Performance of a Wavelet-Based Spectral Procedure for Steady-State Simulation Analysis

Emily K. Lada
SAS Institute, Cary, North Carolina, 27513-8617, USA, Emily.Lada@sas.com
James R. Wilson
Department of Industrial Engineering, North Carolina State University, Campus Box 7906, Raleigh, North Carolina 27695-7906, USA, jwilson@ncsu.edu
Natalie M. Steiger
Maine Business School, University of Maine, Orono, Maine 04469-5723, USA, nsteiger@maine.edu
Jeffrey A. Joines
Department of Textile Engineering, Chemistry, and Science, North Carolina State University, Raleigh, North Carolina 27695-8301, USA, jeffjoines@ncsu.edu

A summary and an analysis are given for an experimental performance evaluation of WASSP, an automated wavelet-based spectral method for constructing an approximate confidence interval on the steady-state mean of a simulation output process such that the delivered confidence interval satisfies user-specified requirements on absolute or relative precision as well as coverage probability. The experimentation involved three difficult test problems, each with an output process exhibiting some combination of the following characteristics: a long warm-up period, a persistent autocorrelation structure, or a highly nonnormal marginal distribution. These problems were used to compare the performance of WASSP with that of the Heidelberger-Welch algorithm and ASAP3, two sequential procedures based respectively on the methods of spectral analysis and nonoverlapping batch means. Concerning efficiency (required sample sizes) and robustness against the statistical anomalies commonly encountered in simulation studies, WASSP outperformed the Heidelberger-Welch procedure and compared favorably with ASAP3.

Key words: simulation, statistical analysis; spectral analysis; steady-state analysis; wavelet analysis
History: Accepted by Susan M. Sanchez, Area Editor for Simulation; received April 2004; revised January 2005, June 2005, July 2005; accepted August 2005.

---

1. Introduction

A nonterminating simulation is one in which interest is focused on long-run (steady-state) average performance measures. Usually in a nonterminating probabilistic simulation, the

objective is to compute point and confidence-interval estimators for some parameter, or characteristic, of the steady-state cumulative distribution function (c.d.f.) of a particular simulation-generated response. Lada and Wilson (2005) develop an automated wavelet-based spectral method for constructing an approximate confidence interval (CI) on the steady-state mean of a simulation output process. This procedure, called WASSP, determines first a batch size and a truncation point (the end of the warm-up period, also called the statistics-clearing time) beyond which successive batch means form an approximately stationary Gaussian process. For this purpose WASSP uses the randomness test of von Neumann (1941) to determine the size of the spacer (the number of ignored observations) preceding each batch that is sufficiently large to ensure the resulting spaced batch means are approximately independent and identically distributed (i.i.d.). In this situation the spacer preceding the first batch must contain the warm-up period and hence defines an appropriate truncation point; and then WASSP uses the univariate normality test of Shapiro and Wilk (1965) to determine a batch size that is sufficiently large to ensure the spaced batch means are approximately normal.

Next WASSP computes the discrete wavelet transform of the bias-corrected log-smoothed-periodogram of the truncated, nonspaced batch means (i.e., the batch means computed from adjacent nonoverlapping batches observed beyond the truncation point); and the resulting wavelet coefficients are denoised by applying a soft-thresholding scheme. Then by computing the inverse discrete wavelet transform of the thresholded wavelet coefficients, WASSP delivers an estimator of the batch means log-spectrum and ultimately the steady-state variance parameter (SSVP) of the original (unbatched) process—i.e., the sum of the covariances at all lags for the original process. Finally WASSP combines the estimator of the SSVP with the grand average of the truncated batch means in a sequential procedure for constructing a CI estimator of the steady-state mean satisfying user-specified requirements on absolute or relative precision as well as coverage probability.

This article contains a summary of some experimental results exemplifying the performance observed in applying WASSP and other selected procedures for steady-state simulation output analysis to a suite of particularly difficult test problems. These test problems were designed to explore the following characteristics of the selected output-analysis procedures:

- the efficiency of each procedure in terms of the sample size required to deliver a CI that is supposed to attain the user-specified levels of precision and coverage probability; and

- the robustness of each procedure against the statistical anomalies commonly encountered in the analysis of outputs generated by large-scale steady-state simulation experiments (in particular, initialization bias, correlation, and nonnormality).

The experimental performance evaluation is focused on the following test problems:

1. the $M/M/1$ queue waiting-time process for which the underlying system has an arrival rate of 0.90, a service rate of 1, and an empty-and-idle initial condition;

2. the first-order autoregressive (AR(1)) process with a lag-one correlation of 0.995, a white-noise variance of 1, a steady-state mean of 100, and an initial value of 0; and

3. the "AR(1)-to-Pareto" (ARTOP) process that has marginals given by a Pareto distribution with lower limit and shape parameter equal to 1 and 2.1, respectively (implying the marginal mean and variance are both finite while the marginal skewness and kurtosis are both infinite), and that is obtained by applying to a standardized, stationary version of process 2 above the composite of (a) the inverse of the specified Pareto c.d.f., and (b) the standard normal c.d.f.

For each of the above test problems, the following criteria were used to evaluate the performance of WASSP and its competitors: (i) the empirical coverage probability of the delivered CIs; (ii) the mean and variance of the half-lengths of the delivered CIs; (iii) the average relative precision of the delivered CIs; and (iv) the mean of the total required sample sizes. Independent replications were performed for each simulation-analysis procedure to construct nominal 90% and 95% CIs satisfying a given relative-precision requirement, which is an upper bound on the acceptable CI half-length specified as a maximum percentage of the magnitude of the final point estimator as detailed in (6) below. Confidence intervals were also constructed for the no-precision case—i.e., the case in which there was no upper bound on the CI half-length so the final CI delivered by WASSP was based on the batch count and batch size required to pass the randomness and normality tests. For each test problem, the theoretical steady-state mean response is available analytically; thus the performances of WASSP and its competitors were evaluated in terms of actual versus nominal CI coverage probabilities as well as sample sizes and half-lengths of the CIs delivered by each procedure. For comparison, the spectral method of Heidelberger and Welch (1983) and the batch-means procedure ASAP3 (Steiger et al. 2005) were also applied to each test problem.

The rest of this article is organized as follows. Section 2 contains the notation required for the performance evaluation together with brief summaries of the Heidelberger-Welch (H&W) and ASAP3 procedures. Section 3 contains detailed descriptions of test problems 1–3 as well as a discussion of the results of applying WASSP and its competitors to these test problems. Finally Section 4 provides a summary of the main findings of this research and recommendations for future work. Lada (2003) and the Online Supplement to the present paper on the journal's Web site provide a complete discussion of the experimental performance evaluation of WASSP. Some preliminary results on the formulation and evaluation of WASSP are presented in Lada et al. (2003, 2004a). A stand-alone Windows-based version of WASSP and a user's manual are available online via Lada et al. (2004b).

2. Simulation-Analysis Methods to Be Compared with WASSP

Although the notation and terminology of Lada and Wilson (2005) is used throughout the article, for completeness this section contains a summary of the most frequently used notation along with overviews of the H&W method and ASAP3. If $\{X_u : u = 1, \ldots, n\}$ is a covariance-stationary simulation output process for which the objective is to compute point and CI estimators of the mean $\mu_X = \mathrm{E}[X_u]$, then the covariance at lag ℓ for this process is $\gamma_X(\ell) = \mathrm{E}[(X_u - \mu_X)(X_{u+\ell} - \mu_X)]$ for $\ell = 0, \pm 1, \pm 2, \ldots$ and $u = 1, 2, \ldots$; and the SSVP of the process is

$$\gamma_X = \sum_{\ell=-\infty}^{\infty} \gamma_X(\ell), \tag{1}$$

where the right-hand side of (1) is assumed to be absolutely convergent so γ_X is well defined. Moreover, let $\overline{X}_j(m) = m^{-1} \sum_{u=(j-1)m+1}^{jm} X_u$ denote the jth batch mean for batches of size m computed from the process $\{X_u : u = 1, \ldots, n\}$ for $j = 1, \ldots, k = \lfloor n/m \rfloor$. Let $\overline{\overline{X}} = \overline{\overline{X}}(m, k) = k^{-1} \sum_{j=1}^{k} \overline{X}_j(m)$ denote the grand mean computed over all k batches of size m.

