Proceedings of the 2012 Winter Simulation Conference
C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, eds.

A NEW PERSPECTIVE ON BATCHED QUANTILE ESTIMATION

Christos Alexopoulos
David Goldsman

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0205, USA

James R. Wilson

Edward P. Fitts Department of Industrial and Systems Engineering
North Carolina State University
Raleigh, NC 27695-7906, USA

ABSTRACT

We study asymptotically valid confidence intervals (CIs) for steady-state quantiles computed from nonoverlapping batches. Asymptotic validity of the CIs is established under conditions that are weaker and more easily verifiable than the usual mixing assumptions. The performance of the CIs is evaluated with a preliminary experimental study. These results form the basis for developing fully sequential procedures that yield CI estimators of steady-state quantiles with user-specified absolute or relative precision.

1 INTRODUCTION

Simulation is perhaps the most widely used tool in the fields of industrial engineering, operations research, and the management sciences. Steady-state simulations play a fundamental role in system design, and they are particularly appropriate for evaluating long-run system performance or risk. For instance, what is the steady-state expected return from a certain financial management strategy, or what is the long-term probability of a default? Although there are now many commercial and public-domain software packages supporting the development of valid and efficient simulation models for complex systems, rarely have these packages been equipped with *comprehensive* facilities for performing rigorous, state-of-the-art statistical analysis of the outputs arising from steady-state simulation experiments. In many large-scale simulation applications, most of the effort is devoted to the development and execution of computer-based models, while relatively little attention is devoted to careful follow-up analysis of the final results. For questions about risk in the context of steady-state simulation analysis, there is not even much literature on the supporting theory, not to mention the lack of implementation of that theory in practical problems.

This paper is a step towards the development of sequential procedures for computing valid point estimators and confidence intervals for steady-state quantiles. While a mean measures central tendency, quantiles are used to measure risk; furthermore, CIs for means measure estimation error, not future risk (Nelson 2008). In many applications of simulation to problems of risk analysis, a typical objective is to estimate quantiles such as the Value at Risk of a portfolio (Glasserman 2004) and the fair value of options.

In the development of effective steady-state simulation analysis procedures, the main obstacle is that generally the associated output processes do not even approximately satisfy the basic assumptions underlying conventional statistical methods—in particular, successive outputs are rarely independent and identically distributed (i.i.d.) normal random variables (e.g., consecutive waiting times in a heavily congested queueing simulation with the empty-and-idle initial condition). Consider the estimation of the steady-state mean

978-1-4673-4781-5/12/$31.00 ©2012 IEEE

$\mu \equiv \lim_{n\to\infty} \mathrm{E}(X_n)$ of a process $\{X_i : i \geq 1\}$ representing successive responses within a single simulation run. If the simulation is in steady-state operation, then the sample mean $\overline{X}_n \equiv n^{-1}\sum_{i=1}^{n} X_i$ based on the observations $\{X_1,\ldots,X_n\}$ is an unbiased estimator of μ. To estimate the precision of $\overline{X}_n$ as a point estimator of μ, we seek an estimator $\widehat{\mathrm{Var}}(\overline{X}_n)$ of $\mathrm{Var}(\overline{X}_n)$; and ultimately we build a CI for μ that is typically of the form $\overline{X}_n \pm q\big[\widehat{\mathrm{Var}}(\overline{X}_n)\big]^{1/2}$, where q is an appropriate critical value. The CI's half-length represents the precision ("margin for error") of its midpoint $\overline{X}_n$, while the coverage probability represents the likelihood of achieving that precision in repeated applications.

If the X_i are i.i.d., then the sample variance $S_n^2 \equiv \sum_{i=1}^{n}(X_i - \overline{X}_n)^2/(n-1)$ is an unbiased estimator of $\mathrm{Var}(X_i)$ so that $\mathrm{Var}(\overline{X}_n)$ can be estimated by S_n^2/n. Otherwise, $\overline{X}_n$ and S_n^2/n can be biased estimators of μ and $\mathrm{Var}(\overline{X}_n)$, respectively. In many practical applications, $\mathrm{E}(S_n^2/n) \ll \mathrm{Var}(\overline{X}_n)$ (Law 2007); and then CIs for μ based on $\overline{X}_n$ and S_n^2 have such significant undercoverage as to make those CIs grossly misleading.

The estimation of $\mathrm{Var}(\overline{X}_n)$ or, almost equivalently, the asymptotic variance parameter $\sigma^2 \equiv \lim_{n\to\infty} n\mathrm{Var}(\overline{X}_n)$, has been the goal of many techniques, including nonoverlapping batch means (NBM) (Fishman 2001), overlapping batch means (OBM) (Meketon and Schmeiser 1984), and standardized time series (STS) (Schruben 1983). Some of these techniques, especially STS, can be used to obtain variance estimators that possess low bias and variance, and hence low mean squared error (MSE) (Alexopoulos et al. 2007a, 2007b). The literature contains several effective sequential procedures based on NBM that deliver CIs for μ with user-specified absolute or relative accuracy (Fishman and Yarberry 1997; Lada, Steiger, and Wilson 2008; Steiger et al. 2005; Tafazzoli et al. 2011a, 2011b, 2011c).

Compared with estimation of the steady-state mean, the development and implementation of automated sequential procedures for estimating steady-state quantiles is much more difficult. Given $p \in (0,1)$ and the marginal cumulative distribution function (c.d.f.) of the target process, $F(x) \equiv \Pr\{X_1 \leq x\}$, $x \in \mathbb{R}$, we define the p quantile as $x_p \equiv F^{-1}(p) \equiv \inf\{x : F(x) \geq p\}$. When the data are identically distributed, the point estimation of x_p is straightforward: sort the observations in order $X_{(1)} \leq \cdots \leq X_{(n)}$ to yield the estimator $\widehat{x}_p = X_{(\lceil np \rceil)}$, where $\lceil \cdot \rceil$ denotes the ceiling function. If the X_i are also independent and $F(\cdot)$ is differentiable at x_p with derivative $F'(x_p) > 0$, then valid large-sample CIs for x_p can also be easily computed. The variate $\sqrt{n}(\widehat{x}_p - x_p)$ is asymptotically normal with mean zero and variance $p(1-p)/[F'(x_p)]^2$ (Hogg, McKean, and Craig 2005); thus an asymptotically valid CI for x_p has the form $\widehat{x}_p \pm q\big[\widehat{\mathrm{Var}}(\widehat{x}_p)\big]^{1/2}$, where $\widehat{\mathrm{Var}}(\widehat{x}_p)$ is an estimator of $\mathrm{Var}(\widehat{x}_p)$ and q is the associated critical value.

