Proceedings of the 2006 Winter Simulation Conference
L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds.

PERFORMANCE EVALUATION OF SPECTRAL PROCEDURES FOR SIMULATION ANALYSIS

Emily K. Lada

SAS Institute Inc.
100 SAS Campus Drive, R5413
Cary, NC 27513-8617, U.S.A.

James R. Wilson

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

ABSTRACT

We summarize an experimental performance evaluation of WASSP and the Heidelberger-Welch (HW) algorithm, two sequential spectral procedures for steady-state simulation analysis. Both procedures approximate the log-smoothed-periodogram of the batch means after suitable data-truncation to eliminate the effects of initialization bias, finally delivering a confidence-interval estimator for the mean response that satisfies user-specified half-length and coverage-probability requirements. HW uses a Cramér–von Mises test for initialization bias based on the method of standardized time series; and then HW fits a quadratic polynomial to the batch-means log-spectrum. In contrast WASSP uses the von Neumann randomness test and the Shapiro-Wilk normality test to obtain an approximately stationary Gaussian batch-means process whose log-spectrum is approximated via wavelets. Moreover, unlike HW, WASSP estimates the final sample size required to satisfy the user's confidence-interval requirements. Regarding closeness of conformance to both confidence-interval requirements, we found that WASSP outperformed HW in the given test problems.

1 INTRODUCTION

In discrete-event simulation, we are often interested in estimating the steady-state mean μ_X of a stochastic output process $\{X_u : u = 1, 2, \ldots\}$ generated by a single, prolonged run of the associated simulation model. Assuming the target process is stationary and given a time series of length n that is part of a single realization of this process, we see that a natural point estimator of μ_X is the sample mean, $\overline{X}(n) = n^{-1}\sum_{u=1}^{n} X_u$. We also require some indication of the precision of this point estimator; and typically we construct a confidence interval (CI) for μ_X with a user-specified probability $1-\alpha$ of covering the point μ_X, where $0 < \alpha < 1$. The CI for μ_X should satisfy two criteria: (i) it is approximately valid—that is, its coverage probability is sufficiently close to the nominal level $1-\alpha$; and (ii) it has acceptable precision—that is, it is narrow enough to be meaningful in the context of the application at hand without being excessively narrow.

In this article we focus on spectral procedures for constructing such CIs. If $\{X_u : u = 1, \ldots, n\}$ is covariance stationary, 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 steady-state variance parameter (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. The power spectrum $p_X(\omega)$ of this process is given by

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

At the 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

$$I\left(\tfrac{\ell}{n}\right) = \tag{3}$$
$$n^{-1}\left(\left\{\sum_{u=1}^{n} X_u \cos\left[\tfrac{2\pi(u-1)\ell}{n}\right]\right\}^2 + \left\{\sum_{u=1}^{n} X_u \sin\left[\tfrac{2\pi(u-1)\ell}{n}\right]\right\}^2\right)$$

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

A spectral procedure for simulation analysis delivers an estimator $\widehat{\gamma}_X$ of γ_X based on (3), from which we compute a $100(1-\alpha)\%$ CI estimator of μ_X having the form

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

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 $\overline{X}(n')$ 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*) $t_{1-\alpha/2,\nu}$ denotes the $1-\alpha/2$ quantile of Student's t-distribution with ν d.f., provided $0 < \alpha < 1$.

In this article we examine the performance of two spectral procedures for steady-state simulation output analysis: the Heidelberger-Welch (HW) procedure (Heidelberger and Welch 1981a, 1981b, 1983) and WASSP (Lada 2003; Lada, Wilson, and Steiger 2003; Lada and Wilson 2006; Lada et al. 2004a, 2004b, 2005). Lada et al. (2004b, 2005) summarize the performance of HW and WASSP when those procedures are applied to problems constituting a kind of "torture test" designed to elicit worst-case behavior. By contrast, in this article we report the performance of HW and WASSP on selected test problems whose probabilistic behavior is more nearly typical of a broad class of steady-state simulation applications.

This article is organized as follows. The WASSP and HW procedures are briefly summarized in Sections 2 and 3, respectively. The results of the experimental performance evaluation are detailed in Section 4. Finally Section 5 recapitulates the main conclusions and recommendations stemming from this work. The slides for the oral presentation of this article are available online via <ftp.ncsu.edu/pub/eos/pub/jwilson/wsc06lada.pdf>.

2 OVERVIEW OF WASSP

The first part of the WASSP algorithm seeks to determine sufficiently large values for the size of each batch and for the size of the *spacer* preceding each batch so as to ensure that the corresponding spaced batch means are approximately independent and identically distributed (i.i.d.). WASSP begins by dividing the initial simulation-generated output process of length $n = 4{,}096$ observations into a set of $k = 256$ adjacent batches each of size $m = 16$ so that the initial spacer size is zero. On each iteration of the first part of WASSP, the randomness test of von Neumann (1941) is used to test the hypothesis that the current spaced batch means constitute a random sample—that is, the spaced batch means are i.i.d. Each time the randomness test is failed, one more batch is added to each spacer (up to a limit of nine batches per spacer); and then the randomness test is reperformed on the new (reduced) set of spaced batch means. If the randomness test is failed with spacers each consisting of nine batches so that the batch count has been reduced to $k = 25$, then the batch size m is increased by the factor $\sqrt{2}$ and the process of testing the batch means for randomness is restarted by computing adjacent batch means of the new batch size (so that the spacer size is reset to zero and the batch count is reset to $k = 256$).

Once the randomness test is passed, the spacer preceding the first batch is assumed to contain the warm-up period and hence is taken to define an appropriate truncation point; and then the second part of the WASSP algorithm seeks to determine a batch size that is sufficiently large to ensure the spaced batch means are approximately normal. For this purpose WASSP uses the univariate normality test of Shapiro and Wilk (1965) to test the composite hypothesis that the current spaced batch means have a common normal distribution whose mean and variance are unspecified.

