/M/1 especially when the throughput is at $x = 0.95$.

3.1.2 Standard Error

An estimator of the standard error $\text{SE}[\widehat{\mathcal{C}}_\alpha(x)] = \sqrt{\text{Var}[\widehat{\mathcal{C}}_\alpha(x)]}$ is provided for each percentile estimator $\widehat{\mathcal{C}}_\alpha(x)$ by the procedure described in Section 2.2.3. Our goal in this section is to evaluate the quality of the SE estimator. Tables 1 and 2 show the results for M/M/1 and D/M/1, respectively. The column labeled “Sample Stdev" is the sample standard deviation of the percentile point estimators calculated from the 100 realizations of the percentile estimator; therefore, it is an unbiased estimator of the true standard error. The “Average SE" column is the average of the 100 standard error estimators $\widehat{\text{SE}}[\widehat{\mathcal{C}}_\alpha(x)]$, each one of which is estimated from within a single macro-replication.

Table 2 shows that for M/M/1, the mean of the standard error estimate in the “Average SE" column is close to, but consistently less than, the unbiased external estimate of the standard deviation found in the “Sample Stdev" column. The underestimation trend is more apparent for the D/M/1. Nevertheless, the estimated standard error $\widehat{\text{SE}}[\widehat{\mathcal{C}}_\alpha(x)]$ provided by the procedure can still give the user a rough idea about how variable the percentile estimator is.

In the absence of any knowledge about the distribution of the percentile estimators, it would be natural to attempt to form a 95% confidence interval for the percentile by using:

$$\widehat{\mathcal{C}}_\alpha(x) \pm 1.96 \times \widehat{\text{SE}}[\widehat{\mathcal{C}}_\alpha(x)]. \qquad (13)$$

For M/M/1, (13) works well in terms of coverage and gives a conservative CI. However, for D/M/1, the coverage probability was lower than the nominal level. This can be explained by underestimation of the standard error, and non-normality of the percentile of cycle time estimator. In the case with D/M/1, it appears that non-normality is the dominant factor.

3.2 Summary of Results

Through experimentation based on queueing models, it has been shown that the proposed procedure is effective in providing accurate and precise percentile estimators. By controlling the relative standard error of the percentile estimators at the upper end of the throughput range, high precision has been achieved for estimators of percentiles throughout the feasible region.

For each percentile estimator, an estimate of the standard error is also provided which gives the user a sense of the size of the variability. However, in the scope of our work, there is not sufficient information to draw any conclusion regarding the distribution (or limiting distribution) of the percentile estimators. Thus, no reliable confidence interval can be created based on the standard error estimation.