If the X_i are dependent and possibly contaminated by an initial transient, then the quantile x_p can be estimated using the data observed on a single run using any of the methods described in Bekki et al. (2010), Chen and Kelton (2006, 2008), Iglehart (1976), Jain and Chlamtac (1985), Jin, Fu, and Xiong (2003), Raatikainen (1987, 1990), and Seila (1982a, 1982b). The relatively sparse simulation literature on this problem reflects the following difficulties: (a) lack of an adequate theoretical basis for some of the existing methods; (b) lack of effective guidelines for using the methods in practice; (c) poor performance of the estimators in industrial-strength applications; and (d) excessive computational or storage requirements.

In Section 3, we focus on the method of nonoverlapping batch quantiles (NBQ), wherein we form batches and use within-batch sample quantiles as the basic observations. Wood and Schmeiser (1995) study quantile estimation based on overlapping batches. For extreme quantiles, one can apply the maximum transformation method (Heidelberger and Lewis 1984) to independently simulated groups of observations by averaging the within-group quantile estimators across groups to yield point and CI estimators of x_p.

The remainder of this article proceeds as follows. Section 2 lays the theoretical foundations of the NBQ methodology based on assumptions that are more applicable and easier to verify that the mixing conditions often imposed in the literature. The preliminary comparisons of the NBQ method with existing methods presented in Section 3 illustrate the potential of the sequential methods under study. Section 4 contains concluding remarks and outlines the next steps in our endeavor. The slides for the oral presentation of this article are available online via www.ise.ncsu.edu/jwilson/files/wsc12nbq.pdf [accessed July 15, 2012].

2 QUANTILE ESTIMATORS BASED ON THE METHOD OF NONOVERLAPPING BATCH QUANTILES (NBQ)

To lay a sufficiently broad foundation for building point and CI estimators of the steady-state p quantile x_p, we assume the stationary simulation output process $\{X_i : i = 0, 1, \ldots\}$ can be expressed as a (measurable, possibly nonlinear) function of a sequence of "shocks" $\{\varepsilon_i : i \in \mathbb{Z}\}$ that are i.i.d. random variables,

$$X_i = G(\ldots, \varepsilon_{i-2}, \varepsilon_{i-1}, \varepsilon_i) \quad \text{for } i = 0, 1, \ldots, \tag{1}$$

so the $\{\varepsilon_i\}$ may be regarded as the stream of random numbers driving the simulation, and the function $G(\cdot)$ represents the operations performed by the simulation model on its probabilistic inputs up to time i so as to generate the corresponding output response X_i. We assume that in a nonempty open interval $\mathscr{D}(x_p)$ containing the desired quantile x_p, the random variable X_i has a probability density function (p.d.f.) $f(x)$ with derivative $f'(x)$ such that

$$f(x_p) > 0 \quad \text{and} \quad \sup\bigl\{\, f(x) + |f'(x)| : x \in \mathscr{D}(x_p)\,\bigr\} < \infty. \tag{2}$$

We also assume that $\{X_i : i = 0, 1, \ldots\}$ satisfies the *geometric-moment contraction* (GMC) condition—i.e., there exist constants $\alpha > 0$, $C > 0$, and $r \in (0, 1)$ such that for the independent input processes $\{\varepsilon_j : j \in \mathbb{Z}\}$ and $\{\varepsilon^*_j : j \in \mathbb{Z}\}$ each consisting of i.i.d. variates, we have

$$\mathrm{E}\Bigl[\bigl|G(\ldots, \varepsilon_{-2}, \varepsilon_{-1}, \varepsilon_0, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i) - G(\ldots, \varepsilon^*_{-2}, \varepsilon^*_{-1}, \varepsilon^*_0, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i)\bigr|^{\alpha}\Bigr] \le Cr^i \quad \text{for } i \ge 0\,. \tag{3}$$

The GMC condition (3) requires that if two paired replications of the simulation model associated with the function $G(\cdot)$ are initialized independently but then use common random numbers after the simulation starting time, then the difference $X_i - X^*_i$ between the matching output responses generated by the two simulations at time i will converge to zero in the mean of order α as the time index $i \to \infty$. If the GMC condition (3) holds, then the difference $X_i - X^*_i$ also converges in probability to zero as $i \to \infty$ (Bickel and Doksum 2007).

As noted by Wu (2005), condition (3) is easier to check than the strong mixing condition (Bradley 2005). The setup (1)–(3) applies to the usual finite-order moving-average and autoregressive processes; and the latter class of processes forms the basis for the autoregressive method of steady-state simulation analysis (Law 2007). Moreover, conditions (1)–(3) are satisfied by a rich diversity of widely used linear and nonlinear processes, including conditional heteroscedastic (ARCH) processes (Engle 1982), random coefficient autoregressive (RCA) processes (Tsay 1987), and threshold autoregressive (TAR) processes (Tong 1990), as well as a broad class of Markov chains (Wu and Woodroofe 2000).

Let $\{X_1, \ldots, X_n\}$ denote a data set from which we wish to build point and CI estimators of x_p using the NBQ method with b nonoverlapping batches each of size m. With the definition

$$I_i(x) \equiv \begin{cases} 1, & \text{if } X_i \le x \\ 0, & \text{otherwise} \end{cases} \quad \text{for } x \in \mathbb{R} \text{ and } i = 1, 2, \ldots,$$

we see that $\{I_i(x_p) : i = 1, 2, \ldots\}$ is a stationary process with $\mathrm{E}[I_i(x_p)] = p$ and $\mathrm{Var}[I_i(x_p)] = p(1-p)$. The GMC condition (3) and Theorem 4 of Wu and Shao (2004) imply that the process $\{I_i(x_p)\}$ satisfies the central limit theorem (CLT)

$$\sqrt{n}\,\bigl[\bar{I}(x_p, n) - p\bigr]/\sigma_{I(x_p)} \underset{n\to\infty}{\Longrightarrow} N(0, 1)\,, \tag{4}$$

