Proceedings of the 1996 Winter Simulation Conference
ed. J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain

EXPERIMENTAL EVALUATION OF CONFIDENCE INTERVAL PROCEDURES IN SEQUENTIAL STEADY-STATE SIMULATION

Donald C. McNickle
Krzysztof Pawlikowski
Gregory Ewing

University of Canterbury
Christchurch
NEW ZEALAND

ABSTRACT

Sequential analysis of simulation output is generally accepted as the most efficient way for securing representativeness of samples of collected observations. In this scenario a simulation experiment is stopped when the relative precision of estimates, defined as the relative width of confidence intervals at an assumed confidence level, reaches the required level. This paper deals with the statistical correctness of the methods proposed for estimating confidence intervals for mean values in sequential steady-state stochastic simulation. We formulate basic rules that should be followed in proper experimental analysis of coverage of different steady-state interval estimators. Our main argument is that such analysis should be done sequentially. The numerical results of our preliminary coverage analysis of the method of Spectral Analysis (SA/HW) and Non-overlapping Batch Means are presented, and compared with those obtained by traditional, non-sequential approaches.

1 INTRODUCTION

Sequential analysis of simulation output is generally accepted as the most efficient way for securing representativeness of samples of collected observations [see, for example, Law and Kelton (1992)]. In this scenario a simulation experiment is stopped when the relative precision of estimates, defined as the relative width of confidence intervals at an assumed confidence level, reaches the required level.

This paper deals with the statistical correctness of the methods proposed for estimating confidence intervals of mean values in sequential steady-state stochastic simulation. The main analytical problems of such simulation studies were discussed in Law (1983) and Pawlikowski (1990). They are caused by correlations between events observed during typical simulated processes. At least a dozen methods have been proposed for estimating confidence intervals of mean values from series of correlated observations collected during such simulation. A survey of the methods until 1990 can be found in Pawlikowski (1990). Newer proposals include those by Fox et al. (1991), Goldsman and Kang (1991), Howard et al. (1992). So far only a few implementations of these methods in an automated sequential simulation framework have been reported [see for example Fox et al. (1991), Heidelberger and Welch (1983), Pawlikowski et al. (1994), Rego and Sunderam (1992), Yau and Pawlikowski (1993)] and incorporated in some simulation packages. The methods are based on different approximations and their quality should be assessed by analysing the properties of the final confidence intervals they generate. A good method should produce narrow and stable confidence intervals, which should of course be valid, ie. they should contain the true value of the estimated performance measure (with the correct probability). Theoretical studies of various estimators of confidence intervals, reported before 1990, are surveyed in Pawlikowski (1990). Newer results can be found for example in Kang and Goldsman (1990).

Unfortunately, no satisfactorily exhaustive comparative studies of these methods have been reported yet, and it is difficult to find a good method for a specific range of applications. Additionally, most studies relate to non-sequential simulation experiments run on single processors. Very little is known about quality of these methods in sequential simulation, and in fast concurrent sequential simulations based on Multiple Replications in Parallel (MRIP), when multiple processors cooperate in production of data – see Pawlikowski et al. (1994).

The theoretical studies of confidence intervals reveal general conditions which have to be satisfied to ensure the validity of the final confidence intervals, but the correctness of any practical implementation of a specific method also has to be tested experimentally. In this paper we formulate a new methodology of such experimental studies of the methods used in sequential stochastic simulation for determining the final precision of results, and present the results of our comparative studies of two selected methods: SA/HW (the method of Spectral

Analysis in its version proposed in Heidelberger and Welch (1981), and the classical method of (non-overlapping) Batch Means, both in sequential simulations on single processors and in sequential simulations on multiple processors in MRIP scenario. Further directions of research in this area are indicated in the Conclusions.