If the process $\{X_u\}$ is covariance-stationary, then the power spectrum $p_X(\omega)$ of this process is given by the cosine transform of the covariance function $\gamma_X(\ell)$,

$$p_X(\omega) = \sum_{\ell=-\infty}^{\infty} \gamma_X(\ell) \cos(2\pi\omega\ell) \quad \text{for} \quad -\tfrac{1}{2} \le \omega \le \tfrac{1}{2}. \tag{2}$$

At frequency $\omega = 0$, equation (2) yields $p_X(0) = \sum_{\ell=-\infty}^{\infty} \gamma_X(\ell) = \gamma_X$. In using a spectral method to analyze the time series $\{X_u : u = 1, \ldots, n\}$ of length n, the first step is to compute

the periodogram

$$I\left(\tfrac{\ell}{n}\right) = \frac{1}{n}\left\{\left[\sum_{u=1}^{n} X_u \cos\left(\tfrac{2\pi(u-1)\ell}{n}\right)\right]^2 + \left[\sum_{u=1}^{n} X_u \sin\left(\tfrac{2\pi(u-1)\ell}{n}\right)\right]^2\right\} \quad \text{for} \quad \ell = 1, \ldots, n-1 \quad (3)$$

as an estimator of $p_X\left(\frac{\ell}{n}\right)$ at the Fourier frequency $\frac{\ell}{n}$ cycles per time unit for $\ell = 1, \ldots, n-1$. An appropriate extrapolation of (3) to zero frequency then yields an estimator of $p_X(0)$.

Both the H&W procedure and WASSP are designed to deliver a spectral estimator $\widehat{\gamma}_X$ of γ_X for computing a $100(1-\beta)\%$ CI estimator of μ_X having the form

$$\overline{\overline{X}} \pm H, \quad \text{with half-length} \quad H = t_{1-\beta/2,\nu}\sqrt{\widehat{\gamma}_X \big/ n'}, \quad (4)$$

where: (a) n' is the length of the truncated output process after deleting (if necessary) a warm-up period containing initialization bias; (b) the grand mean $\overline{\overline{X}}$ and the SSVP estimator $\widehat{\gamma}_X$ are computed from the truncated output process; (c) ν denotes the "effective" degrees of freedom (d.f.) associated with $\widehat{\gamma}_X$; and (d) $t_{1-\beta/2,\nu}$ denotes the $1-\beta/2$ quantile of Student's t-distribution with ν d.f., provided $0 < \beta < 1$.

In WASSP the user may specify that the CI (4) must satisfy a precision requirement expressed in terms of either of the following quantities:

- a maximum acceptable half-length H^* (for an absolute-precision requirement) so that if the latest CI (4) computed by WASSP satisfies the stopping rule

  $$H \le H^*, \quad (5)$$

  then WASSP delivers (4) as the final CI estimator for μ_X and terminates; or

- a maximum acceptable fraction r^* of the magnitude of the CI midpoint (for a relative-precision requirement) so that if the latest CI (4) computed by WASSP satisfies the stopping rule

  $$H \le r^*\left|\overline{\overline{X}}\right|, \quad (6)$$

  then WASSP delivers (4) as the final CI estimator for μ_X and terminates.

Because the H&W procedure was apparently designed only for use with a relative-precision specification, in this article all the experiments are based on a stopping rule of the form (6).

2.1. Overview of Heidelberger and Welch's Spectral Method

Heidelberger and Welch (1981ab, 1983) develop a spectral method for steady-state simulation analysis in which they use standard regression techniques to estimate the power spectrum (2) of the given output process at zero frequency. Heidelberger and Welch estimate γ_X by fitting a quadratic polynomial to the logarithm of a smoothed version of the periodogram (3) for the given output process over the frequency range between 0 and $\frac{1}{2}$ cycles per time unit (excluding the endpoints), where the smoothing operation consists of averaging nonoverlapping pairs of periodogram values. The resulting SSVP estimator is then used to compute a CI of the form (4) for μ_X.

Comparing the performance of WASSP and the H&W procedure is complicated because the latter requires the user to specify an upper limit $t_{\max}$ on the allowable length of a given test process. (To avoid confusion in this section and throughout the rest of the article, the notation of Heidelberger and Welch is always used when referring to the H&W procedure.) For a fair comparison of WASSP with the H&W procedure, first WASSP is applied to the test process so as to obtain not only the corresponding WASSP-generated CI of the form (4) but also a complete (untruncated) time series $\{X_u : n = 1, \ldots, n\}$ to which the (partially) sequential version of the H&W procedure can be applied after taking $t_{\max} = n$, the length of the simulation-generated time series, for the current replication of the H&W procedure.

Heidelberger and Welch (1983) describe a scheme for batching data prior to applying their spectral method, and this scheme is used in the implementation of the H&W procedure. The batch count k for H&W is always in the range $L \le k \le 2L$, where the value $L = 200$ is used to conform to the recommendations of Heidelberger and Welch (1983). Within each replication, let t_i denote the "time"—i.e., the current (untruncated) sample size—at the ith checkpoint in the analysis of a given output process, where $t_1 = \left\lceil 0.15\, t_{\max} \right\rceil$ and $t_i = \min\left\{ \left\lceil 1.5\, t_{i-1} \right\rceil, t_{\max} \right\}$ for $i = 2, 3, \ldots$. If $t_i \ge L$ and the assignment $b_i = \left\lfloor \log_2\{(t_i - 1)/L\} \right\rfloor$ is made, then at the ith checkpoint the batch size m_i and the number of batches k_i are given by $m_i = 2^{b_i}$ and $k_i = \left\lfloor t_i/m_i \right\rfloor$, respectively.

The version of H&W examined in this article uses the method for detecting and eliminating initialization bias described in Heidelberger and Welch (1983). At the ith checkpoint (for $i = 1, 2, \ldots$), H&W tests the null hypothesis that the untruncated batch-means process $\left\{ \overline{X}_j(m_i) : j = 1, \ldots, k_i \right\}$ is covariance-stationary by computing the Cramér–von Mises (CVM) test statistic, $\mathrm{CVM}(m_i, k_i) = \left[\sum_{j=0}^{k_i - 1} D_{ij}^2 \right] \Big/ \left[k_i^2 \widehat{p}_{\overline{X}(m_i)}(0) \right]$, where: (a) for each fre-

quency ω in a neighborhood of zero, let $\widehat{p}_{\overline{X}(m_i)}(\omega)$ denote the H&W estimator of the power spectrum $p_{\overline{X}(m_i)}(\omega)$ of the untruncated batch means process, with $p_{\overline{X}(m_i)}(\omega)$ defined similarly to the power spectrum (2) of the original (unbatched) process; and (b) let $D_{i0} = 0$ and $D_{ij} = \sum_{u=1}^{j} \left[\overline{X}_u(m_i) - \overline{\overline{X}}(m_i, k_i) \right]$ for $j = 1, \ldots, k_i$.