Once the normality test is passed, the adjacent (nonspaced) batch means of the current batch size computed beyond the truncation point are assumed to constitute an approximately stationary Gaussian (normal) process. In the third and final part of WASSP, a wavelet-based estimator of the corresponding batch-means log-spectrum is computed over its full frequency range (that is, from $-\frac{1}{2}$ to $\frac{1}{2}$ cycles per unit of time) as follows. WASSP smooths the periodogram (3) of the batch means by computing a multipoint moving average (consisting of seven points by default); then WASSP applies a logarithmic transformation to the smoothed periodogram and corrects for the bias induced by this transformation. Next, WASSP computes the discrete wavelet transform of the bias-corrected log-smoothed-periodogram of the batch means, where a soft-thresholding scheme is used to obtain a parsimonious, denoised set of wavelet coefficient estimators. Finally, WASSP computes the inverse discrete wavelet transform of the thresholded wavelet coefficients to recover the wavelet-based approximation to the batch-means log-spectrum.

The third part of WASSP also yields a CI estimator of μ_X that satisfies the user-specified precision requirement and approximately achieves the user-specified coverage probability. From the wavelet-based estimator of the log-spectrum of the truncated batch means, WASSP computes an estimator of the SSVP defined in (1); and then WASSP computes a CI of the form (4), where the midpoint of the CI is the grand average of all the adjacent (nonspaced) batch means computed beyond the truncation point. If the CI fails to satisfy the user-specified precision requirement, then WASSP performs the following operations: (i) estimation of the total sample size required to satisfy the precision requirement; (ii) computation of a new set of adjacent (nonspaced) batch means after obtaining additional data if necessary and skipping the observations in the initial spacer; (iii) computation of a new estimate of the SSVP from the wavelet-based approximation to the log-smoothed-periodogram for the latest set of adjacent (nonspaced) batch means; (iv) construction of the CI (4) from the latest set of adjacent (nonspaced) batch means; and (v) evaluation of the precision requirement. This cycle of estimating the SSVP, computing the CI (4), and testing the precision requirement—that is, the third

part of WASSP—is performed iteratively until the precision requirement is finally satisfied. In the final CI (4) delivered by WASSP, we have $\nu = 6$ d.f. by default.

WASSP requires the following user-supplied inputs:

- a simulation-generated output process $\{X_u : u = 1, \ldots, n\}$ from which the steady-state expected response μ_X is to be estimated;
- the desired CI coverage probability $1 - \alpha$, where $0 < \alpha < 1$; and
- an absolute or relative precision requirement specifying the final CI half-length in terms of (i) a maximum acceptable half-length H^* (for an absolute precision requirement); or (ii) a maximum acceptable fraction r^* of the magnitude of the CI midpoint (for a relative precision requirement).

WASSP delivers the following outputs: (i) a nominal $100(1-\alpha)\%$ CI for μ_X that satisfies the specified absolute or relative precision requirement, provided no additional simulation-generated observations are required; or (ii) a larger value of n, the size of the total sample to be supplied to WASSP when it is executed again. Thus WASSP finally terminates when the midpoint $\overline{X}(n')$ and the half-length H of the latest CI of the form (4) satisfy the stopping rule

$$H \leq \begin{cases} H^*, & \text{for an absolute precision reqt.,} \\ r^* \left| \overline{X}(n') \right|, & \text{for a relative precision reqt.,} \\ \infty, & \text{for no precision requirement.} \end{cases} \quad (5)$$

A formal algorithmic statement of WASSP is given in Lada and Wilson (2006). A stand-alone Windows-based version of WASSP and a user's manual are available online via Lada et al. (2004a).

3 OVERVIEW OF THE HEIDELBERGER-WELCH (HW) PROCEDURE

Heidelberger and Welch (1981a, 1981b, 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 HW procedure is complicated because the latter procedure requires the user to specify an upper limit $t_{\max}$ on the allowable length of a given test process to which HW is to be applied. (To avoid confusion in this section and throughout the rest of the article, the notation of Heidelberger and Welch (1981a, 1981b, 1983) is always used when referring to the HW procedure.) For a fair comparison of WASSP with the HW 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 $\{X_u : u = 1, \ldots, n\}$ to which the (partially) sequential version of the HW procedure can be applied after taking $t_{\max} = n$, the length of the simulation-generated time series, for the current replication of the HW procedure.

Heidelberger and Welch (1983) describe a scheme for batching data prior to applying their spectral method, and this scheme is used in our implementation of the HW procedure. The batch count k for the HW procedure is always in the range $L \leq k \leq 2L$, where the value $L = 200$ is used to conform to the recommendations of Heidelberger and Welch (1983). Within each replication of the HW procedure, 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 $t_1 = \lceil 0.15\, t_{\max} \rceil$ and $t_i = \min\{\lceil 1.5\, t_{i-1} \rceil, t_{\max}\}$ for $i = 2, 3, \ldots$. If $t_i \geq L$ and the assignment $b_i = \lfloor \log_2\{(t_i - 1)/L\} \rfloor$ 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 $k_i = \lfloor t_i/m_i \rfloor$, respectively.

The version of the HW procedure examined in this article uses the method for detecting and eliminating initialization bias described in Heidelberger and Welch (1983). At the ith checkpoint (for $i = 1, 2, \ldots$), the HW procedure tests the null hypothesis that the untruncated batch-means process (currently consisting of k_i batch means with batch size m_i) is covariance stationary by computing the corresponding Cramér–von Mises (CVM) test statistic, $\text{CVM}(m_i, k_i)$.

If the current untruncated batch-means process is covariance stationary, then under widely applicable conditions as $m_i \to \infty$ and $k_i \to \infty$, the asymptotic distribution of $\text{CVM}(m_i, k_i)$ is equal to the distribution of $\text{CVM}(\mathcal{B}) = \int_0^1 \mathcal{B}^2(u)\, du$, where $\{\mathcal{B}(u) : u \in [0, 1]\}$ is a Brownian bridge process. Thus if the current batch-means process with batch size m_i and batch count k_i is covariance stationary, then the asymptotic 0.9 quantile of the CVM test statistic is $\text{CVM}(\mathcal{B})_{0.9} = 0.3473$; see Table 1 of Anderson and Darling (1952). If $\text{CVM}(m_i, k_i) > 0.3473$, then the CVM test has detected nonstationarity (initialization bias) in the untruncated sequence of batch means so the HW procedure 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, the HW procedure 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 consecutive times at the ith checkpoint so that the CVM test detects nonstationarity even after deleting the first half of the batch-means sequence, then the HW procedure 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 the precision requirement (5) is computed from the truncated batch means; or (*b*) the untruncated sample size required by the HW procedure 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 the HW procedure terminates without delivering a CI; otherwise the HW procedure terminates after delivering a CI of the form (4) that might not satisfy (5).