where:

$$\bar{I}(x_p, n) \equiv n^{-1} \sum_{i=1}^{n} I_i(x_p)$$

is the sample mean of the data set $\{I_i(x_p) : i = 1, \ldots, n\}$;

$$\sigma^2_{I(x_p)} = \lim_{n\to\infty} n\mathrm{Var}\big[\bar{I}(x_p, n)\big] = p(1-p) \sum_{\ell=-\infty}^{\infty} \rho_{I(x_p)}(\ell) < \infty \tag{5}$$

is the variance parameter for the process $\{I_i(x_p)\}$; $\rho_{I(x_p)}(\ell) \equiv \mathrm{Corr}\big[I_i(x_p), I_{i+\ell}(x_p)\big]$ denotes the lag-ℓ correlation for $\ell \in \mathbb{Z}$; and $\underset{n\to\infty}{\Longrightarrow}$ denotes weak convergence as $n \to \infty$.

Next we consider relevant statistics computed from b nonoverlapping batches of the original sequence of observations $\{X_i\}$, where the m is the size of each batch; and we let $m \to \infty$ while keeping b fixed. From the jth batch of observations $\{X_{(j-1)m+1}, \ldots, X_{jm}\}$ (where $j = 1, \ldots, b$), we define the associated batch means,

$$\bar{I}_j(x_p, m) \equiv m^{-1} \sum_{i=1}^{m} I_{(j-1)m+i}(x_p) \quad \text{for } j = 1, \ldots, b\,.$$

Moreover we sort the jth batch of observations $\{X_{(j-1)m+1}, \ldots, X_{jm}\}$ in ascending order to yield the order statistics

$$X_{j,(1)} \le X_{j,(2)} \le \cdots \le X_{j,(m)}\,;$$

and the associated quantile estimator based on the jth batch of size m is

$$\widehat{x}_p(j, m) = X_{j,(\lceil mp \rceil)}\,.$$

From Theorem 4 of Wu (2005), we see that with probability 1 (w.p.1), $\lim_{m\to\infty} \widehat{x}_p(j,m) = x_p$ and $\lim_{m\to\infty} \bar{I}_j(x_p, m) = p$ for $j = 1, \ldots, b$ so that we also have the usual batch means–type results:

$$\lim_{m\to\infty} b^{-1} \sum_{j=1}^{b} \widehat{x}_p(j,m) = x_p \quad \text{and} \quad \lim_{m\to\infty} b^{-1} \sum_{j=1}^{b} \bar{I}_j(x_p, m) = p \quad \text{w.p.1.}$$

Since the far right-hand side of (5) is absolutely convergent, we see that $\lim_{|\ell|\to\infty} \rho_{I(x_p)}(\ell) = 0$; and if we take

$$\rho_{\bar{I}(x_p,m)}(\ell) \equiv \mathrm{Corr}\big[\bar{I}_j(x_p,m), \bar{I}_{j+\ell}(x_p,m)\big] \quad \text{for all } \ell \in \mathbb{Z}\,,$$

then we see that $\lim_{m\to\infty} \rho_{\bar{I}(x_p,m)}(\ell) = \lim_{m\to\infty} m^{-1} \sum_{j=-m+1}^{m-1} (1 - |j|/m) \rho_{I(x_p)}(\ell m + j) = 0$ for $\ell \in \mathbb{Z}$. Exploiting (4) and the continuous mapping theorem (Theorem 2.7 of Billingsley 1999) along the lines of the proof of Theorem 1 in Steiger and Wilson (2001), we see that the $b \times 1$ vector of batch means $\big[\bar{I}_1(x_p,m), \ldots, \bar{I}_b(x_p,m)\big]^{\mathrm{T}}$ satisfies the multivariate CLT

$$\sqrt{m}\,\big[\bar{I}_1(x_p,m) - p, \ldots, \bar{I}_b(x_p,m) - p\big]^{\mathrm{T}} \underset{m\to\infty}{\Longrightarrow} N_b\big[\mathbf{0}_b, \sigma^2_{I(x_p)} \mathbf{I}_b\big]\,,$$

where $\mathbf{0}_b$ is the $b \times 1$ vector of zeros and $\mathbf{I}_b$ is the $b \times b$ identity matrix.

From Theorem 4 of Wu (2005), we see that

$$\begin{aligned}\sqrt{m}\,\big[\widehat{x}_p(1,m) - x_p, \ldots, \widehat{x}_p(b,m) - x_p\big]^{\mathrm{T}} = -\big[\sqrt{m}/f(x_p)\big]\,&\big[\bar{I}_1(x_p,m) - p, \ldots, \bar{I}_b(x_p,m) - p\big]^{\mathrm{T}} \\ &+ \big[R_1(m), \ldots, R_b(m)\big]^{\mathrm{T}},\end{aligned}$$

where

$$R_j(m) = \big[\sqrt{m}/f(x_p)\big] \cdot O\big[m^{-3/4}(\log m)^{3/2}\big] = O\big[m^{-1/4}(\log m)^{3/2}\big] \underset{m\to\infty}{\longrightarrow} 0 \quad \text{for } j = 1, \ldots, b \quad \text{w.p.1.}$$

Combining the multivariate CLT for $\left[\bar{I}_1(x_p,m),\ldots,\bar{I}_b(x_p,m)\right]^{\mathrm{T}}$ with the last two displays and applying Slutsky's Theorem (Bickel and Doksum 2007), we obtain a multivariate CLT for batch quantiles under the GMC condition (3):

$$\sqrt{m}\left[\widehat{x}_p(1,m)-x_p,\ldots,\widehat{x}_p(b,m)-x_p\right]^{\mathrm{T}} \underset{m\to\infty}{\Longrightarrow} N_b\Big\{\mathbf{0}_b,\left[\sigma^2_{I(x_p)}\big/f^2(x_p)\right]\mathbf{I}_b\Big\}. \tag{6}$$

Muñoz (2010) obtained a comparable result for Markov chains that obey a certain functional CLT.