If the untruncated batch-means process $\left\{ \overline{X}_j(m_i) : j = 1, \ldots, k_i \right\}$ is covariance-stationary, then under widely applicable conditions as $m_i \to \infty$ and $k_i \to \infty$, the asymptotic distribution of $\mathrm{CVM}(m_i, k_i)$ is equal to the c.d.f. of $\mathrm{CVM}(\mathcal{B}) = \int_0^1 \mathcal{B}^2(u)\, \mathrm{d}u$, where $\{\mathcal{B}(u) : u \in [0, 1]\}$ is a Brownian-bridge process. Thus if the untruncated batch means-process is covariance-stationary, then the asymptotic 0.9 quantile of the CVM test statistic is $\mathrm{CVM}(\mathcal{B})_{0.9} = 0.3473$; see Table 1 of Anderson and Darling (1952). If $\mathrm{CVM}(m_i, k_i) > 0.3473$, then the CVM test has detected nonstationarity (initialization bias) in the untruncated sequence of batch means so H&W deletes the initial 10% of this sequence and recomputes the CVM test statistic from the truncated sequence of batch means.

After each repetition of the CVM test that detects nonstationarity at the ith checkpoint, H&W tries to delete an additional 10% of the current untruncated sequence of batch means before repeating the CVM test on the truncated batch means. If the CVM test is failed six times, then H&W tries to advance to the next checkpoint so the current (untruncated) sample size is increased by 50% before the batch size, batch count, and untruncated batch-means sequence are all updated. The CVM test is repeated at successive checkpoints with warm-up periods (truncation points) ranging from 0% to 50% of the untruncated batch-means sequence until either (a) the CVM test is passed and a CI of the form (4) satisfying (6) is computed from the truncated batch means; or (b) the untruncated sample size required by H&W reaches the upper limit $t_{\max}$. If case (b) holds, then the CVM test is performed one last time. If the final CVM test for case (b) is failed, then H&W terminates without delivering a CI; otherwise H&W terminates after delivering a CI of the form (4) that might not satisfy (6). In conformance with the recommendations of Heidelberger and Welch (1981ab, 1983), in this article the batch-means log-spectrum is estimated by fitting a quadratic polynomial to the first 25 points on the log-smoothed-periodogram of the batch means. Thus in the H&W-generated CI of the form (4), the quantity ν denoting the effective degrees of freedom is given by $\nu = 7$ d.f.

2.2. Overview of ASAP3

Steiger et al. (2005) formulated ASAP3 as an improved variant of the batch-means algorithms ASAP (Steiger and Wilson 2002) and ASAP2 (Steiger et al. 2002) for steady-state simulation analysis. ASAP3 operates as follows: the batch size is progressively increased until spaced groups of four adjacent batch means pass the Shapiro-Wilk test for four-dimensional normality, where the spacer preceding each group also consists of four adjacent batch means; and then after skipping the first spacer as the warm-up period, ASAP3 fits a first-order autoregressive (AR(1)) time series model to the truncated, nonspaced batch means. If necessary, the batch size is further increased until the autoregressive parameter in the AR(1) model does not significantly exceed 0.8. Next ASAP3 computes the terms of an inverse Cornish-Fisher expansion for the classical batch-means t-ratio based on the AR(1) parameter estimates; finally ASAP3 delivers a correlation-adjusted CI based on this expansion. ASAP3 is a sequential procedure designed to deliver a CI satisfying a user-specified precision requirement of the form (5) or (6).

3. Test Problems Used in the Performance Evaluation

3.1. The $M/M/1$ Queue Waiting-Time Process

For the first test problem, let X_u denote the waiting time in the queue for the uth customer ($u = 1, 2, \ldots$) in a single-server queueing system with i.i.d. exponential interarrival times having mean 10/9 (so the arrival rate $\lambda = 0.9$); i.i.d. exponential service times having mean 1 (so the service rate $\mu = 1$); steady-state server utilization $\tau = \lambda/\mu = 0.9$; and an empty-and-idle initial condition (so $X_1 = 0$). The steady-state mean for this process is $\mu_X = 9.0$.

The selected $M/M/1$ queue waiting-time process is a particularly difficult test case for the following reasons. (a) Because the system starts empty and idle, both the magnitude and duration of the initial transient in the process $\{X_u : u = 1, 2, \ldots\}$ are pronounced. (b) Once the system has reached steady-state operation, the autocorrelation function of the appropriately truncated process $\{X_u\}$ decays very slowly with increasing lags. (c) The steady-state marginal distribution of waiting times is markedly nonnormal, having an atom at zero and an exponential tail. It follows from (a)–(c) that the $M/M/1$ queue waiting-time process is a suitable test problem for thorough evaluation of the effectiveness of WASSP's independence and normality tests in determining both an appropriate batch size and an

appropriate truncation point beyond which successive batch means approximately constitute a covariance-stationary Gaussian process.

If $\{X_u\}$ is in steady-state operation, then the associated power spectrum is given by

$$p_{_X}(\omega) = \frac{\tau^3(2-\tau)}{(1-\tau)^2\mu^2} + \frac{1-\tau^2}{\pi\mu^2}\int_0^r \frac{t^{5/2}(r-t)^{1/2}[\cos(2\pi\omega)-t]}{(1-t)^3[1-2t\cos(2\pi\omega)+t^2]}\,\mathrm{d}t \tag{7}$$

for $\omega \in \left[-\frac{1}{2}, \frac{1}{2}\right]$, where $r = 4\tau\big/(1+\tau)^2$. Since the literature seems to lack readily available computing formulas for $p_{_X}(\omega)$, the result (7) is derived in Appendix A of the Online Supplement and in Appendix D of Lada (2003).

Figure 1 displays $\ln\left[p_{_X}(\omega)\right]$, the log-spectrum of the original (unbatched) process $\{X_u\}$, for $\omega \in \left[-\frac{1}{2}, \frac{1}{2}\right]$. WASSP, however, estimates $\ln\left[p_{\overline{X}(m)}(\omega)\right]$, the log-spectrum of the batch means process $\left\{\overline{X}_j(m)\right\}$. Figure 1 provides a general idea of the shape of $\ln\left[p_{\overline{X}(m)}(\omega)\right]$ since the peakedness of this function at zero frequency depends on the peakedness of $\ln\left[p_{_X}(\omega)\right]$ at that point. While $\ln\left[p_{\overline{X}(m)}(\omega)\right]$ will be less peaked than $\ln\left[p_{_X}(\omega)\right]$ because of the averaging operation performed on each batch, the batch-means log-spectrum will still be sharply peaked; and this characteristic enables assessment of the robustness of WASSP's wavelet-based technique for estimating $\gamma_{_X}$.

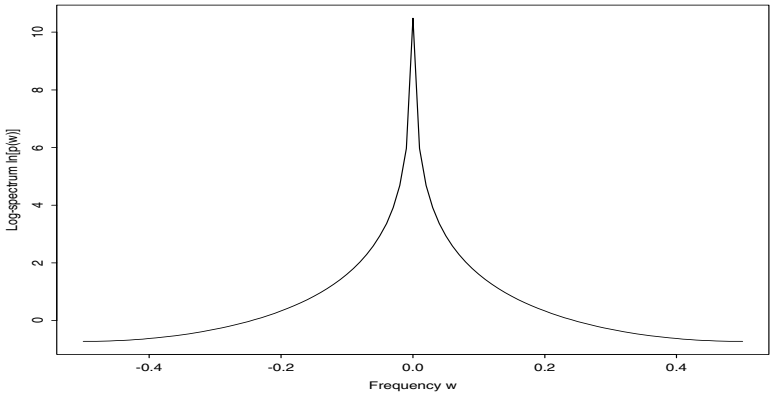

Figure 1: Log-Spectrum $\ln[p_{_X}(\omega)]$ of the Steady-State $M/M/1$ Queue Waiting-Time Process for Frequency $\omega \in \left[-\frac{1}{2}, \frac{1}{2}\right]$