In conformance with the recommendations of Heidelberger and Welch (1981a, 1981b, 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 HW-generated CI of the form (4), the quantity ν denoting the effective degrees of freedom is given by $\nu = 7$ d.f.

4 EXPERIMENTAL PERFORMANCE EVALUATION

In Lada et al. (2004b, 2005), the suite of test problems is deliberately selected to provide relatively extreme examples of nonnormal, correlated simulation output processes that in most cases are contaminated by initialization bias—namely, (*a*) the $M/M/1$ queue waiting time process with a steady-state server utilization of 0.9 and an empty-and-idle initial condition; (*b*) the AR(1) process with autoregressive parameter value of 0.995, steady-state mean of 100, steady-state variance of 100.25, and initial condition of zero; and (*c*) the Autoregressive-to-Pareto process obtained by transformation of a stationary version of the AR(1) process above such that the transformed process has a Pareto marginal distribution with finite mean and variance but with infinite skewness and kurtosis.

By contrast Lada, Steiger, and Wilson (2006) identify several test problems that are more nearly typical of practical applications but are particularly difficult for many steady-state simulation analysis procedures:

1. the response (cycle) times in the central server model 3 of Law and Carson (1979);
2. queue waiting times in the $M/M/1$/LIFO queue with server utilization of 0.8;
3. queue waiting times in the $M/H_2/1$ queueing system with hyperexponential service times having a coefficient of variation of 2.0 and a server utilization of 0.8;
4. queue waiting times for the $M/M/1$ queue with server utilization 0.8;
5. total time spent waiting in a queue for each customer passing through the $M/M/1/M/1$ tandem queueing system with server utilization 0.8 at each work station; and
6. the reward process associated with a two-state discrete-time Markov chain with high positive correlation.

The stochastic behaviors exhibited by test problems 1–6 are typical of many steady-state simulation applications and will enable us to make a direct comparison of the performance of HW and WASSP.

WASSP and HW were applied to $G = 400$ independent replications of each test problem so as to deliver up to 400 independent instances of CIs with nominal coverage probabilities of 90% and 95% and with several nominal levels of relative precision. We consider the generation of CIs for the steady-state mean response in a specific configuration of a given test problem that is defined by particular values of the nominal CI coverage and relative precision—for example, nominal 90% CIs for the mean queue waiting time in the $M/M/1$ queue with nominal relative precision of $\pm 7.5\%$. On each replication of the given test-problem configuration, WASSP was executed to deliver another CI and to determine the overall sample to be supplied to the corresponding run of the HW procedure. This approach was necessary because HW has no facility for determining its own sample size required to satisfy given CI coverage and relative-precision requirements. An advantage of this approach to performing paired runs of WASSP and HW is that it sharpened our performance comparison because both procedures were applied to exactly the same data sets for each test-problem configuration reported in this paper.

From our previous computational experience as detailed in Lada et al. (2005), we found that the coverage probabilities delivered by HW could be significantly affected by the premature termination of the algorithm on those replications for which there was insufficient data to generate a CI satisfying the precision requirement. Thus to characterize fully the performance of both WASSP and HW in terms of conformance to the user's requirements on relative precision and coverage probability, we computed the following estimated probabilities for each procedure applied to each test-problem configuration:

- the *net CI coverage*, defined as the ratio Q/G, where Q denotes the number of CIs that simultaneously covered the steady-state mean and satisfied the precision requirement (5); and
- the *satisfied coverage*, defined as the ratio Q/R, where R denotes the number of CIs that satisfied the relative precision requirement (5).

Thus the "net CI coverage" for a particular procedure applied to a particular test-problem configuration estimates the probability that the procedure will deliver a CI covering μ_X and satisfying the user's relative precision requirement. By contrast, the "satisfied coverage" for a particular procedure and test-problem configuration estimates the conditional coverage probability of the CI delivered by the procedure, given that the delivered CI is sufficiently narrow to satisfy the relevant precision requirement. For WASSP these two performance measures always coincide by design; for HW, however, these statistics may differ significantly. Because $G = 400$, the coverage estimators for the CIs delivered by WASSP have standard errors of approximately 1.5% and 1% for nominal coverage levels of 90% and 95%, respectively; but we cannot make a similar statement about the standard errors of the empirical coverage probabilities delivered by HW.

To complete our comparison of the performance of WASSP and HW, for each test-problem configuration we computed estimates of the bias, variance, and mean squared error of the final point estimator $\overline{X}(n')$ delivered by each procedure, when we restricted consideration to those CIs satisfying the relevant precision requirement. For the uth replication of a given procedure on a particular test-problem configuration that delivered a CI satisfying the precision requirement, let $\overline{X}_u(n'_u)$ denote the resulting grand average of the truncated output process consisting of n'_u observations, where $u = 1, \ldots, R$. The bias of $\overline{X}(n')$ is estimated by $\widehat{\text{Bias}}[\,\overline{X}(n')\,] = [R^{-1}\sum_{u=1}^{R}\overline{X}_u(n'_u)] - \mu_X$; and the corresponding mean squared error is estimated by $\widehat{\text{MSE}}[\,\overline{X}(n')\,] = R^{-1}\sum_{u=1}^{R}[\overline{X}_u(n'_u) - \mu_X]^2$. The variance of $\overline{X}(n')$ is estimated by $\widehat{\text{Var}}[\,\overline{X}(n')\,]$, the (unbiased) sample variance of the truncated batch means $\{\overline{X}_u(n'_u) : u = 1, \ldots, R\}$.

Lada et al. (2005) find that the Crámer–von Mises test used by HW is not effective in detecting and eliminating initialization bias in test problems for which the initial transient period is exceptionally long or pronounced. In this article, estimation of the bias, variance, and mean squared error of $\overline{X}(n')$ in the selected test-problem configurations allowed us to assess the effectiveness of the methods used by WASSP and HW to eliminate initialization bias when those methods are applied to test problems with less extreme transients.