2 EXPERIMENTAL ANALYSIS OF COVERAGE

In any performance evaluation studies of dynamic systems by means of stochastic discrete-event simulation the final estimates should be determined together with their statistical errors, which are usually measured by the half-width of the final confidence intervals. Restricting our attention to estimators of means, let us assume that we estimate theoretical mean $\mu = EX$ by

$$\overline{X}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where $x_1, x_2, ..., x_n$ are observations collected during simulation. Then, one should also determine

$$P(\overline{X}(n)-\Delta \le \mu \le \overline{X}(n)+\Delta) = 1 - \alpha$$

i.e. the confidence interval (c.i.) of μ, at a given confidence level $1-\alpha$, $0<\alpha<1$. Δ is the half-width of the c.i. Typically, this will be $\Delta = t_{\kappa,1-\alpha/2}\hat{\sigma}[\overline{X}(n)]$, where $\hat{\sigma}^2[\overline{X}(n)]$ is an estimator of the variance of $\overline{X}(n)$, with κ degrees of freedom and $t_{\kappa,1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the Student t-distribution.

Various estimators of $\hat{\sigma}[\overline{X}(n)]$ have been proposed. This in sequel has created the need for a quality assessment of these estimators.

Let us note that in an ideal case the final c.i. would contain μ with the probability $1-\alpha$, or equivalently, if an experiment is repeated many times, one would expect to have μ contained in about $(1-\alpha)100\%$ of final confidence intervals. *Coverage of confidence intervals* is defined as the frequency with which the final confidence intervals $(\overline{X}(n) -\Delta \le \mu \le \overline{X}(n) +\Delta)$ contain the true value μ. While some interesting results have been achieved in theoretical studies of coverage [see eg. Glynn (1982), Kang and Goldsman (1990), Schruben (1980)], experimental analysis of coverage is still required for assessing the quality of practical implementations of methods used for determining confidence intervals in steady-state simulation. Of course, such analysis is limited to analytically tractable systems, since the value of μ has to be known.

As for any other point estimate, the coverage can be determined together with its c.i. :

$$\left(c - z_{1-\alpha/2}\sqrt{\frac{c(1-c)}{n_c}}\ ,\ c + z_{1-\alpha/2}\sqrt{\frac{c(1-c)}{n_c}}\right) \quad (1)$$

where c is the coverage, $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution and n_c is the (suitably large) number of replicated experiments in the coverage analysis. In this study, n_c is at least 600, so the use of a normal approximation seems to be justified.

Then, an estimator of $\hat{\sigma}^2[\overline{X}(n)]$ used for determining the c.i. of μ can be considered as valid, ie. producing valid $100(1-\alpha)\%$ confidence intervals of μ, if the upper bound of the confidence interval of the coverage c in Equation (1) equals at least $(1-\alpha)$; see Sauer (1977).

Results of experimental coverage analysis have been reported in many publications, although majority of these results are related to simulations on single processors. Very little is known about coverage of estimators that could be used in parallel simulation executed in the MRIP scenario [Pawlikowski et al. (1994)]. It is strange, but while sequential simulation is generally recognised as the only way of producing results with the required precision since "*..no procedure in which the run length is fixed before the simulation begins can be relied upon to produce a c.i. that covers the true steady-state mean with the desired probability level*" [Law and Kelton (1982), Law and Kelton (1992)], even the original advocates of sequential simulation have applied non-sequential (fixed-sample size) approaches in their simulation studies of coverage. Certainly, if one accepts the arguments for the sequential approach as the only practical one then also such meta-simulation experiments as those for coverage analysis should be run sequentially!

In addition, most reported results on coverage were based on 50-200 replications [see for example Adam (1983), Fishman (1978), Heidelberger and Welch (1983), Kelton and Law (1984), Law and Carson (1979), Law and Kelton (1982), Lavenberg and Sauer (1977), Sauer (1979), Schriber and Andrews (1981), Schruben (1983)], which obviously puts in question the statistical representativeness of such experimental data. In all these cases, the estimates of coverage were based on only a few bad confidence intervals, ie. ones which did not contain μ.

This issue could be solved by requiring that the coverage can be estimated only after a minimum number of bad confidence intervals has been recorded.