From the batch means $\left\{\overline{X}_j(m) : j = 1, \ldots, k\right\}$, the associated periodogram $I_{\overline{X}(m)}\left(\frac{\ell}{k}\right)$ is computed at frequency $\frac{\ell}{k}$ (for $\ell = 1, \ldots, k-1$) in the same way the periodogram (3) is computed from the original (unbatched) process; in WASSP the batch-means periodogram is smoothed by computing a moving average of $A = 2a+1$ points, where $a \in \{2, 3, 4, 5\}$. As explained in Section 3.3.3 of Lada (2003) and in Section 4.4.1 of Lada and Wilson (2005),

at zero frequency the resulting smoothed-periodogram value equals $a^{-1}\sum_{\ell=1}^{a} I_{\overline{X}(m)}\left(\frac{\ell}{k}\right)$. As the overall sample size $n \to \infty$ with a fixed batch size m so the batch count $k \to \infty$, the smoothed-periodogram value at zero frequency is (i) asymptotically independent of the grand average of the batch means; and (ii) approximately a chi-squared random variable with $2a$ d.f. that has been scaled by the multiplier $p_{\overline{X}(m)}(0)\big/(2a)$, where $p_{\overline{X}(m)}(0) = p_X(0)\big/m = \gamma_X\big/m$. Thus WASSP's $100(1-\beta)\%$ CI for μ_X of the form (4) is based on the $1-\beta/2$ quantile of Student's t-distribution with $\nu = 2a$ d.f. The user selects the smoothing parameter $A \in \{5, 7, 9, 11\}$, with the default being $A = 7$ so WASSP's CIs are based on $\nu = 6$ d.f. by default.

Table 1 shows the performance of WASSP for the $M/M/1$ queue waiting-time process using the smoothing parameter values $A = 5, 7, 9,$ and 11. The results are based on 1,000 independent replications of nominal 90% CIs. For each coverage estimator in Table 1, the standard error is less than 1%. Table 1 shows that the coverage probability usually decreased as A increased. This behavior is due to the target process having a power spectrum with a sharp peak at zero frequency. As A was increased, WASSP's estimate of the batch-means power spectrum near zero frequency became flatter than it should have been; this behavior ultimately resulted in underestimation of γ_X. In general, for small-sample cases (specifically, if one were only interested in generating an initial, or pilot, CI for the steady-state mean of a particular process without imposing a precision requirement), it might be desirable to change the smoothing parameter from the default value $A = 7$ to $A = 5$. However, for the $\pm 7.5\%$ precision case, there was significant CI overcoverage for $A = 5$. On the basis of all the experimentation with WASSP on the $M/M/1$ queue and other test problems, it is recommended to use the default value $A = 7$ for the smoothing parameter in applications of WASSP involving a nontrivial precision requirement and in the absence of additional information relevant to setting A.

For the $M/M/1$ queue waiting-time process, Table 2 shows a comparison of the performance of the following procedures: WASSP (using $A = 7$), ASAP3, and H&W as specified in Heidelberger and Welch (1983). Analyzing multiple replications of each test process required a special version of the ASAP3 software. Because of the extensive disk-space requirements of the simulation-generated data sets processed by this software, at most 400 replications of ASAP3 could be performed for the given test processes; thus the reported coverage probabilities for ASAP3 had standard errors of approximately 1.5% and 1% for nominal coverages of 90% and 95%, respectively. Since 1,000 replications were performed for WASSP and H&W,

Table 1: Performance of WASSP Using Different Values of the Smoothing Parameter A in the $M/M/1$ Queue Waiting-Time Process Based on 1,000 Replications of 90% CIs

Precision Requirement	Performance Measure	Smoothing Parameter $A = 5$	$A = 7$	$A = 9$	$A = 11$
None	CI coverage	88.8%	87.7%	86.1%	84.2%
	Avg. sample size	18,369	18,090	17,696	18,369
	Avg. relative prec.	0.378	0.341	0.323	0.297
	Avg. CI half-length	3.40	3.07	2.91	2.67
	Var. CI half-length	2.65	2.00	1.62	1.25
±15%	CI coverage	89.6%	87.2%	83.5%	82.8%
	Avg. sample size	114,710	92,049	79,824	68,533
	Avg. relative prec.	0.122	0.123	0.124	0.128
	Avg. CI half-length	1.10	1.11	1.12	1.15
	Var. CI half-length	0.0414	0.0387	0.0381	0.0340
±7.5%	CI coverage	93.6%	90.4%	88.5%	91.5%
	Avg. sample size	467,370	388,000	341,380	322,990
	Avg. relative prec.	0.065	0.065	0.065	0.066
	Avg. CI half-length	0.585	0.587	0.586	0.591
	Var. CI half-length	0.0072	0.0072	0.0067	0.0060

the coverage probabilities for both procedures had standard errors of approximately 0.95% and 0.69% for nominal coverages of 90% and 95%, respectively.

To account properly for situations in which H&W runs out of data before satisfying the precision requirement, Table 2 includes the following CI coverage probabilities:

1. the *net CI coverage*, defined as the fraction of all replications performed in which the delivered CI not only covered the steady-state mean μ_X but also satisfied the precision requirement (6); and

2. the *coverage of CIs satisfying the precision requirement*, defined as the fraction of only those replications satisfying the precision requirement in which the delivered CI covered μ_X.

Although in theory H&W could terminate without delivering a CI, in practice this behavior was never observed in our experimentation. Table 2 reveals that in the no-precision case, WASSP and ASAP3 yielded similar results in terms of CI coverage; however, in this case WASSP's average sample size was 42% smaller than that of ASAP3.

Table 2 shows that the net-coverage probabilities of the CIs delivered by H&W were significantly below not only their specified nominal levels but also the net-coverage probabilities

Table 2: Performance of WASSP (Using $A = 7$), ASAP3, and H&W in the $M/M/1$ Queue Waiting-Time Process

Precision		Nominal 90% CIs			Nominal 95% CIs		
Requirement	Performance Measure	WASSP	H&W	ASAP3	WASSP	H&W	ASAP3
None	# replications	1,000	1,000	400	1,000	1,000	400
	Net CI coverage	87.7%	67.8%	87.5%	93.4%	76.2%	91.5%
	Avg. sample size	18,090	2,714	31,181	17,971	2,696	31,181
	Avg. relative prec.	0.341	0.450	0.239	0.444	0.576	0.290
	Avg. CI half-length	3.07	4.05	2.07	4.00	5.18	2.52
	Var. CI half-length	2.00	4.46	0.348	3.70	8.00	0.535
	# replications satisfying prec. reqt.	1,000	1,000	400	1,000	1,000	400
	Coverage of CIs satisfying prec. reqt.	87.7%	67.8%	87.5%	93.4%	76.2%	91.5%
±15%	# replications	1,000	1,000	400	1,000	1,000	400
	Net CI coverage	87.2%	76.0%	91%	93%	83.4%	95.5%
	Avg. sample size	92,049	62,112	103,742	143,920	98,838	140,052
	Avg. relative prec.	0.123	0.128	0.134	0.126	0.129	0.136
	Avg. CI half-length	1.11	1.15	1.18	1.13	1.16	1.21
	Var. CI half-length	0.0387	0.0406	0.0259	0.0314	0.0347	0.0205
	# replications satisfying prec. reqt.	1,000	939	400	1,000	944	400
	Coverage of CIs satisfying prec. reqt.	87.2%	80.9%	91%	93%	88.4%	95.5%
±7.5%	# replications	1,000	1,000	400	1,000	1,000	400
	Net CI coverage	90.4%	77%	89.5%	97%	83.5%	94%
	Avg. sample size	388,000	275,610	287,568	598,020	431,590	382,958
	Avg. relative prec.	0.065	0.066	0.070	0.066	0.066	0.070
	Avg. CI half-length	0.587	0.590	0.627	0.595	0.590	0.632
	Var. CI half-length	0.0072	0.0072	0.0023	0.0056	0.0078	0.0020
	# replications satisfying prec. reqt.	1,000	918	400	1,000	917	400
	Coverage of CIs satisfying prec. reqt.	90.4%	83.9 %	89.5%	97%	91.1%	94%

of the CIs delivered by WASSP and ASAP3. For example in the case of nominal 95% CIs with a required precision of ±7.5%, H&W delivered a net CI coverage probability of 83.5% while WASSP delivered a net CI coverage probability of 97%.