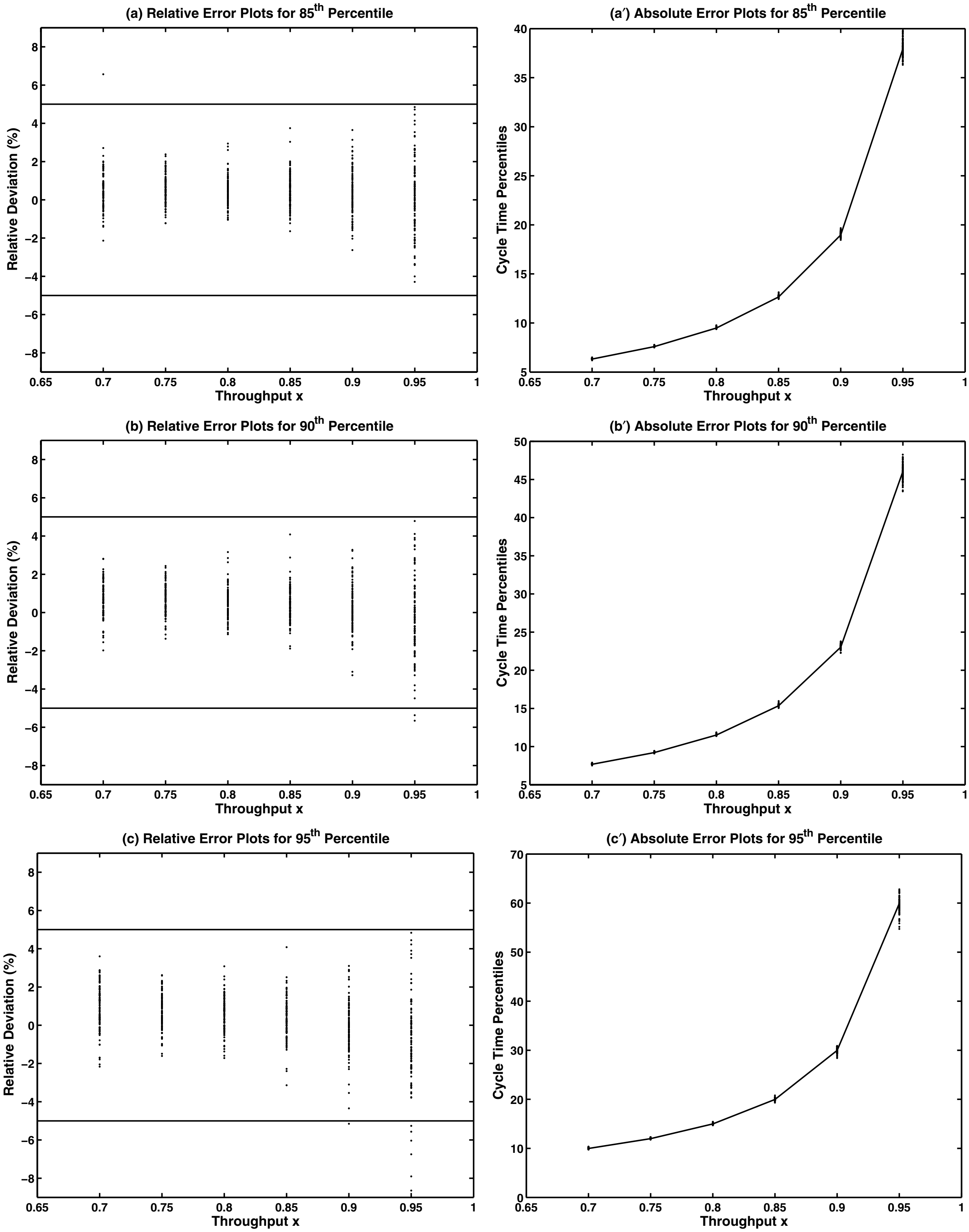

Figure 4: Plots of the Percentile Estimates for M/M/1 (100 Macro-replications)