4.1 Central Server Model 3 of Law and Carson (1979)

Central server model 3 of Law and Carson (1979) is one of four computer-system models used by Lada, Steiger, and Wilson (2006) to compare the performance of WASSP with that of several batch-means procedures. This model consists of a CPU (the central server) and two peripheral units so that there are $M = 3$ service centers in this system. The system has $N = 8$ jobs in it at all times. When a job is finished at the CPU, it is routed to a peripheral unit. Specifically, it is sent to peripheral unit 2 with probability $p_2 = 0.9$ or to peripheral unit 3 with probability $p_3 = 0.10$. After getting service from the ℓth peripheral unit at rate μ_ℓ (so that the corresponding service time is exponentially distributed with mean $1/\mu_\ell$), the job leaves the system and is immediately replaced by a job joining the CPU queue. Law and Carson (1979) use the service rates $\mu_1 = 1.0$, $\mu_2 = 0.45$, and $\mu_3 = 0.05$ in this system. The process of interest is the response (cycle, sojourn) time X_i of the ith job for $i = 1, 2, \ldots$, where the job's response (cycle, sojourn) time is the delay between its arrival at the CPU queue and its departure from the system. In this system, the steady-state mean response time is $\mu_X = 18.279$. The system's initial condition consisted of 5, 1, and 2 jobs at service centers 1, 2, and 3, respectively. Table 1 summarizes the results obtained for this test problem. In all the following tables, the label "None" for the precision requirement identifies results obtained using the last option in the stopping rule (5) so that in effect each analysis procedure delivered a CI without checking on a precision requirement to be satisfied by that CI.

Table 1 shows that at all levels of precision, the CIs delivered by WASSP exhibited net coverages slightly above the corresponding nominal levels. By contrast the CIs delivered by HW exhibited net coverages slightly below the corresponding nominal levels for all reported relative precision levels except ±1%, where significant losses in net CI coverage were observed. In particular, the nominal 90% and 95% CIs with relative precision of ±1% delivered by HW had net CI coverages of 79% and 84.75%, respectively; and in each case, only 370 replications of HW delivered a CI satisfying the precision requirement. On each of the other 30 paired replications of WASSP and HW that were performed to obtain nominal 90% and 95% CIs with relative precision of ±1%, HW finally delivered a CI; but to satisfy the precision requirement HW required a larger sample than that provided by the matching run of WASSP. Although the "satisfied coverage" probabilities for nominal 90% and 95% CIs delivered by HW were 85.43% and 91.75%, respectively, we concluded that the resulting net CI coverages delivered by HW were unacceptably low.

Further examination of Table 1 revealed that at all levels of precision, the estimated absolute bias, variance, and mean squared error of $\overline{X}(n')$ were larger for HW than the corresponding statistics for WASSP. However, once a precision requirement was imposed and the sample size increased, we observed substantial reductions in the bias, variance, and mean squared error of $\overline{X}(n')$ for both procedures.

4.2 $M/M/1$/LIFO Queue

Table 2 summarizes our results for the queue waiting time process in the $M/M/1$/LIFO queueing system with mean

Table 1: Performance of Spectral Procedures for the Response Time Process in Central Server Model 3 of Law and Carson (1979) Based on 400 Independent Replications of Nominal 90% and 95% CIs

Prec. Reqt.	Performance Measure	90% CIs WASSP	90% CIs HW	95% CIs WASSP	95% CIs HW
None	Net CI coverage	93.0%	87.5%	96.5%	94.25%
	Avg. sample size	79,075	11,862	79,075	11,862
	Avg. CI half-length	0.5220	1.084	0.6580	1.3529
	Var. CI half-length	0.1880	0.5178	0.2990	0.8067
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.1119	0.5615	0.1119	0.5615
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.1127	0.5607	0.1127	0.5607
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0115	0.0630	0.0115	0.0630
	# reps. satisfying	400	400	400	400
	Satisfied coverage	93.0%	87.5%	96.5%	94.25%
±15%	Net CI coverage	93.0%	87.75%	96.5%	94.25%
	Avg. sample size	79,066	11,904	79,056	11,971
	Avg. CI half-length	0.5210	1.0398	0.6460	1.2335
	Var. CI half-length	0.1830	0.3478	0.2450	0.3797
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.1133	0.4563	0.1025	0.4466
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.1141	0.4546	0.1033	0.4419
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0101	0.0671	0.0077	0.0833
	# reps. satisfying	400	400	400	400
	Satisfied coverage	93.0%	87.75%	96.5%	94.25%
±7.5%	Net CI coverage	93.0%	88.0%	96.5%	93.0%
	Avg. sample size	79,188	12,562	79,355	13,539
	Avg. CI half-length	0.4770	0.8475	0.5780	0.9346
	Var. CI half-length	0.0840	0.0972	0.1000	0.0837
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.0806	0.3130	0.0793	0.2573
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.0812	0.3089	0.0798	0.2542
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0148	0.0752	0.0145	0.0653
	# reps. satisfying	400	400	400	398
	Satisfied coverage	93.0%	88.0%	96.5%	93.47%
±1%	Net CI coverage	92.5%	79.0%	97.0%	84.75%
	Avg. sample size	272,670	178,805	430,818	282,368
	Avg. CI half-length	0.1600	0.2058	0.1610	0.2057
	Var. CI half-length	0.0004	0.0389	0.0004	0.0339
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.0080	0.0251	0.0053	0.0214
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.0079	0.0252	0.0052	0.0213
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0084	0.0117	0.0094	0.0147
	# reps. satisfying	400	370	400	370
	Satisfied coverage	92.5%	85.43%	97.0%	91.85%

Table 2: Performance of Spectral Procedures for the $M/M/1$/LIFO Queue Waiting Time Process Based on 400 Independent Replications of Nominal 90% and 95% CIs