For a clearer indication of its asymptotic performance as the required relative precision $r^* \to 0$, WASSP was also applied to the $M/M/1$ process with relative-precision requirements of 3.75% and 1.875%. For 1,000 replications at the 3.75% precision level, WASSP delivered (a) a net CI coverage probability of 94% and an average relative precision of 3.4% for nominal 90% CIs; and (b) a net CI coverage probability of 97.7% and an average relative precision of 3.4% for nominal 95% CIs. For 400 replications of nominal 90% CIs at the 1.875% precision level, WASSP delivered a net CI coverage of 94% and an average relative precision of 1.71%. These results, along with those in Table 2, suggest that in the given $M/M/1$ queue waiting-time process, the CI coverage delivered by WASSP should stabilize slightly above the nominal level as the precision requirement approaches zero.

To investigate the causes of the poor performance of H&W relative to WASSP, estimates

were obtained for the bias, variance, and mean squared error of the final point estimator $\overline{\overline{X}}(m,k)$ delivered by both procedures for each selected combination of CI nominal coverage and required precision. (ASAP3 was omitted from this analysis because its performance is thoroughly examined in Steiger et al. 2005; moreover, since the other two procedures operated on exactly the same data sets as explained in the second paragraph of Section 2.1, it was natural to limit the comparison to those procedures.) The following statistics were computed for WASSP and the H&W procedure:

$$\widehat{\text{Bias}}[\,\overline{\overline{X}}(m,k)] = \left[\frac{1}{R}\sum_{u=1}^{R}\overline{\overline{X}}_u(m_u,k_u)\right] - \mu_X, \quad \widehat{\text{MSE}}[\,\overline{\overline{X}}(m,k)\,] = \frac{1}{R}\sum_{u=1}^{R}[\,\overline{\overline{X}}_u(m_u,k_u) - \mu_X\,]^2, \quad (8)$$

where: (a) the variable R denotes the number of replications with the selected nominal coverage probability that satisfied the given precision requirement; (b) for the uth such replication ($u = 1,\ldots,R$), the random variable $\overline{\overline{X}}_u(m_u,k_u)$ denotes the associated grand average of the truncated batch means based on k_u batches of size m_u; and (c) the random variable $\widehat{\text{Var}}\big[\,\overline{\overline{X}}(m,k)\,\big]$ denotes the sample variance of the truncated batch means $\big\{\overline{\overline{X}}_u(m_u,k_u) : u = 1,\ldots,R\big\}$.

Table 3 shows the estimated absolute bias, variance, and mean squared error for the final point estimators delivered by WASSP and H&W in the $M/M/1$ queue waiting-time process. In the case of no precision requirement, all three performance measures for the final point estimator $\overline{\overline{X}}(m,k)$ delivered by H&W were substantially larger than the corresponding quantities for WASSP. In the no-precision case, the CVM test failed to yield significant reductions in initialization bias when it was used with H&W. Moreover, Table 2 shows that in the no-precision case, H&W required much smaller final sample sizes than did WASSP; in Table 3, this behavior is reflected in much larger point-estimator variances for H&W compared with the corresponding quantities for WASSP.

In those applications of H&W without a precision requirement, the CVM test was often passed at relatively small values of both the total (untruncated) sample size t_i and the associated truncation point; as a result, the truncated time series used to construct the delivered CIs of the form (4) were in general neither sufficiently long nor sufficiently free of initialization bias to yield accurate estimates of μ_X. Table 3 shows that once a precision requirement was imposed on each procedure and the sample size began to increase, the bias, variance, and mean squared error of $\overline{\overline{X}}(m,k)$ began to decrease for both procedures.

Table 3: Mean Squared Error, Variance, and Absolute Bias of $\overline{\overline{X}}(m,k)$ Based on 1,000 Runs the $M/M/1$ Queue Waiting-Process

Precision Requirement	Performance Measure	Nominal 90% CIs WASSP	Nominal 90% CIs H&W	Nominal 95% CIs WASSP	Nominal 95% CIs H&W
None	$\widehat{\text{MSE}}[\,\overline{\overline{X}}(m,k)\,]$	2.58	10.7	2.97	12.4
	$\widehat{\text{Var}}[\,\overline{\overline{X}}(m,k)\,]$	2.53	10.3	2.97	12.3
	$\lvert\widehat{\text{Bias}}[\,\overline{\overline{X}}(m,k)\,]\rvert$	0.231	0.601	0.0728	0.288
±15%	$\widehat{\text{MSE}}[\,\overline{\overline{X}}(m,k)\,]$	0.629	0.850	0.417	0.550
	$\widehat{\text{Var}}[\,\overline{\overline{X}}(m,k)\,]$	0.570	0.768	0.382	0.505
	$\lvert\widehat{\text{Bias}}[\,\overline{\overline{X}}(m,k)\,]\rvert$	0.244	0.288	0.186	0.212
±7.5%	$\widehat{\text{MSE}}[\,\overline{\overline{X}}(m,k)\,]$	0.116	0.187	0.0703	0.314
	$\widehat{\text{Var}}[\,\overline{\overline{X}}(m,k)\,]$	0.113	0.180	0.0694	0.308
	$\lvert\widehat{\text{Bias}}[\,\overline{\overline{X}}(m,k)\,]\rvert$	0.0574	0.0872	0.0300	0.0819

However, Tables 2 and 3 also show that WASSP outperformed H&W with respect to point-estimator accuracy and precision as well as CI coverage, precision, and stability.

Remark. The Online Supplement contains an extensive comparison of the performance of two versions of H&W: (*a*) the original version without the CVM test as formulated in Heidelberger and Welch (1981a); and (*b*) the extended version with the CVM test as described in Heidelberger and Welch (1983) and as used throughout the present paper. Even with a nontrivial precision requirement, incorporating the CVM test did not significantly improve the performance of H&W in terms of any of the following criteria: net CI coverage; CI half-length; and bias, variance, and mean squared error of the final point estimator. Although the overall performance of the extended H&W procedure (with the CVM test) was found to be only slightly better than the performance of the original H&W procedure (without the CVM test) in some experiments, the extended H&W procedure is referenced and used far more frequently than is the original (Pawlikowski 1990).

3.2. The First-Order Autoregressive (AR(1)) Process

If $\{\delta_u : u = 1, 2, \ldots\} \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma_\delta^2\right)$ is a white-noise process, then a first-order autoregressive (AR(1)) process $\{X_u : u = 1, 2, \ldots, \}$ with the starting value X_0 can be generated as

$$X_u = \mu_X + \rho(X_{u-1} - \mu_X) + \delta_u, \quad \text{for} \quad u = 1, 2, \ldots, \tag{9}$$

where μ_X is the mean and ρ is the lag-one correlation of the process in steady-state operation. The parameters of the process (9) were assigned as follows: the mean $\mu_X = 100$, the autoregressive parameter $\rho = 0.995$, and the white-noise variance $\sigma_\delta^2 = 1$. Moreover, the initial condition $X_0 = 0$ was used to obtain the analogue of the "empty-and-idle" initial condition for the $M/M/1$ queue. The most difficult aspects of this test process are its exceptionally long initial-transient period and its persistent autocorrelation structure. On the other hand, the batch means computed from this process are always multivariate normal.

The spectrum of the steady-state AR(1) process (9) is $p_{_X}(\omega) = \sigma_\delta^2 \Big/ \Big[1 - 2\rho\cos(2\pi\omega) + \rho^2\Big]$ for $\omega \in \Big[-\frac{1}{2}, \frac{1}{2}\Big]$; see Section 4.3 of Lada (2003). The variance of the process is $\sigma_X^2 = \sigma_\delta^2 \Big/ \Big(1 - \rho^2\Big) = 100.25$ while the SSVP is $\gamma_{_X} = p_{_X}(0) = \sigma_\delta^2 \Big/ (1-\rho)^2 = 40{,}000$. For $\omega \in \Big[-\frac{1}{2}, \frac{1}{2}\Big]$, the log-spectrum $\ln\Big[p_{_X}(\omega)\Big]$ exhibits peakedness at zero frequency similar to that exhibited by its counterpart for the $M/M/1$ waiting-time process; this property resulted in more pronounced underestimation of the SSVP with increasing values of WASSP's smoothing parameter A.