It follows from (6) that in terms of the sample mean and variance of the nonoverlapping batch quantiles,

$$\widetilde{x}_p(b,m)\equiv b^{-1}\sum_{j=1}^{b}\widehat{x}_p(j,m) \quad \text{and} \quad S^2_{\widehat{x}_p}(b,m)\equiv(b-1)^{-1}\sum_{j=1}^{b}\left[\widehat{x}_p(j,m)-\widetilde{x}_p(b,m)\right]^2,$$

as $m\to\infty$ an asymptotically valid $100(1-\alpha)\%$ NBQ CI for x_p has the form

$$\widetilde{x}_p(b,m)\pm t_{\alpha/2,b-1}S_{\widehat{x}_p}(b,m)\big/\sqrt{b}, \tag{7}$$

where $t_{\beta,\nu}$ is the $1-\beta$ quantile of Student's t distribution with ν degrees of freedom.

Paralleling the situation with the classical $100(1-\alpha)\%$ NBM CI for μ, the NBQ CI for x_p may require adjustments to handle the following anomalies that are due to the finite batch size m used in practice: (a) slowly declining bias in the $\{\widehat{x}_p(j,m)\}$ of the form $O\big[m^{-3/4}(\log m)^{3/2}\big]$; (b) residual nonnormality (specifically, nonzero skewness) of the $\{\widehat{x}_p(j,m)\}$; and (c) residual correlation between the $\{\widehat{x}_p(j,m)\}$. The skewness and correlation adjustments developed for recent NBM methods (Lada et al. 2008; Tafazzoli et al. 2011a, 2011b, 2011c) may be adapted to resolve issues (b) and (c), but (a) may require the formulation of an appropriate resampling method for dependent data (Lahiri 2003; Shao and Tu 1995)—e.g., a computationally efficient version of the jackknife-after-bootstrap method.

3 COMPARISONS WITH EXISTING METHODS

The fixed-sample-size methods of Iglehart (1976), Moore (1980), and Seila (1982a, 1982b) assume that the underlying process $\{X_i\}$ is regenerative, and those methods are based on independent sample quantiles obtained within regenerative cycles. In particular, Seila's method uses batches of regenerative cycles and jackknifing within each batch to reduce the bias of the average quantile estimator. The indirect method of Bekki et al. (2010) estimates the Cornish–Fisher expansion (Cornish and Fisher 1937) for an individual response X_i based on a standard normal random variable using the first four sample moments of the observations $\{X_i : i=1,\ldots,n\}$. Hence this method can estimate several quantiles simultaneously without data sorting. However, sample moments computed from highly correlated data can exhibit slow convergence to the true moments of the response; this phenomenon is evident from sample sizes of the same order of magnitude as in Table 1 below.

The first sequential method for estimating the steady-state quantile x_p was proposed by Raatikainen (1990). Within each nonoverlapping batch of observations, quantile estimates are computed by the extended P^2 algorithm (Jain and Chlamtac 1985; Raatikainen 1987), which approximates the marginal c.d.f. $F(\cdot)$ using a piecewise quadratic curve and then inverts the approximated c.d.f. to obtain a point estimator of x_p. The CI estimator for x_p exploits a univariate analogue of (6), spectral estimation of the variance parameter $\sigma^2_{I(x_p)}$ based on the method of Heidelberger and Welch (1981), and estimation of $f(x_p)$ using the approximation to $F(\cdot)$. This method has several drawbacks. (a) While the P^2 method avoids sorting and has low storage requirements, no conditions on the $\{X_i\}$ are established that are sufficient to ensure the final point estimator of x_p is consistent. (b) The CI for x_p requires estimating $f(x_p)$, which is problematic because the latter operation is based on a piecewise quadratic approximation to $F(\cdot)$ in a neighborhood of x_p. (c) The P^2 algorithm ignores the task of batch size selection, an archetypal problem in this area of study. (d) Our numerical experiments indicate that recent efficient sorting techniques and inexpensive storage outweigh the advantages of the P^2 algorithm that existed in the early 1990s.

The sequential algorithms of Chen and Kelton (2006) are based on a small number of replicate runs (typically set to 3). Loosely speaking, each iteration of the first run of their *zoom-in* (ZI) algorithm obtains lower and upper bounds on each quantile estimate and discards data outside the range of the computed bounds. Thus the run progressively zooms-in towards the unknown quantile. The first run terminates based on several rules, while the remaining runs use the bounds computed in the first run (though this results in correlated runs). The *quasi-independent* (QI) algorithm attempts to create approximately independent data by progressively spacing observations used to compute a quantile estimate within a replication. Although the ZI algorithm outperformed the QI algorithm in systems with a high degree of autocorrelation, the ZI algorithm's reliance on several user-defined parameters makes it difficult to convert to a fully automated procedure requiring minimal user intervention.

The following example illustrates the potential of the NBQ method. An ideal sequential procedure should progressively increase both the batch size and the number of batches in order to achieve the nominal CI coverage and precision.

Example 1 Consider a stationary M/M/1 queueing system with interarrival rate $\lambda = 0.8$ and service rate $\omega = 1$, and let X_i be the time-in-system of entity i. It is well known that the distribution of X_i is exponential with rate $\omega - \lambda = 0.2$; hence the response X_i has mean 5 and p quantile $x_p = -5\ln(1-p)$, $0 < p < 1$.

Table 1 contains experimental results based on 1000 independent replications of the QI algorithm in Chen and Kelton (2006) with $\varepsilon = 0.005$ such that $\Pr\{\widehat{x}_{p-\varepsilon} \le x_p \le \widehat{x}_{p+\varepsilon}\} \ge 0.95$, where the $p \pm \varepsilon$ quantile estimates are based on quasi-independent data. Each replication involves 3 independent runs; and the number of quasi-independent observations (38,416) required by each run was obtained from the formula $\lceil 1.96^2 p(1-p)/\varepsilon^2 \rceil$ for $p = 0.5$. The first row contains point estimates, the second row gives the average CI half-lengths, and the third row displays the estimated CI coverages. All algorithms were coded in Matlab and were executed on a Condor Unix cluster.