Prec. Reqt.	Performance Measure	90% CIs WASSP	90% CIs HW	95% CIs WASSP	95% CIs HW
None	Net CI coverage	93.0%	89.8%	96.0%	94.0%
	Avg. sample size	125,517	19,916	124,202	19,916
	Avg. CI half-length	0.2650	0.5936	0.3350	0.7408
	Var. CI half-length	0.0230	0.1379	0.0310	0.2148
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.0198	0.1076	0.0198	0.1076
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.0195	0.1149	0.0196	0.1080
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0171	0.0138	0.0172	0.0205
	# reps. satisfying	400	400	400	400
	Satisfied coverage	93.0%	89.8%	96.0%	94.0%
±15%	Net CI coverage	90.7%	85.46%	95.2%	88.5%
	Avg. sample size	124,512	26,847	126,682	35,524
	Avg. CI half-length	0.2490	0.3713	0.2960	0.3847
	Var. CI half-length	0.0110	0.0061	0.0110	0.0043
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.0190	0.0563	0.0166	0.0421
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.0185	0.0552	0.0164	0.0402
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0166	0.0366	0.0132	0.0456
	# reps. satisfying	400	400	400	394
	Satisfied coverage	90.7%	85.46%	95.2%	90.05%
±7.5%	Net CI coverage	90.2%	79.5%	96.2%	85.25%
	Avg. sample size	152,355	80,098	194,590	120,135
	Avg. CI half-length	0.1860	0.2088	0.1990	0.2128
	Var. CI half-length	0.0020	0.0018	0.0010	0.0024
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.0113	0.0240	0.0091	0.0164
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.0112	0.0233	0.0082	0.0156
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0104	0.0281	0.0109	0.0260
	# reps. satisfying	400	386	400	375
	Satisfied coverage	90.2%	82.59%	96.2%	91.18%
±3.75%	Net CI coverage	89.0%	78.25%	94.0%	84.0%
	Avg. sample size	444,190	306,055	695,017	473,765
	Avg. CI half-length	0.1030	0.1170	0.1030	0.1160
	Var. CI half-length	0.0002	0.0033	0.0002	0.0034
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.0037	0.0071	0.0021	0.0050
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.0036	0.0070	0.0021	0.0047
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0093	0.0106	0.0085	0.0164
	# reps. satisfying	400	369	400	375
	Satisfied coverage	89.0%	84.97%	94.0%	89.78%

interarrival time of 1.0, mean service time of 0.8, and an empty-and-idle initial condition. Thus in steady-state operation this system has server utilization $\tau = 0.8$ and mean queue waiting time $\mu_X = 3.20$.

Based on the bias, variance, and mean squared error results, we concluded that both WASSP and HW effectively detected and eliminated the initialization bias for this system. In all configurations of this test problem for which a nonvacuous precision requirement was specified, WASSP exhibited substantially better conformance to the nominal CI coverage probabilities than did HW. For example, at the ±7.5% precision level, the net CI coverages for WASSP were 90.2% for nominal 90% CIs and 96.2% for nominal 95% CIs, while the corresponding net coverages for HW were 79.5% and 85.25%, respectively. At the precision levels of ±7.5% and ±3.75%, we concluded that HW delivered CIs with unacceptably low net coverage probabilities. One cause of HW's poor performance in these cases was the inability of HW consistently to deliver a CI satisfying the precision requirement using the sample sizes provided by WASSP. Another noteworthy cause of HW's poor performance in this test problem was the unacceptably low coverage probabilities of the CIs satisfying the precision requirement. For example in the case of nominal 90% CIs with relative precision ±3.75%, only 369 replications of HW delivered CIs satisfying the precision requirement;

moreover the "satisfied coverage" for the latter CIs was only 84.97%

4.3 $M/H_2/1$ Queue

Table 3 summarizes our results for the $M/H_2/1$ queueing system with an empty-and-idle initial condition, a mean interarrival time of 1.0, and a hyperexponential service-time distribution that is a mixture of two exponential distributions such that the service times have mean 0.8 and coefficient of variation 2.0. (See Lada, Steiger, and Wilson (2006) further discussion of this service-time distribution.) Thus in steady-state operation this system has server utilization $\tau = 0.8$ and mean queue waiting time $\mu_X = 8.0$.

From Table 3 we concluded that at all reported levels of precision, WASSP outperformed HW in terms of conformance to the nominal CI coverage probabilities. For the case of nominal 90% CIs with a required precision of $\pm 3.75\%$, HW delivered $R = 349$ CIs with acceptable precision; and because 85.1% of those CIs actually covered the true steady-state mean, the net coverage probability for HW was only 74% while the corresponding figure for WASSP was 93%. Furthermore, in the no precision case HW had significant point-estimator bias, indicating that the CVM test was not effective in detecting and eliminating that bias.

Table 3: Performance of Spectral Procedures for the $M/H_2/1$ Queue Waiting Time Process Based on 400 Independent Replications of Nominal 90% and 95% CIs

Prec. Reqt.	Performance Measure	90% CIs WASSP	90% CIs HW	95% CIs WASSP	95% CIs HW
None	Net CI coverage	91.0%	76.25%	93.0%	81.5%
	Avg. sample size	23,221	3,663	22,230	3,663
	Avg. CI half-length	2.7040	4.2225	3.4560	5.2700
	Var. CI half-length	1.7720	5.9377	2.9820	9.2492
	$\widehat{\text{MSE}}[\overline{X}(n')]$	1.6838	8.3086	1.6838	8.3086
	$\widehat{\text{Var}}[\overline{X}(n')]$	1.6729	8.1947	1.6729	8.1947
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.1228	0.3666	0.1228	0.3666
	# reps. satisfying	400	400	400	400
	Satisfied coverage	91.0%	76.25%	93.0%	81.5%
$\pm 15\%$	Net CI coverage	88.3%	82.25%	94.5%	85.5%
	Avg. sample size	78,691	57,739	138,960	90,973
	Avg. CI half-length	0.9930	0.9893	0.9940	1.0170
	Var. CI half-length	0.0300	0.0258	0.0290	0.0204
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.4091	0.5033	0.1961	0.3120
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.3838	0.4789	0.1902	0.2959
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.1619	0.1601	0.0798	0.1300
	# reps. satisfying	400	392	400	380
	Satisfied coverage	88.3%	84.18%	94.5%	90.26%
$\pm 7.5\%$	Net CI coverage	91.0%	78.0%	95.7%	85.0%
	Avg. sample size	330,580	233,430	519,990	343,980
	Avg. CI half-length	0.5160	0.5095	0.5280	0.5153
	Var. CI half-length	0.0060	0.0046	0.0050	0.0038
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.0828	0.1251	0.0537	0.0772
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.0821	0.1231	0.0528	0.0754
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0296	0.0482	0.0315	0.0445
	# reps. satisfying	400	377	400	363
	Satisfied coverage	91.0%	83.02%	95.7%	93.75%
$\pm 3.75\%$	Net CI coverage	93.0%	74.0%	98.0%	81.25%
	Avg. sample size	1,283,400	882,880	2,006,800	1,429,800
	Avg. CI half-length	0.2700	0.2586	0.2700	0.2556
	Var. CI half-length	0.0009	0.0009	0.0009	0.0010
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.0195	0.0340	0.0128	0.0215
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.0195	0.0336	0.0127	0.0211
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0078	0.0227	0.0131	0.0201
	# reps. satisfying	400	349	400	357
	Satisfied coverage	93.0%	85.1%	98.0%	91.04%

4.4 $M/M/1$ Queue

Table 4 summarizes our results for the $M/M/1$ queueing system with empty-and-idle initial condition, mean interarrival time 1.0, and mean service time 0.8. Thus in steady-state operation, this system has server utilization $\tau = 0.8$ and mean waiting time in the queue $\mu_X = 3.2$.