For the given AR(1) process, Table 4 shows a comparison of the performance of WASSP (using $A = 7$), ASAP3, and the H&W spectral method. In some cases, the actual precision levels of the CIs delivered by these procedures were significantly smaller than the corresponding nominal levels. For example, in the case of 90% CIs with no precision requirement, WASSP, H&W, and ASAP3 delivered CIs with average relative precision levels of 5.3%, 13.4%, and 2.3%, respectively; as a consequence of this behavior, the results for the relative precision levels of ±15% and ±7.5% were essentially the same as for the no-precision case. To provide a meaningful side-by-side comparison of the performance of WASSP, H&W, and ASAP3 in this test process, Table 4 displays the results for the following levels of relative precision: no precision, ±3.75%, ±1.875%, and ±0.9375%.

Regarding conformance to the precision and coverage-probability requirements for the delivered CIs as summarized in Table 4, WASSP outperformed ASAP3 in the cases of no precision and ±3.75% precision while requiring substantially smaller sample sizes than ASAP3 required. For the precision levels ±1.875% and ±0.9375%, WASSP and ASAP3 achieved reasonable conformance to the requested precision levels but exhibited significant CI overcoverage while requiring roughly the same sample sizes. The cause of this overcoverage is the subject of ongoing research.

The results for H&W were obtained in the same way as described in Section 3.1. For the no-precision case, H&W-based CIs with nominal coverage probabilities of 90% and 95% had net coverage probabilities of 46.9% and 65.9%, respectively. For the case of nominal 90%

Table 4: Performance of WASSP (Using $A = 7$), ASAP3, and H&W in the AR(1) Process

Precision		Nominal 90% CIs			Nominal 95% CIs		
Requirement	Performance Measure	WASSP	H&W	ASAP3	WASSP	H&W	ASAP3
None	# replications	1,000	1,000	400	1,000	1,000	400
	Net CI coverage	90.9%	46.9%	95.5%	94.5%	65.9%	98.8%
	Avg. sample size	9,866	1,480	41,076	9,824	1,474	41,076
	Avg. relative prec.	0.053	0.134	0.023	0.067	0.167	0.028
	Avg. CI half-length	5.30	13.4	2.33	6.73	16.7	2.83
	Var. CI half-length	1.83	3.03	0.170	2.88	4.55	0.270
	# replications satisfying prec. reqt.	1,000	1,000	400	1,000	1,000	400
	Coverage of CIs satisfying prec. reqt.	90.9%	46.9%	95.5%	94.5%	65.9%	98.8%
±3.75%	# replications	1,000	1,000	400	1,000	1,000	400
	Net CI coverage	87%	13.7%	95.5%	95%	29.1%	98.8%
	Avg. sample size	13,535	13,281	41,076	21,099	20,176	41,208
	Avg. relative prec.	0.032	0.051	0.023	0.033	0.046	0.028
	Avg. CI half-length	3.21	5.09	2.33	3.28	4.57	2.82
	Var. CI half-length	0.142	1.67	0.170	0.153	1.904	0.257
	# replications satisfying prec. reqt.	1,000	149	400	1,000	310	400
	Coverage of CIs satisfying prec. reqt.	87%	92.0%	95.5%	95%	93.9%	98.8%
±1.875%	# replications	1,000	1,000	400	1,000	1,000	400
	Net CI coverage	93.5%	60.8%	95.5%	97.7%	75.6%	99.3%
	Avg. sample size	57,449	50,152	68,474	90,371	73,249	101,526
	Avg. relative prec.	0.017	0.018	0.018	0.017	0.017	0.018
	Avg. CI half-length	1.65	1.77	1.76	1.66	1.69	1.77
	Var. CI half-length	0.0423	0.104	0.0134	0.0429	0.0784	0.0120
	# replications satisfying prec. reqt.	1,000	697	400	1,000	816	400
	Coverage of CIs satisfying prec. reqt.	93.5%	87.2%	95.5%	97.7%	92.7%	99.3%
±0.9375%	# replications	1,000	1,000	400	1,000	1,000	400
	Net CI coverage	94%	72.5%	94.3%	98%	74.6%	97.3%
	Avg. sample size	229,730	173,700	213,826	333,050	255,180	254,920
	Avg. relative prec.	0.083	0.084	0.090	0.087	0.085	0.090
	Avg. CI half-length	0.830	0.838	0.894	0.867	0.854	0.896
	Var. CI half-length	0.0105	0.0201	0.0026	0.0115	0.0253	0.0021
	# replications satisfying prec. reqt.	1,000	841	400	1,000	817	400
	Coverage of CIs satisfying prec. reqt.	94%	86.2%	94.3%	98%	91.3%	97.3%

CIs with a required precision of ±3.75%, H&W delivered 149 CIs with acceptable precision; since 92% of those CIs actually covered μ_X, the net CI coverage for the H&W procedure in this case was 13.7%. Overall, H&W was judged to have broken down completely in the given AR(1) process.

Table 5 summarizes the absolute bias, variance, and mean-squared-error statistics for WASSP and H&W in the given AR(1) process. In the no-precision case H&W had significant point-estimator bias, and thus the CVM test was not effective in detecting and eliminating that bias. For both WASSP and H&W, the bias of $\overline{\overline{X}}(m, k)$ shown in Table 5 represents a combination of two different effects. First, $\overline{\overline{X}}(m, k)$ is influenced in general by residual initialization bias—after all, there is no unique, well-defined end of the warm-up period for

the AR(1) process. Second, the truncation point (final spacer size) S and the truncated simulation run length $n' = mk$ (as determined by WASSP or H&W) are random variables so $\overline{\overline{X}}(m,k) = \left(\sum_{u=S+1}^{S+n'} X_u\right) \Big/ n'$ is a ratio of two random variables. Thus for the reasons detailed in Section 2.1 of Lada et al. (2004a) and in Section 4.2.1 of Lada (2003), the truncated grand mean $\overline{\overline{X}}(m,k)$ can also exhibit significant ratio-estimator bias due to (a) randomness of the truncation point S and the truncated sample size n', or (b) an insufficiently large value of the truncated sample size n'.

Table 5: Mean Squared Error, Variance, and Absolute Bias of $\overline{\overline{X}}(m,k)$ Based on 1,000 Runs of WASSP and H&W in the AR(1) Process

Precision Requirement	Performance Measure	Nominal 90% CIs		Nominal 95% CIs	
		WASSP	H&W	WASSP	H&W
None	$\widehat{\text{MSE}}[\,\overline{\overline{X}}(m,k)\,]$	8.75	224.	8.06	225.
	$\widehat{\text{Var}}[\,\overline{\overline{X}}(m,k)\,]$	7.76	22.5	7.26	23.3
	$\lvert\widehat{\text{Bias}}[\,\overline{\overline{X}}(m,k)\,]\rvert$	0.996	14.2	0.896	14.2
±3.75%	$\widehat{\text{MSE}}[\,\overline{\overline{X}}(m,k)\,]$	4.37	4.21	2.63	2.55
	$\widehat{\text{Var}}[\,\overline{\overline{X}}(m,k)\,]$	4.15	2.16	2.56	1.91
	$\lvert\widehat{\text{Bias}}[\,\overline{\overline{X}}(m,k)\,]\rvert$	0.471	1.43	0.276	0.800
±1.875%	$\widehat{\text{MSE}}[\,\overline{\overline{X}}(m,k)\,]$	0.870	1.15	0.518	0.747
	$\widehat{\text{Var}}[\,\overline{\overline{X}}(m,k)\,]$	0.861	0.924	0.517	0.663
	$\lvert\widehat{\text{Bias}}[\,\overline{\overline{X}}(m,k)\,]\rvert$	0.0938	0.479	0.0332	0.290
±0.9375%	$\widehat{\text{MSE}}[\,\overline{\overline{X}}(m,k)\,]$	0.186	0.297	0.110	0.179
	$\widehat{\text{Var}}[\,\overline{\overline{X}}(m,k)\,]$	0.186	0.279	0.110	0.171
	$\lvert\widehat{\text{Bias}}[\,\overline{\overline{X}}(m,k)\,]\rvert$	0.020	0.135	0.010	0.0917

For the given AR(1) process, Tables 4 and 5 show the performance of WASSP and ASAP3 was acceptable, but the performance of H&W was unacceptable with respect to point-estimator accuracy and precision as well as net CI coverage, precision, and stability.