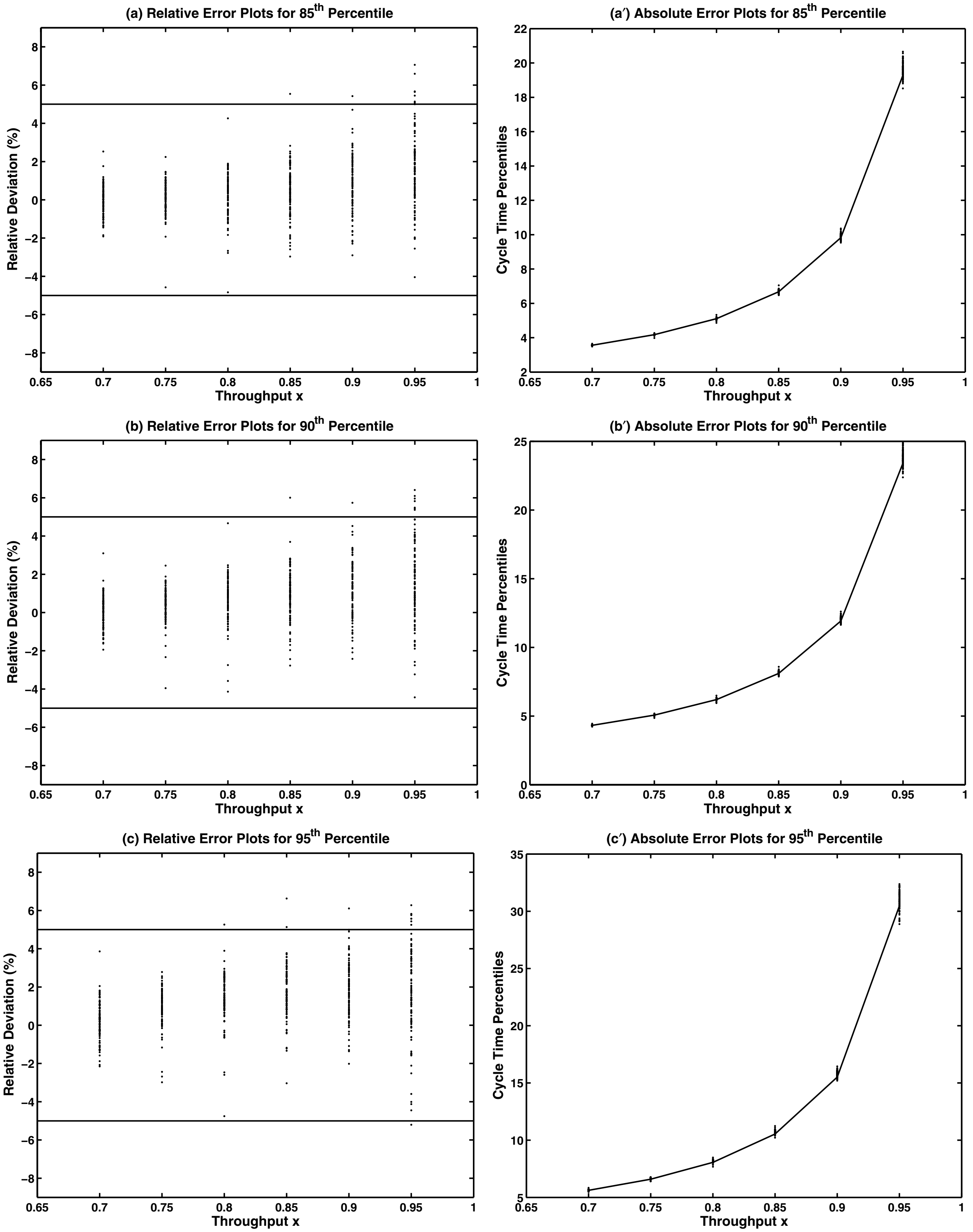

Figure 5: Plots of the Percentile Estimates for D/M/1 (100 Macro-replications)

Table 1: Estimated Standard Errors of Percentile Estimates for M/M/1

	85^{th} Percentile		90^{th} Percentile		95^{th} Percentile	
TH x	Sample Stdev	Average SE	Sample Stdev	Average SE	Sample Stdev	Average SE
0.70	0.055	0.053	0.072	0.067	0.117	0.105
0.75	0.056	0.053	.071	0.067	0.112	0.106
0.80	0.070	0.064	0.089	0.080	0.137	0.125
0.85	0.110	0.100	0.139	0.126	0.213	0.190
0.90	0.225	0.213	0.282	0.270	0.430	0.399
0.95	0.759	0.734	0.950	0.926	1.439	1.361

Table 2: Estimated Standard Errors of Percentile Estimates for D/M/1

	85^{th} Percentile		90^{th} Percentile		95^{th} Percentile	
TH x	Sample Stdev	Average SE	Sample Stdev	Average SE	Sample Stdev	Average SE
0.70	0.028	0.024	0.035	0.029	0.056	0.046
0.75	0.035	0.026	0.044	0.032	0.065	0.048
0.80	0.060	0.042	0.074	0.052	0.103	0.079
0.85	0.090	0.065	0.106	0.082	0.142	0.125
0.90	0.144	0.105	0.171	0.132	0.224	0.195
0.95	0.399	0.373	0.503	0.471	0.733	1.696

4 CONCLUSIONS

Estimating percentiles of cycle time is difficult due to the high variability of percentile estimators and the diversity of cycle time distributions. This paper proposes a new methodology for estimating multiple cycle time percentiles throughout the throughput range of interest based on a single set of simulation runs. It has been shown that for M/M/1 and D/M/1 system the multistage procedure developed provides good point estimators for percentiles of cycle time without requiring unreasonable computational effort.