Table 1: Experimental results for the QI algorithm in Chen and Kelton (2006) for various quantiles of the time-in-system for an M/M/1 system with traffic intensity 0.8. The estimates are based on 1000 independent replications. The sample average number of observations required per replication was 50,879,967.

	p						
	0.1	0.3	0.5	0.7	0.9	0.95	0.99
Estimate	0.527	1.783	3.466	6.019	11.512	14.980	23.029
Half-length	0.015	0.023	0.045	0.069	0.137	0.199	0.456
Coverage	0.955	0.943	0.942	0.963	0.963	0.946	0.941

To establish a basis for applying the NBQ method to sojourn times in the steady-state M/M/1 queue, we attempted to verify the GMC condition (3) analytically. Let $X_i \equiv G(\ldots, \varepsilon_{-2}, \varepsilon_{-1}, \varepsilon_0, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i)$ and $X_i^* \equiv G(\ldots, \varepsilon_{-2}^*, \varepsilon_{-1}^*, \varepsilon_0^*, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i)$ for $i = 1, 2, \ldots$. Although we have been able to establish that there is a random cutoff time T such that $|X_i - X_i^*| = 0$ for $i > T$ and $\Pr\{0 < T < \infty\} = 1$ and although this appears to be a *stronger* result than the GMC condition (3), unfortunately we have been unable to verify the latter condition rigorously for the process $\{X_i\}$. Instead we attempted to provide some convincing empirical evidence that the GMC condition is satisfied for the simulation-generated process at hand. Visual evidence supporting the GMC condition is given in Figure 1.

We do not claim that Figure 1 constitutes definitive evidence of the validity of the GMC condition; but we believe that it provides good evidence of the phenomenon mentioned in the previous paragraph—namely, that beyond a certain time random time T that in this case is concentrated in the vicinity of the customer index $i \approx 375$, the difference $\mathrm{E}\big[\,|X_i - X_i^*|^\alpha\,\big]$ drops suddenly to zero; and prior to time T, the decline in the log-transformed response $\ln\big\{\mathrm{E}\big[\,|X_i - X_i^*|^\alpha\,\big]\big\}$ appears to be a nearly linear function of the customer index i. For the linear regression performed on the latter time series based on 4,000 independent replications

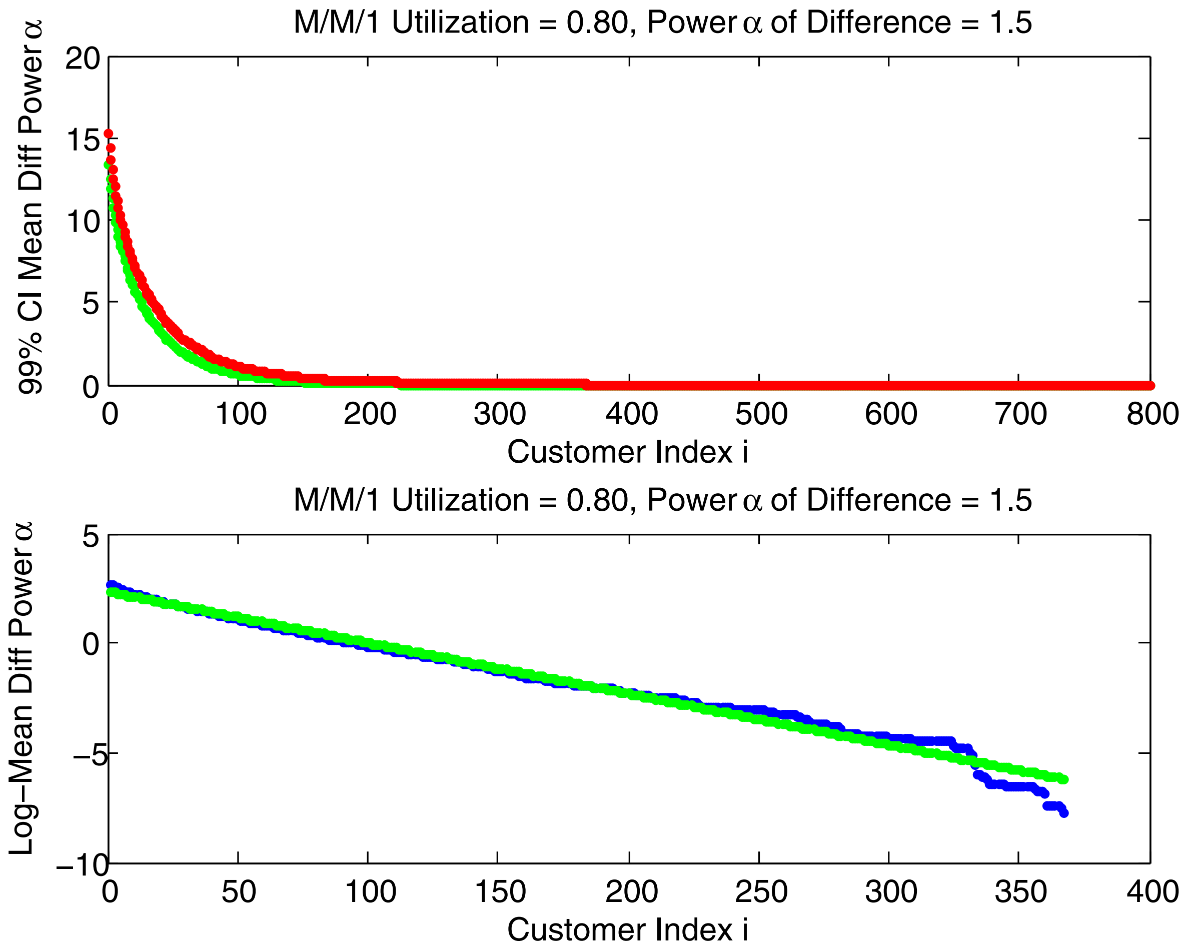

Figure 1: Empirical verification of GMC condition (3) with $\alpha = 1.5$ for sojourn times in an M/M/1 queue with server utilization 0.8. Upper panel is a plot of 99% CIs for $\mathrm{E}\big[\,\big|X_i - X_i^*\big|^{\alpha}\,\big]$ vs. customer index i, where $i = 0, 1, \ldots, 800$. Lower panel is a plot of $\ln\big\{\,\mathrm{E}\big[\,\big|X_i - X_i^*\big|^{\alpha}\,\big]\,\big\}$ vs. i for $i = 0, 1, \ldots, 375$.

of the stationary sojourn-time processes $\{X_i\}$ and $\{X_i^*\}$, we obtained a sample squared coefficient of correlation $R^2 = 0.979$ with the corresponding least squares estimates $\ln(C) \approx 2.3340$ (so that $\widehat{C} = 10.32$) and $\ln(r) \approx -0.0233$ (so that $\widehat{r} = 0.977$). On the basis of this graphical and statistical evidence, we believe that it is appropriate to apply the NBQ method to sojourn times in the M/M/1 queue. It is also noteworthy that in a complex, large-scale simulation model, performing a meaningful empirical verification of the GMC condition is substantially more straightforward than any attempt to verify conditions such as phi-mixing, strong mixing, or the assumption of a functional central limit theorem for the simulation-generated process at hand.