3.3. The AR(1)-to-Pareto (ARTOP) Process

The "AR(1)-to-Pareto," or ARTOP process, is defined as follows. Let $\{Z_u : u = 1, 2, \ldots\}$ be a stationary AR(1) process with $N(0,1)$ marginals and lag-one correlation ρ, which can be generated by the relation $Z_u = \rho Z_{u-1} + \delta_u$, where $Z_0 \sim N(0,1)$ and $\{\delta_u : u = 1, 2, \ldots\} \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma_\delta^2\right)$ is a white-noise process with variance $\sigma_\delta^2 = 1 - \rho^2$. If $\{X_u : u =$

$1, 2, \ldots\}$ is an ARTOP process with marginal c.d.f.

$$F_X(x) \equiv \Pr\{X \le x\} = \begin{cases} 1 - (\xi/x)^{\vartheta}, & x \ge \xi, \\ 0, & x < \xi, \end{cases} \tag{10}$$

where $\xi > 0$ is a location parameter and $\vartheta > 0$ is a shape parameter, then the $\{X_u\}$ are generated from the $\{Z_u\}$ as follows. For all real z, let $\Phi(z) = (2\pi)^{-1/2} \int_{-\infty}^{z} \exp\Big(-\frac{1}{2}w^2\Big)\, dw$ denote the c.d.f. of the $N(0,1)$ distribution. Let $\Big\{\mathcal{R}_u = \Phi(Z_u) : u = 1, 2, \ldots\Big\}$ denote a sequence of correlated random numbers that is supplied to the inverse of the Pareto c.d.f. (10) to yield the ARTOP process,

$$X_u = F_X^{-1}(\mathcal{R}_u) = F_X^{-1}\left[\Phi(Z_u)\right] = \xi \Big/ [1 - \Phi(Z_u)]^{1/\vartheta}, \quad u = 1, 2, \ldots . \tag{11}$$

The mean and the variance of the ARTOP process (11) are $\mu_X = \mathrm{E}[X_u] = \vartheta\xi(\vartheta - 1)^{-1}$ (for $\vartheta > 1$) and $\sigma_X^2 = \xi^2\vartheta(\vartheta - 1)^{-2}(\vartheta - 2)^{-1}$ (for $\vartheta > 2$), respectively (Lada 2003).

The parameters of the Pareto distribution (10) were assigned the values $\vartheta = 2.1$ and $\xi = 1$, and the lag-one correlation in the base process $\{Z_u\}$ was assigned the value $\rho = 0.995$. This yields a test process whose marginal distribution has mean, variance, skewness, and kurtosis respectively given by $\mu_X = 1.91$, $\sigma_X^2 = 17.4$, $\mathrm{E}\Big\{[(X_u - \mu_X)/\sigma_X]^3\Big\} = \infty$, and $\mathrm{E}\Big\{[(X_u - \mu_X)/\sigma_X]^4\Big\} = \infty$. The most difficult aspects of this test process are its highly nonnormal marginals and persistent autocorrelation structure. With the initial condition $Z_0 \sim N(0,1)$, the process started in steady-state operation and therefore had no warm-up period.

Table 6 shows the performance of WASSP for the given ARTOP process using the smoothing parameter values $A = 5, 7$, and 9. The results are based on 400 independent replications of nominal 90% CIs. Table 6 reveals that the CI coverage decreased in general as A increased. For nominal 90% CIs with $A = 7$ and $A = 9$, the resulting coverage probabilities were judged to be unacceptable at all three precision levels. While there was significant undercoverage when the value $A = 5$ was used in the small-sample cases, the CI coverage probabilities approached the nominal level as the sample size increased. It is not entirely clear at this point why $A = 5$ produced the best results for this process.

For the ARTOP process (11), Table 7 shows a comparison of the performance of WASSP (using the default $A = 7$), ASAP3, and H&W. For the no-precision and $\pm 15\%$ cases, ASAP3 outperformed WASSP in terms of CI coverage. In these cases, however, ASAP3 required substantially larger sample sizes on average than did WASSP. In the case of nominal 90% CIs with $\pm 7.5\%$ precision, the coverage probability for WASSP was similar to the coverage

Table 6: Performance of WASSP Using Different Values of the Smoothing Parameter A in the ARTOP Process Based on 400 Replications of 90% CIs

Precision Requirement	Performance Measure	Smoothing Parameter $A = 5$	$A = 7$	$A = 9$
None	CI coverage	84.2%	79%	78%
	Avg. sample size	19,880	22,512	22,512
	Avg. relative prec.	0.271	0.235	0.218
	Avg. CI half-length	0.518	0.448	0.416
	Var. CI half-length	0.0774	0.0544	0.0441
±15%	CI coverage	77.3%	71.5%	72.3%
	Avg. sample size	79,095	66,158	54,551
	Avg. relative prec.	0.115	0.117	0.120
	Avg. CI half-length	0.220	0.223	0.230
	Var. CI half-length	0.0020	0.0018	0.0018
±7.5%	CI coverage	89%	85.3%	82.5%
	Avg. sample size	430,430	345,870	272,670
	Avg. relative prec.	0.060	0.061	0.062
	Avg. CI half-length	0.115	0.116	0.118
	Var. CI half-length	0.0005	0.0005	0.0005

Table 7: Performance of WASSP (Using $A = 7$), ASAP3, and H&W in the ARTOP Process

Precision Requirement	Performance Measure	Nominal 90% CIs WASSP	H&W	ASAP3	Nominal 95% CIs WASSP	H&W	ASAP3
None	# replications	400	400	400	400	400	400
	Net CI coverage	79%	67%	85.5%	87%	75.5%	90.8%
	Avg. sample size	22,512	2,982	114,053	19,012	2,555	114,053
	Avg. relative prec.	0.235	0.373	0.091	0.295	0.492	0.109
	Avg. CI half-length	0.448	0.712	0.173	0.564	0.939	0.207
	Var. CI half-length	0.054	0.684	0.00977	0.083	1.888	0.0144
	# replications satisfying prec. reqt.	400	400	400	400	400	400
	Coverage of CIs satisfying prec. reqt.	79%	67%	85.5%	87%	75.5%	90.8%
±15%	# replications	400	400	400	400	400	400
	Net CI coverage	71.5%	70%	85.5%	81%	82.8%	90.8%
	Avg. sample size	66,158	39,781	117,092	95,488	72,093	120,660
	Avg. relative prec.	0.117	0.123	0.087	0.117	0.126	0.101
	Avg. CI half-length	0.223	0.234	0.163	0.223	0.241	0.190
	Var. CI half-length	0.002	0.002	0.00248	0.002	0.019	0.00239
	# replications satisfying prec. reqt.	400	389	400	400	394	400
	Coverage of CIs satisfying prec. reqt.	71.5%	72%	85.5%	81%	84%	90.8%
±7.5%	# replications	400	400	400	400	400	400
	Net CI coverage	85.3%	81%	84%	91.5%	73.5%	90.3%
	Avg. sample size	345,870	208,570	186,517	520,750	348,470	255,512
	Avg. relative prec.	0.061	0.063	0.068	0.063	0.063	0.070
	Avg. CI half-length	0.116	0.120	0.127	0.120	0.120	0.131
	Var. CI half-length	5.0E-4	3.0E-4	2.10E-4	4.0E-4	3.0E-4	1.18E-4
	# replications satisfying prec. reqt.	400	395	400	400	334	400
	Coverage of CIs satisfying prec. reqt.	85.3%	82%	84%	91.5%	88.1%	90.3%

probability for ASAP3; however, in this case WASSP's average required sample size was 104% larger than that of ASAP3. To supplement the results in Table 7, WASSP was also applied to the ARTOP process with a relative-precision requirement of 3.75%. For 400 replications of nominal 90% CIs at the 3.75% level, WASSP delivered a net CI coverage probability of 91.25% and an average relative precision of 3.27%. These additional results suggest that in the given ARTOP process, the CI coverage delivered by WASSP should stabilize very close to the nominal level as the precision requirement approaches zero.