Table 4 shows that in the no precision case, the sample sizes for HW were significantly smaller than the sample sizes for WASSP; thus for the HW procedure, the estimated variance and mean squared error of $\overline{X}(n')$ were much larger than the corresponding figures for WASSP. Once a precision requirement was imposed and the sample size began to increase, the bias, variance, and mean squared error of $\overline{X}(n')$ began to decrease for both procedures.

Table 4 also shows that in the cases where a precision requirement was specified, WASSP outperformed HW in terms of conformance to the nominal coverage probability levels. For example, in the case of nominal 95% CIs at the $\pm 3.75\%$ precision level, HW delivered only $R = 375$ CIs satisfying the precision requirement so that HW's net coverage probability was 78.25%. The net coverage probability for WASSP for the same case was 94%.

4.5 $M/M/1/M/1$ Queue

Table 5 summarizes our results for the queue waiting time process in the $M/M/1/M/1$ queueing system—that is, the system with two $M/M/1$ queues in series. This system has mean interarrival time of 1.0, mean service time of 0.8 at each server, and an empty-and-idle initial condition. Thus in steady-state operation, each server has utilization $\tau = 0.8$, and expected total waiting time in both queues is $\mu_X = 6.4$.

We concluded from Table 5 that in the $\pm 15\%$ and $\pm 7.5\%$ precision cases, WASSP significantly outperformed HW; in particular, the net coverage probabilities delivered by HW were significantly below the corresponding nominal levels. However, at the $\pm 15\%$ and $\pm 7.5\%$ precision levels, both methods appeared to effectively detect and eliminate the initialization bias.

4.6 Two-State Discrete-Time Markov Chain

The final test process is a real-valued "reward" function defined on an irreducible aperiodic discrete-time Markov chain (DTMC) with a high positive correlation structure.

Table 4: Performance of Spectral Procedures for the $M/M/1$ Queue Waiting Time Process Based on 400 Independent Replications of Nominal 90% and 95% CIs

Prec. Reqt.	Performance Measure	90% CIs WASSP	90% CIs HW	95% CIs WASSP	95% CIs HW
None	Net CI coverage	91.3%	90.5%	96.5%	90.75%
	Avg. sample size	38,275	6,197	35,074	5,631
	Avg. CI half-length	0.5400	1.1380	0.7060	1.5067
	Var. CI half-length	0.1110	0.6464	0.1870	1.2423
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.0908	0.3486	0.0928	0.4543
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.0880	0.3486	0.0904	0.4514
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0551	0.0290	0.0519	0.0632
	# reps. satisfying	400	400	400	400
	Satisfied coverage	91.3%	90.5%	96.5%	90.75%
±15%	Net CI coverage	89.0%	80.0%	96.3%	88.75%
	Avg. sample size	42,497	19,173	54,578	30,737
	Avg. CI half-length	0.3720	0.3970	0.3840	0.4016
	Var. CI half-length	0.0070	0.0041	0.0060	0.0035
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.0450	0.0817	0.0321	0.0476
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.0448	0.0793	0.0320	0.0462
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0168	0.0513	0.0123	0.0384
	# reps. satisfying	400	387	400	383
	Satisfied coverage	89.0%	82.69%	96.3%	92.95%
±7.5%	Net CI coverage	88.5%	79.0%	96.3%	86.75%
	Avg. sample size	117,540	77,971	179,840	123,370
	Avg. CI half-length	0.2000	0.2034	0.2020	0.2035
	Var. CI half-length	0.0010	0.0006	0.0008	0.0007
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.0166	0.0242	0.0091	0.0144
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.0158	0.0228	0.0088	0.0138
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0290	0.0378	0.0196	0.0252
	# reps. satisfying	400	381	400	382
	Satisfied coverage	88.5%	82.94%	96.3%	90.84%
±3.75%	Net CI coverage	94.0%	78.25%	97.2%	87.5%
	Avg. sample size	465,160	296,990	710,070	482,450
	Avg. CI half-length	0.1030	0.1024	0.1040	0.1042
	Var. CI half-length	0.0002	0.0002	0.0002	0.0001
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.0033	0.0062	0.0018	0.0029
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.0033	0.0062	0.0018	0.0028
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0048	0.0042	0.0068	0.0064
	# reps. satisfying	400	375	400	367
	Satisfied coverage	94.0%	83.73%	97.2%	95.37%

In particular, the selected DTMC is the two-state chain $\{Z_i : i = 0, 1, \ldots\}$ on the state space $\{0, 1\}$ with one-step transition probability matrix

$$\mathbf{P} = \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix}.$$

The steady-state marginal distribution for this two-state chain is given by $\boldsymbol{\pi} = (0.5, 0.5)$; and the initial condition Z_0 is sampled from $\boldsymbol{\pi}$ so that the process $\{Z_i\}$ starts in steady-state operation. The associated output ("reward") process $\{X_i = h(Z_i) : i = 0, 1, \ldots\}$ with cost vector $[h(0), h(1)]^{\mathrm{T}} = (5, 10)^{\mathrm{T}}$ has steady-state mean $\mu_X = 7.5$. We chose this as the final test process so as to include in the performance evaluation at least one stochastic model with a discrete steady-state distribution that has a smaller kurtosis than the normal distribution. This is a difficult case for any procedure requiring approximately i.i.d. normal data.