Table 2 displays experimental results based on 10,000 independent replications for the NBQ method with a fixed number of batches $b = 32$ and progressively increasing batch sizes m. For each value of m, the first row displays the averages of the point estimates for x_p and the CI half-lengths based on (7), while the second row displays the point estimates $mS^2_{\widehat{x}_p}(b,m)$ of the variance parameter $\sigma^2_{x_p} \equiv \sigma^2_{I(x_p)}\big/f^2(x_p)$ in (6) and the estimated coverages of the CIs (in parentheses). When we use batch sizes that are sufficiently large to yield valid CIs for the steady-state mean μ (namely, $m = 4{,}096$ or $8{,}192$ as discussed in Alexopoulos and Seila 1998), the quantile-estimation problems outlined at the end of §2 are clearly revealed by the relatively low coverages of the CI estimators of the selected steady-state quantiles and the relatively large biases of the associated variance estimators. The boxed entries indicate that for a batch size of $m = 16{,}384$, the variance estimators appear to converge to their (unknown) limits, and the 95% CIs seem to attain the nominal coverage.

An effective sequential procedure based on the NBQ method should be able to detect that with the batch size $m = 16{,}384$, the batched quantiles for each p under consideration have become approximately i.i.d. Consider the extreme case of $p = 0.99$. Since the sample variance of the batched quantiles is roughly $211{,}338/16{,}384 = 12.9$, an estimate of the number of batches of size 16,384 required to yield a 95% CI that is as tight as the CI produced by the QI algorithm is $\lceil 1.96^2 \times 12.9/0.456^2 \rceil = 239$; this corresponds to an approximate sample size of $16{,}384 \times 239 = 3{,}195{,}776$, which is almost 13 times smaller than the average sample size required by the QI algorithm. Indeed, an experiment with 256 batches of size 16,384 yielded 95% NBQ CIs with estimated coverage of 0.945 and average half-length 0.4436. The NBQ method becomes even more effective for smaller quantiles. For instance, the computation of an estimate for the median ($p = 0.5$) with the same absolute precision as the QI-based estimate in Table 1 would require about 76 batches of size 16,834; this corresponds to an approximate sample size of 1,245,184.

Table 2: Experimental evaluation of the NBQ method for various quantiles of the time-in-system for a stationary M/M/1 system with traffic intensity 0.8. The estimates are based on 10,000 independent replications with $b = 32$ batches and increasing batch sizes. For each batch size m, the CI on the first line has the form $\widehat{x}_p \pm H$, while the second line has the form $\widehat{\mathrm{Var}}[\widehat{x}_p]$ (estimated CI coverage).

				p			
m	0.1	0.3	0.5	0.7	0.9	0.95	0.99
512	0.585 ± 0.100	1.960 ± 0.302	3.780 ± 0.578	6.462 ± 0.989	11.474 ± 1.652	13.861 ± 1.881	17.272 ± 2.083
	39.4 (0.813)	359 (0.844)	1317 (0.873)	3852 (0.915)	10752 (0.917)	13937 (0.691)	17097 (0.010)
1024	0.552 ± 0.050	1.862 ± 0.167	3.614 ± 0.344	6.266 ± 0.662	11.714 ± 1.404	14.637 ± 1.745	19.187 ± 2.101
	19.9 (0.875)	220 (0.890)	930 (0.906)	3448 (0.924)	15525 (0.941)	23986 (0.877)	34764 (0.106)
2048	0.539 ± 0.032	1.821 ± 0.107	3.537 ± 0.219	6.143 ± 0.422	11.716 ± 1.049	15.032 ± 1.480	20.836 ± 2.058
	16.5 (0.911)	181 (0.917)	753 (0.925)	2802 (0.935)	17339 (0.949)	34529 (0.940)	66709 (0.424)
4096	0.533 ± 0.022	1.802 ± 0.073	3.501 ± 0.148	6.080 ± 0.283	11.627 ± 0.712	15.095 ± 1.114	22.038 ± 1.917
	15.3 (0.935)	166 (0.933)	690 (0.939)	2531 (0.944)	15991 (0.953)	39115 (0.951)	115816 (0.752)
8192	0.530 ± 0.015	1.793 ± 0.051	3.484 ± 0.103	6.051 ± 0.197	11.572 ± 0.484	15.052 ± 0.766	22.741 ± 1.675
	14.9 (0.941)	161 (0.943)	669 (0.944)	2444 (0.946)	14793 (0.947)	36999 (0.951)	176727 (0.897)
16384	0.528 ± 0.011	1.787 ± 0.035	3.474 ± 0.072	6.034 ± 0.137	11.539 ± 0.335	15.013 ± 0.522	23.001 ± 1.295
	14.5 (0.943)	157 (0.946)	651 (0.945)	2365 (0.946)	14116 (0.948)	34375 (0.948)	211338 (0.930)
32768	0.527 ± 0.008	1.785 ± 0.025	3.470 ± 0.051	6.027 ± 0.096	11.526 ± 0.234	14.995 ± 0.363	23.046 ± 0.923
	14.3 (0.943)	156 (0.946)	643 (0.943)	2335 (0.945)	13776 (0.947)	33166 (0.941)	214527 (0.950)
65536	0.527 ± 0.005	1.784 ± 0.018	3.468 ± 0.036	6.027 ± 0.068	11.519 ± 0.165	14.987 ± 0.255	23.038 ± 0.634
	14.3 (0.951)	155 (0.949)	640 (0.951)	2322 (0.952)	13681 (0.950)	32834 (0.951)	202817 (0.951)
131072	0.527 ± 0.004	1.784 ± 0.012	3.467 ± 0.025	6.022 ± 0.048	11.517 ± 0.116	14.984 ± 0.180	23.033 ± 0.440
	14.3 (0.951)	155 (0.952)	640 (0.951)	2322 (0.954)	13653 (0.950)	32639 (0.951)	195643 (0.948)

4 CONCLUSIONS

This paper obtained a central limit theorem for steady-state quantiles based on widely applicable conditions that are easier to verify than the typical, often-imposed mixing conditions. The preliminary experimental results in Section 3 illustrated the potential savings of a well-conceived sequential method over methods in the literature.