Examination of Table 7 revealed some noteworthy differences between WASSP and H&W, especially in the no-precision and ±7.5% precision cases. For example, in the no-precision case the empirical coverage probabilities for nominal 90% and 95% CIs delivered by H&W were 67% and 75.5%, respectively, while the corresponding figures for WASSP were 79% and 87%, respectively. Furthermore, for the case of nominal 95% CIs with a required precision of ±7.5%, H&W delivered $R = 334$ CIs with acceptable precision; since 88.1% of those CIs actually covered μ_X, the net-coverage probability for H&W was only 73.5% while the corresponding figure for WASSP was 91.5%. For the experiments summarized in Table 7, WASSP generally outperformed H&W and achieved marginally acceptable CI coverage probabilities for the precision requirement of ±7.5%. Because the ARTOP process was started in steady-state, there was no need to examine the mean squared error and absolute bias of $\overline{\overline{X}}(m, k)$ for any of the output-analysis procedures.

Finally as detailed in Section 2.1, H&W requires specification of an upper limit on the allowable length of the test process. In many practical applications, users do not have enough information to set this limit; therefore for the given ARTOP process (as well as for the other test problems discussed in this article) the upper limit for H&W was taken to be the final sample size required by WASSP. Consequently, the H&W results presented in this article represent the scenario in which the user has selected a suitable upper limit on the run length.

4. Conclusions and Recommendations

In the experimental performance evaluation summarized in Section 3, three extraordinarily difficult test processes were used to compare WASSP, the H&W spectral method, and ASAP3 with respect to their efficiency and the robustness of the CIs delivered by these procedures. WASSP outperformed H&W in many respects. The results of Lada and Wilson (2005) together with the results in this article provide some evidence that WASSP represents an

advance in spectral methods for simulation output analysis.

Comparison of WASSP and ASAP3 is less clear-cut, with neither procedure dominating the other in the given experiments. Both WASSP and ASAP3 were designed to deliver point and confidence-interval estimators for the steady-state mean of a simulation output process. WASSP, however, also provides an estimator of the SSVP with reasonably stable behavior as well as an estimator of the entire power spectrum of the delivered set of batch means. This additional information can be useful for validating results generated by WASSP and in planning follow-up experiments. Furthermore, it may be possible to use the wavelet-based estimator of the batch-means power spectrum as part of an adaptive version of WASSP in which an appropriate value of the smoothing parameter A might based on an initial study of the shape of the estimated power spectrum of the batch means. This possibility is the subject of ongoing research.

The experimental results detailed in Section 3 provide substantial evidence of WASSP's ability to deliver approximately valid CIs for the steady-state mean of a simulation-generated process with relative precision levels and nominal coverage probabilities often arising in practical applications. Nevertheless, we will continue our experimental investigation of the efficiency and robustness of WASSP when it is applied to interesting test problems—including processes with long-range dependence as well as queuing-network models with multiple customer classes, probabilistic routing, subnetwork capacity constraints, and workstation utilizations that are commonly encountered in certain application domains.

Acknowledgements

The authors thank Stephen D. Roberts and Charles E. Smith (North Carolina State University); and David Goldsman (Georgia Tech) for many enlightening discussions on this article. This research was partially supported by NSF grant DMI-9900164 and by the American Association of University Women (AAUW) through an AAUW Educational Foundation Engineering Dissertation Fellowship.

References

Anderson, T. W., D. A. Darling. 1952. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Statist.* **23** 193–212.

Heidelberger, P., P. D. Welch. 1981a. A spectral method for confidence interval generation and run length control in simulations. *Comm. ACM* **24** 233–245.

Heidelberger, P., P. D. Welch. 1981b. Adaptive spectral methods for simulation output analysis. *IBM J. Res. Develop.* **25** 860–876.

Heidelberger, P., P. D. Welch. 1983. Simulation run length control in the presence of an initial transient. *Oper. Res.* **31** 1109–1144.

Lada, E. K. 2003. A wavelet-based procedure for steady-state simulation output analysis. Ph.D. Dissertation, Operations Research Program, North Carolina State University, Raleigh, NC. `www.lib.ncsu.edu/theses/available/etd-04032003-141616/unrestricted/etd.pdf` [accessed July 6, 2005].

Lada, E. K., J. R. Wilson. 2005. A wavelet-based spectral procedure for steady-state simulation analysis. *European J. Oper. Res.* to appear.

Lada, E. K., J. R. Wilson, N. M. Steiger. 2003. A wavelet-based spectral method for steady-state simulation analysis. S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, eds. *Proceedings of the 2003 Winter Simulation Conference*. Institute of Electrical and Electronics Engineers, Piscataway, NJ. 422–430. `www.informs-sim.org/wsc03papers/052.pdf` [accessed July 6, 2005].

Lada, E. K., J. R. Wilson, N. M. Steiger, J. A. Joines. 2004a. Performance evaluation of a wavelet-based spectral method for steady-state simulation analysis. R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds. *Proceedings of the 2004 Winter Simulation Conference*. Institute of Electrical and Electronics Engineers, Piscataway, NJ. 694–702. `www.informs-sim.org/wsc04papers/084.pdf` [accessed July 6, 2005].

Lada, E. K., J. R. Wilson, N. M. Steiger, J. A. Joines. 2004b. User's manual for WASSP version 1 [online]. Department of Industrial Engineering, North Carolina State University, Raleigh, NC. `ftp.ncsu.edu/pub/eos/pub/jwilson/wasspman.pdf`. [accessed July 6, 2005].

Pawlikowski, K. 1990. Steady-state simulation of queueing processes: a survey of problems and solutions. *ACM Comput. Surveys* **22** 123–170.

Shapiro, S. S., M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* **52** 591–611.

Steiger, N. M., E. K. Lada, J. R. Wilson, C. Alexopoulos, D. Goldsman, F. Zouaoui. 2002.

ASAP2: An improved batch means procedure for simulation output analysis. E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, eds. *Proceedings of the 2002 Winter Simulation Conference*, Institute of Electrical and Electronics Engineers, Piscataway, NJ. 336–344. `www.informs-sim.org/wsc02papers/043.pdf` [accessed July 6, 2005].

Steiger, N. M., E. K. Lada, J. R. Wilson, J. A. Joines, C. Alexopoulos, D. Goldsman. 2005. ASAP3: a batch means procedure for steady-state simulation analysis. *ACM Trans. Model. Comput. Simulation* **15** 39-73.

Steiger, N. M., J. R. Wilson. 2002. An improved batch means procedure for simulation output analysis. *Management Sci.* **48** 1569–1586.

von Neumann, J. 1941. Distribution of the ratio of the mean square successive difference to the variance. *Ann. Math. Statist.* **12** 367–395.