Additional work will focus on applying the proposed procedure on full factory simulation models and evaluate the goodness of the percentile estimation.

ACKNOWLEDGMENTS

This research was supported by National Science Grant DMI-0140385 and Semiconductor Research Corporation Grant 2004-OJ-1225. Additional thanks go to Professors John Fowler and Gerald Mackulak from Arizona State University.

REFERENCES

Ashkar, F., B. Bobée, D. Leroux, and D. Morisette. 1988. The generalized method of moments as applied to the generalized gamma distribution. *Stochastic Hydrology and Hydraulics* 2: 161–174.

Bates, D. M., and D. G. Watts. 1988. *Nonlinear Regression Analysis and Its Applications*. New York: John Wiley & Sons Inc.

Chen, E. J., and W. D. Kelton. 1999. Simulation-based estimation of quantiles. In *Proceedings of the 1999 Winter Simulation Conference*, ed. P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, 428–434. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available via <http://www.informs-cs.org/wsc99papers/059.PDF> [accessed June 29, 2005].

Cheng, R. C. H., and J. P. C. Kleijnen. 1999. Improved design of queueing simulation experiments with highly heteroscedastic responses. *Operations Research* 47: 762–777.

McNeill, J. E., G. T. Mackulak, and J. W. Fowler. 2003. Indirect estimation of cycle time quantiles from discrete event simulation models using the Cornish-Fisher expansion. *Proceeding of the 2003 Winter Simulation Conference*, ed. S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 1377–1382. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available via <http://www.informs-cs.org/wsc03papers/173.pdf> [accessed June 29, 2005].

Rose, O. 1999. Estimation of the cycle time distribution of a wafer fab by a simple simulation model. In *Proceedings of the 1999 International Conference on Semiconductor Manufacturing Operational Modeling and Simulation*. 133–138

Stacy, E. W. 1962. A generalization of the gamma distribution. *The Journal of Mathematical Statistics* 33: 1187–1192.

Whitt, W. 1989. Planning queueing simulations. *Management Science* 35: 1341–1366.

Yang, F., B. E. Ankenman, and B. L. Nelson. 2004. Efficient generation of cycle time-throughput curves through simulation and metamodeling. Working paper, Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, Illinois.

Yang, F., B. E. Ankenman, and B. L. Nelson. 2005. Estimation of cycle-time percentile curves in manufacturing simulation. Working paper, Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, Illinois.

AUTHOR BIOGRAPHIES

FENG YANG is a Ph.D. candidate in the Department of Industrial Engineering and Management Sciences at Northwestern University. Her dissertation topic is the efficient generation of cycle time-throughput curves for manufacturing systems via simulation and metamodeling. She has also participated in the design and implement of simulation tools for General Motors. Her research interests include experimental design and quality control, simulation of manufacturing systems, and simulation input analysis. Her e-mail and web addresses are <ffyang@northwestern.edu> and <www.iems.northwestern.edu/~ffyang/>.

BRUCE E. ANKENMAN is an Associate Professor in the Department of Industrial Engineering and Management Sciences at the McCormick School of Engineering at Northwestern University. His current research interests include response surface methodology, design of experiments, robust design, experiments involving variance components and dispersion effects, and design for simulation experiments. He is a past chair of the Quality Statistics and Reliability Section of INFORMS, is an Associate Editor for *Naval Research Logistics* and is a Department Editor for *IIE Transactions: Quality and Reliability Engineering*. His e-mail address is <ankenman@northwestern.edu>, and his web page is <www.iems.northwestern.edu/~bea>.

BARRY L. NELSON is the Krebs Professor of Industrial Engineering and Management Sciences at Northwestern University, and is Director of the Master of Engineering Management Program there. His research centers on the design and analysis of computer simulation experiments on models of stochastic systems. He has published numerous papers and two books. Nelson has served the profession as the Simulation Area Editor of *Operations Research* and President of the INFORMS (then TIMS) College on Simulation. He has held many positions for the Winter Simulation Conference, including Program Chair in 1997 and currently Chair of its Board of Directors. His e-mail and web addresses are <nelsonb@northwestern.edu> and <www.iems.northwestern.edu/~nelsonb/>.