The development of effective sequential procedures involves various additional problems we plan to address. First, we plan to resolve the simulation start-up problem due to the simulation's initial condition. In practice, it is usually impossible to start a simulation in steady state; instead users often do the following: (a) start the simulation in some convenient initial condition that may not be typical of steady-state operation; and (b) select the warm-up period (whose statistics are discarded) so that beyond the warm-up point, the selected parameters (e.g., quantiles) of the observations are sufficiently close to the respective steady-state values.

In the context of estimating the steady-state p quantile x_p via batching, one can employ an approach based on spaced batches of observations. From each spaced batch, a separate batch quantile estimator is computed. By (6), the batch quantiles are asymptotically unbiased, normal, and independent as the batch size and the spacer size increase. The randomness test of von Neumann (1941) can be successively applied to spaced batch quantiles with progressively increasing batch sizes and interbatch spacer sizes so that when the randomness test is finally passed, the resulting spaced batch quantiles are approximately independent of each other and of the simulation's initial condition. Recall that the bias of the batch quantiles depends not only on initialization effects but also on the batch size; handling this latter source of bias is discussed in Section 2.

Second, we are studying theoretical properties of the estimators $\widetilde{x}_p(b,m)$ and $S^2_{\widehat{x}_p}(b,m)$ as both the batch size m and the batch count b increase. It turns out that the sample variance $S^2_{\widehat{x}_p}(b,m)$ is not a consistent estimator of the asymptotic variance parameter $\sigma^2_{\widehat{x}_p} \equiv \sigma^2_{I(x_p)} / f^2(x_p)$. These properties will guide us towards the derivation of batching sequences that produce estimators of $\sigma^2_{\widehat{x}_p}$ with minimum asymptotic MSE. While such results have been established for variance estimators relative to the steady-state mean (Chien, Goldsman, and Melamed 1997; Damerdji 1994, 1995), no such results are known for quantile estimation.

REFERENCES

Alexopoulos, C., N. T. Argon, D. Goldsman, N. M. Steiger, G. Tokol, and J. R. Wilson. 2007a. "Efficient Computation of Overlapping Variance Estimators for Simulation". *INFORMS Journal on Computing* 19 (3): 314–327.

Alexopoulos, C., N. T. Argon, D. Goldsman, G. Tokol, and J. R. Wilson. 2007b. "Overlapping Variance Estimators for Simulation". *Operations Research* 55 (6): 1090–1103.

Alexopoulos, C., and A. F. Seila. 1998. "Output Data Analysis". In *Handbook of Simulation*, edited by J. Banks, 225–272. New York: John Wiley & Sons.

Bekki, J. M., J. W. Fowler, G. T. Mackulak, and B. L. Nelson. 2010. "Indirect Cycle Time Quantile Estimation Using the Cornish–Fisher Expansion". *IIE Transactions* 42 (1): 31–44.

Bickel, P. J., and K. A. Doksum. 2007. *Mathematical Statistics: Basic Ideas and Selected Topics*. 2nd ed. Upper Saddle River, NJ: Pearson Prentice Hall.

Billingsley, P. 1999. *Convergence of Probability Measures*. 2nd ed. New York: John Wiley & Sons.

Bradley, R. C. 2005. "Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions". *Probability Surveys* 2:107–144.

Chen, E. J., and W. D. Kelton. 2006. "Quantile and Tolerance-Interval Estimation in Simulation". *European Journal of Operational Research* 168:520–540.

Chen, E. J., and W. D. Kelton. 2008. "Estimating Steady-State Distributions via Simulation-Generated Histograms". *Computers and Operations Research* 35 (4): 1003–1016.

Chien, C.-H., D. Goldsman, and B. Melamed. 1997. "Large-Sample Results for Batch Means". *Management Science* 43:1288–1295.

Cornish, E. A., and R. A. Fisher. 1937. "Moments and Cumulants in the Specification of Distributions". *Revue de l'Institut International de Statistique* 5:307–320.

Damerdji, H. 1994. "Strong Consistency of the Variance Estimator in Steady-State Simulation Output Analysis". *Mathematics of Operations Research* 19:494–512.

Damerdji, H. 1995. "Mean-Square Consistency of the Variance Estimator in Steady-State Simulation Output Analysis". *Operations Research* 43 (2): 282–291.

Engle, R. F. 1982. "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation". *Econometrica* 50 (4):987–1008.

Fishman, G. S. 2001. *Discrete-Event Simulation: Modeling, Programming, and Analysis*. New York: Springer-Verlag.

Fishman, G. S., and L. S. Yarberry. 1997. "An Implementation of the Batch Means Method". *INFORMS Journal on Computing* 9:296–310.

Glasserman, P. 2004. *Monte Carlo Methods in Financial Engineering*. Heidelberg, Germany: Springer-Verlag.

Heidelberger, P., and P. A. W. Lewis. 1984. "Quantile Estimation in Dependent Sequences". *Operations Research* 32:185–209.

Heidelberger, P., and P. D. Welch. 1981. "A Spectral Method for Confidence Interval Generation and Run Length Control in Simulations". *Communications of the ACM* 24:233–245.

Hogg, R. V., J. W. McKean, and A. T. Craig. 2005. *Introduction to Mathematical Statistics*. 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall.

Iglehart, D. L. 1976. "Simulating Stable Stochastic Systems, VI: Quantile Estimation". *Journal of the Association for Computing Machinery* 23:347–360.

Jain, R., and I. Chlamtac. 1985. "The P^2 Algorithm for Dynamic Calculation of Quantiles and Histograms without Storing Observations". *Communications of the ACM* 28 (10): 1076–1085.

Jin, X., M. C. Fu, and X. Xiong. 2003. "Probabilistic Error Bounds for Simulation Quantile Estimators". *Management Science* 49:230–246.