Table 5: Performance of Spectral Procedures for the $M/M/1/M/1$ Queue Waiting Time Process Based on 400 Independent Replications of Nominal 90% and 95% CIs

Prec. Reqt.	Performance Measure	90% CIs WASSP	90% CIs HW	95% CIs WASSP	95% CIs HW
None	Net CI coverage	92.0%	88.75%	96.3%	91.0%
	Avg. sample size	46,462	6,594	46,462	6,594
	Avg. CI half-length	0.7840	1.642	0.9880	2.0494
	Var. CI half-length	0.2260	0.9224	0.3580	1.4369
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.1627	0.8964	0.1627	0.8964
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.1625	0.9030	0.1625	0.9030
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0357	0.0434	0.0373	0.0434
	# reps. satisfying	400	400	400	400
	Satisfied coverage	92.0%	88.75%	96.3%	91.0%
±15%	Net CI coverage	92.0%	85.25%	96.3%	89.25%
	Avg. sample size	48,064	15,033	52,349	21,855
	Avg. CI half-length	0.6400	0.7757	0.7050	0.7928
	Var. CI half-length	0.0440	0.0195	0.0360	0.0165
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.1106	0.2576	0.0993	0.1717
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.1090	0.2528	0.0969	0.1679
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0489	0.0839	0.0565	0.0692
	# reps. satisfying	400	394	400	387
	Satisfied coverage	92.0%	86.55%	96.3%	92.29%
±7.5%	Net CI coverage	89.0%	80.25%	96.5%	85.0%
	Avg. sample size	82,680	52,700	124,368	82,338
	Avg. CI half-length	0.3920	0.4177	0.4000	0.4167
	Var. CI half-length	0.0050	0.0063	0.0040	0.0064
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.0560	0.0838	0.0373	0.0605
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.0553	0.0817	0.0362	0.0589
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0361	0.0504	0.0308	0.0409
	# reps. satisfying	400	375	400	378
	Satisfied coverage	89.0%	85.83%	96.5%	90.21%

From Table 6, we concluded that for the precision levels of ±7.5%, ±3.75%, and ±1.875%, WASSP outperformed HW; and the net coverage probabilities delivered by HW were significantly below the corresponding nominal levels. As observed in the previous test problems, the anomalous behavior of HW at the more stringent levels of precision was caused by the relatively small number of replications that delivered CIs satisfying the relevant precision requirement—for example, in the case of nominal 95% CIs with relative precision of ±7.5%, only $R = 346$ CIs satisfied the precision requirement; and although 92.68% of those CIs actually covered the steady-state mean, the net CI coverage was only 80.0% in this case. We judged the performance of HW to be unacceptable in this test problem; and we concluded that for this problem, there was a substantial probability that HW would require larger sample sizes than WASSP would require to deliver valid CIs.

Table 6: Performance of Spectral Procedures for the Two-State DTMC with High Positive Correlation Structure Based on 400 Independent Replications of Nominal 90% and 95% CIs

Prec. Reqt.	Performance Measure	90% CIs WASSP	90% CIs HW	95% CIs WASSP	95% CIs HW
None	Net CI coverage	93.0%	89.5%	97.0%	95.25%
	Avg. sample size	9,762	1,465	9,762	1,465
	Avg. CI half-length	0.6010	1.1952	0.7570	1.4917
	Var. CI half-length	0.0380	0.0463	0.0600	0.0722
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.0914	0.4215	0.0914	0.4215
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.0914	0.4213	0.0914	0.4213
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0271	0.0615	0.0271	0.0615
	# reps. satisfying	400	400	400	400
	Satisfied coverage	93.0%	89.5%	97.0%	95.25%
±15%	Net CI coverage	93.0%	87.8%	97.0%	93.5%
	Avg. sample size	9,729	2,524	9,691	3,832
	Avg. CI half-length	0.5960	0.9393	0.7320	0.9512
	Var. CI half-length	0.0330	0.0186	0.0370	0.0239
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.0917	0.3401	0.0917	0.2503
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.0917	0.3396	0.0916	0.2450
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0261	0.0598	0.0273	0.0801
	# reps. satisfying	400	400	400	395
	Satisfied coverage	93.0%	87.8%	97.0%	94.9%
±7.5%	Net CI coverage	92.6%	77.5%	96.0%	80.0%
	Avg. sample size	11,581	8,026	15,972	17,718
	Avg. CI half-length	0.4750	0.5284	0.4810	0.4941
	Var. CI half-length	0.0050	0.0253	0.0040	0.0229
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.0674	0.1153	0.0535	0.0683
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.0670	0.1150	0.0531	0.0671
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0307	0.0318	0.0275	0.0365
	# reps. satisfying	400	351	400	346
	Satisfied coverage	92.6%	88.48%	96.0%	92.68%
±3.75%	Net CI coverage	89.7%	80.0%	96.7%	84.5%
	Avg. sample size	41,083	28,223	65,321	49,085
	Avg. CI half-length	0.2390	0.2835	0.2410	0.2625
	Var. CI half-length	0.0010	0.0297	0.0010	0.0220
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.0205	0.0363	0.0119	0.0241
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.0154	0.0358	0.0120	0.0240
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0120	0.0283	0.0089	0.0144
	# reps. satisfying	400	371	400	364
	Satisfied coverage	89.7%	86.46%	96.7%	93.0%
±1.875%	Net CI coverage	93.0%	82.0%	97.5%	84.5%
	Avg. sample size	164,545	109,325	256,620	169,928
	Avg. CI half-length	0.1220	0.1636	0.1230	0.1547
	Var. CI half-length	0.0003	0.0342	0.0003	0.0278
	$\widehat{\text{MSE}}[\overline{X}(n')]$	0.0040	0.0152	0.0026	0.0127
	$\widehat{\text{Var}}[\overline{X}(n')]$	0.0040	0.0151	0.0026	0.0126
	$\lvert\widehat{\text{Bias}}[\overline{X}(n')]\rvert$	0.0047	0.0140	0.0062	0.0113
	# reps. satisfying	400	375	400	362
	Satisfied coverage	93.0%	87.7%	97.5%	93.55%

5 CONCLUSIONS

In all the test processes considered in this article, we concluded that WASSP substantially outperformed HW with respect to conformance to the given requirements on the coverage probability and relative precision of the delivered CIs. Although we found that HW's method for eliminating initialization bias was much more effective in these test processes than in the test processes of Lada et al. (2005), the main problem with HW's performance occurred when HW terminated prematurely, delivering a CI that did not satisfy the precision requirement. This drawback is due to the lack of an effective stopping rule for estimating the final sample size required to ensure normal (nonpremature) termination of HW. By contrast, we concluded that in the selected test processes, WASSP's stopping rule consistently yielded samples of a size sufficient to enable WASSP to deliver a valid CI having the required precision and coverage probability.