Lada, E. K., N. M. Steiger, and J. R. Wilson. 2008. "SBatch: A Spaced Batch Means Procedure for Steady-State Simulation Analysis". *Journal of Simulation* 2 (3): 170–185.

Lahiri, S. N. 2003. *Resampling Methods for Dependent Data*. New York: Springer-Verlag.

Law, A. M. 2007. *Simulation Modeling and Analysis*. 4th ed. New York: McGraw-Hill.

Meketon, M. S., and B. W. Schmeiser. 1984. "Overlapping Batch Means: Something for Nothing?". In *Proceedings of the 1984 Winter Simulation Conference*, edited by S. Sheppard, U. W. Pooch, and C. D. Pegden, 227–230. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Moore, L. W. 1980. *Quantile Estimation Methods in Regenerative Processes*. Ph. D. thesis, Department of Statistics, University of North Carolina, Chapel Hill, NC.

Muñoz, D. F. 2010. "On the Validity of the Batch Quantile Method for Markov Chains". *Operations Research Letters* 38 (3): 223–226.

Nelson, B. L. 2008. "The MORE Plot: Displaying Measures of Risk & Error from Simulation Output". In *Proceedings of the 2008 Winter Simulation Conference*, edited by S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler, 413–416. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Raatikainen, K. E. E. 1987. "Simultaneous Estimation of Several Percentiles". *Simulation* 49:159–163.

Raatikainen, K. E. E. 1990. "Sequential Procedure for Simultaneous Estimation of Several Percentiles". *Transactions of the Society for Computer Simulation* 7 (1): 21–44.

Schruben, L. W. 1983. "Confidence Interval Estimation Using Standardized Time Series". *Operations Research* 31:1090–1108.

Seila, A. F. 1982a. "A Batching Approach to Quantile Estimation in Regenerative Simulations". *Management Science* 28 (5): 573–581.

Seila, A. F. 1982b. "Estimation of Percentiles in Discrete Event Simulation". *Simulation* 6:193–200.

Shao, J., and D. Tu. 1995. *The Jackknife and Bootstrap*. New York: Springer-Verlag.

Steiger, N. M., E. K. Lada, J. R. Wilson, J. A. Joines, C. Alexopoulos, and D. Goldsman. 2005. "ASAP3: A Batch Means Procedure for Steady-State Simulation Analysis". *ACM Transactions on Modeling and Computer Simulation* 15 (1): 39–73.

Steiger, N. M., and J. R. Wilson. 2001. "Convergence Properties of the Batch-Means Method for Simulation Output Analysis". *INFORMS Journal on Computing* 13 (4): 277–293.

Tafazzoli, A., N. M. Steiger, and J. R. Wilson. 2011a. "N-Skart: A Nonsequential Skewness- and Autoregression-Adjusted Batch Means Procedure for Simulation Analysis". *IEEE Transactions on Automatic Control* 56 (2): 254–264.

Tafazzoli, A., and J. R. Wilson. 2011b. "Skart: A Skewness- and Autoregression-Adjusted Batch Means Procedure for Simulation Analysis". *IIE Transactions* 43 (2): 110–128.

Tafazzoli, A., J. R. Wilson, E. K. Lada, and N. M. Steiger. 2011c. "Performance of Skart: A Skewness- and Autoregression-Adjusted Batch-Means Procedure for Simulation Analysis". *INFORMS Journal on Computing* 23:297–314.

Tong, H. 1990. *Nonlinear Time Series: A Dynamical System Approach*. New York: Oxford University Press.

Tsay, R. S. 1987. "Conditional Heteroscedastic Time Series Models". *Journal of the American Statistical Association* 82 (398):590–604.

von Neumann, J. 1941. "Distribution of the Ratio of the Mean Square Successive Difference to the Variance". *Annals of Mathematical Statistics* 12:367–395.

Wood, D. C., and B. W. Schmeiser. 1995. "Overlapping Batch Quantiles". In *Proceedings of the 1995 Winter Simulation Conference*, edited by C. Alexopoulos, K. Kang, W. R. Lilegdon, and D. Goldsman, 303–308. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Wu, W. B. 2005. "On the Bahadur Representation of Sample Quantiles for Dependent Sequences". *Annals of Statistics* 33 (4): 1924–1963.

Wu, W. B., and X. Shao. 2004. "Limit Theorems for Iterated Random Functions". *Journal of Applied Probability* 41:425–436.

Wu, W. B., and M. Woodroofe. 2000. "A Central Limit Theorem for Iterated Random Functions". *Journal of Applied Probability* 37:748–755.

AUTHOR BIOGRAPHIES

CHRISTOS ALEXOPOULOS is an associate professor in the H. Milton Stewart School of Industrial and Systems Engineering at the Georgia Institute of Technology. His research interests are in the areas of simulation, statistics, and optimization of stochastic systems. He is a member of INFORMS and an active participant in the Winter Simulation Conference, having been Proceedings Co-Editor in 1995, Associate Program Chair in 2006, and a member of the Board of Directors since 2008. He is also an Area Editor of the *ACM Transactions on Modeling and Computer Simulation*. His e-mail address is christos@isye.gatech.edu, and his Web page is www.isye.gatech.edu/∼christos.

DAVID GOLDSMAN is a professor in the H. Milton Stewart School of Industrial and Systems Engineering at the Georgia Institute of Technology. His research interests include simulation output analysis, ranking and selection, and healthcare simulation. He was Program Chair of the Winter Simulation Conference in 1995 and a member of the WSC Board of Directors between 2001–2009. He is currently a trustee of the WSC Foundation. His e-mail address is sman@gatech.edu, and his Web page is www.isye.gatech.edu/∼sman.

JAMES R. WILSON is a professor in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. His current research interests are focused on probabilistic and statistical issues in the design and analysis of simulation experiments, with special emphasis on applications in healthcare and production. As a WSC participant, he served as proceedings editor (1986), associate program chair (1991), and program chair (1992). During the period 1997–2004, he was a member of the WSC Board of Directors. He is a member of ACM, ASA, and SCS; and he is a Fellow of IIE and INFORMS. His e-mail address is jwilson@ncsu.edu, and his Web page is www.ise.ncsu.edu/jwilson.