We also found that the lack of an effective stopping rule in the HW procedure cannot completely explain the deficiencies in HW's performance in the selected test processes. In many of the test processes considered in this article, we found that even when we restricted consideration to HW's CIs satisfying the precision requirement, the "satisfied coverage" probabilities were still unacceptably low. Because HW appeared to handle initialization bias well in these test problems, we concluded that HW's performance deficiencies were caused by certain properties of its spectral estimator $\widehat{\gamma}_X$ of the variance parameter γ_X defined by Equation (1); in particular we found that $\widehat{\gamma}_X$ may possess a large bias (due to lack of fit in estimating the log-spectrum) as well as a large variance (due to inadequate smoothing of the periodogram). WASSP's wavelet-based estimator of γ_X was specifically designed to avoid these pitfalls.

From the results of the performance evaluation detailed in this article, we concluded that as the precision requirement tended to zero, there was a substantial probability that the HW procedure would require larger sample sizes to deliver a valid CI than WASSP would require. This conclusion is based on the significant CI undercoverage observed with the HW procedure when it was supplied with samples of the size required by WASSP to deliver valid CIs.

As detailed in Lada, Steiger, and Wilson (2006), spectral methods (such as WASSP and HW) appear to require larger average sample sizes in some simulation applications than batch-means procedures (such as ASAP3) require. On pp. 723–724 of Lada, Steiger, and Wilson (2006), possible causes of this phenomenon are suggested. Recent experimentation, however, has revealed the need for a more complete explanation of the larger average sample sizes required by spectral methods. We believe the primary cause of this phenomenon is that the spectral estimators of the variance parameter are neither sufficiently accurate nor sufficiently stable in comparison with their counterparts based on batch means. WASSP's wavelet-based spectral estimator has approximately a scaled chi-squared distribution with only $\nu = 6$ degrees of freedom; and HW's quadratic-regression-based spectral estimator has approximately a scaled chi-

squared distribution with $\nu = 7$ degrees of freedom. In contrast, the effective degrees of freedom associated with the CIs delivered by ASAP3 always satisfy $\nu \geq 55$; see p. 70 of Steiger et al. (2005). We believe that more accurate and stable estimators of the variance parameter are required to obtain substantial improvements in the sampling efficiency of spectral methods. A definitive resolution of this problem and its implementation in future versions of WASSP are the subjects of ongoing research. Additional experimental results, follow-up papers and revised software, will be available on the website <www.ie.ncsu.edu/jwilson>.

REFERENCES

Anderson, T. W., and D. A. Darling. 1952. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics* 23 (2): 193–212.

Heidelberger, P., and P. D. Welch. 1981a. A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM* 24 (4): 233–245.

Heidelberger, P., and P. D. Welch. 1981b. Adaptive spectral methods for simulation output analysis. *IBM Journal of Research and Development* 25 (6): 860–876.

Heidelberger, P., and P. D. Welch. 1983. Simulation run length control in the presence of an initial transient. *Operations Research* 31 (6): 1109–1144.

Lada, E. K. 2003. A wavelet-based procedure for steady-state simulation output analysis. Doctoral dissertation, Graduate Program in Operations Research, North Carolina State University, Raleigh, North Carolina. Available online via <www.lib.ncsu.edu/theses/available/etd-04032003-141616/unrestricted/etd.pdf> [accessed July 15, 2006].

Lada, E. K., and J. R. Wilson. 2006. A wavelet-based spectral procedure for steady-state simulation analysis. *European Journal of Operational Research* 174:1769–1801.

Lada, E. K., J. R. Wilson, and N. M. Steiger. 2003. A wavelet-based spectral method for steady-state simulation analysis. In *Proceedings of the 2003 Winter Simulation Conference*, ed. S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 422–430. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available online via <www.informs-sim.org/wsc03papers/052.pdf> [accessed July 15, 2006].

Lada, E. K., J. R. Wilson, N. M. Steiger, and J. A. Joines. 2004a. User's manual for WASSP Version 1 [online]. Edward P. Fitts Department of Industrial and Systems Engineering, NC State University, Raleigh, NC. Available online via <ftp.ncsu.edu/pub/eos/pub/jwilson/wasspman.pdf> [accessed July 15, 2006].

Lada, E. K., J. R. Wilson, N. M. Steiger, and J. A. Joines. 2004b. Performance evaluation of a wavelet-based spectral method for steady-state simulation analysis. In *Proceedings of the 2004 Winter Simulation Conference*, ed. R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, 694–702. Available online via <www.informs-sim.org/wsc04papers/084.pdf> [accessed July 15, 2006].

Lada, E. K., J. R. Wilson, N. M. Steiger, and J. A. Joines. 2005. Performance of a wavelet-based spectral procedure for steady-state simulation analysis. *INFORMS Journal on Computing* to appear. Available online via <ftp.ncsu.edu/pub/eos/pub/jwilson/lada05joc.pdf> [accessed July 15, 2006].

Lada, E. K., N. M. Steiger, and J. R. Wilson. 2006. Performance evaluation of recent procedures for steady-state simulation analysis. *IIE Transactions* 38 (9): 711–727. Available online via <ftp.ncsu.edu/pub/eos/pub/jwilson/lada06iie.pdf> [accessed July 15, 2006].

Law, A. M., and J. S. Carson. 1979. A sequential procedure for determining the length of a steady-state simulation. *Operations Research* 27 (5): 1011–1025.

Shapiro, S. S., and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52 (3–4): 591–611.

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 output analysis. *ACM Transactions on Modeling and Computer Simulation* 15 (1): 39–73. Available online via <ftp.ncsu.edu/pub/eos/pub/jwilson/steiger05tomacs.pdf> [accessed July 16, 2006].

von Neumann, J. 1941. Distribution of the ratio of the mean square successive difference to the variance. *The Annals of Mathematical Statistics* 12 (4): 367–395.

AUTHOR BIOGRAPHIES

EMILY K. LADA is an operations research development tester at the SAS Institute. She is a member of IIE and INFORMS. Her e-mail address is <Emily.Lada@sas.com>.

JAMES R. WILSON is the Edgar S. Woolard Professor and Head of the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. He is a member of AAUW, ACM, and ASA, 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>.