It is also generally known that sequential steady-state simulation can produce very inaccurate estimates if the stopping criterion is only accidentally temporarily satisfied. Sensible practise is to ensure that estimates do not come from simulation runs that are too short. Thus, this effect should be similarly treated, and eliminated, in coverage analysis.

Recognising the significance of all these three factors, we have applied the following rules in experimental analysis of coverage of interval estimators from stochastic simulation:

R1. Coverage should be analysed sequentially, ie. analysis of coverage should be stopped when the relative precision (the relative half-width of c.i.) of the estimated coverage falls below an assumed level.

R2. An estimate of coverage has to be calculated from a representative sample of data, ie. the coverage analysis can start only after a minimum number of bad confidence intervals has been recorded.

R3. Results from simulation runs that are abnormally short should be not taken into account.

Details of our implementation of these rules for studying the quality of the final steady-state interval estimators of means in traditional simulation on single processors as well as fast concurrent simulation on multiple processors is discussed in the next section.

3 NUMERICAL RESULTS

Implementing the rules of coverage analysis formulated in the previous section, we must select (i) the minimum number of bad confidence intervals, Nmin, which have to be recorded before the sequential analysis of coverage can start, and (ii) the minimum sufficient length of simulation to produce valid steady-state estimates. The way in which we approached these problems is illustrated in Fig.1 and 2.

The results presented there are for SA/HW, the method of Spectral Analysis in its version proposed in Heidelberger and Welch (1981). Our implementation of this method for simulations on single processors followed exactly procedures specified in Pawlikowski (1990), including the procedure described there for detecting the length of the initial transient period. Its generalisation for simulations in the MRIP scenario was described in Pawlikowski et al. (1994).

All reported results were obtained by stopping simulations when the final steady-state results reached a (relative) precision of 5% or less, at the 0.95 confidence level. All series of replicated simulations were executed using strictly non-overlapping sequences of pseudo-random numbers generated by a linked sequence of congruential generators listed in Law and Kelton (1992).

The results presented in Fig.1 and 2 were obtained by running multiple independent replications of sequential steady-state simulation of M/D/1/∞ queuing system on P=1 and 4 processors, respectively. The estimated parameter is the mean time that a customer spends in the queue. In all four cases the analysis of coverage was initiated after observing N_{min} bad confidence intervals. This happened after about 300 independent replications in the case of N_{min}=30 [Fig.1(a) and 2(a)] and after about 2000 independent replications in the case of N_{min}=200 [Fig.1(a) and 2(a)]. At this stage the lengths of executed simulations were analysed, and the results obtained from simulation runs shorter than L_{min} (one standard deviation below the average number of observations needed to stop sequential simulation with the required precision at the assumed confidence level) were discarded (hence the sudden improvement in coverage). L_{min} was also later used as the criterion for rejecting/accepting results from any additional replication in cases where sequential analysis of coverage had to be continued.

One can see that in all four cases filtering out too short simulation runs removes significant noise. On the other hand, continued instability of the coverage observed after N_{min}=30 bad confidence intervals have been collected shows that, possibly due to strong asymmetry of the sample distribution, many more than 30 bad confidence intervals had to be recorded to secure representativeness in the analysed data. This conclusion, on the basis of similar results we obtained for other queuing systems, suggested that many more replications were needed than used in previous studies. For this reason, in our further analysis of coverage we assumed N_{min}=200.

The results of sequential coverage analysis for SA/HW in simulations executed on a single processor, as well as for simulations in MRIP scenario for P=2 and 4 processors, are presented in Fig.3 (a)-(c), respectively. The performance of the SA/HW method improves with the number of processors or, equivalently, the number of independent simulation engines, used in the MRIP scenario. Another attractive feature of SA/HW is its good performance when simulating heavily loaded systems, ie. in the region where other methods usually fail. The "safe" degree of parallelisation for SA/HW has yet to be determined.

For comparison, Fig. 3(d) shows results one could obtain applying traditional fixed-sample-size approach, and estimating the coverage on the basis of first 200 replications. It is evident that the results obtained from traditional analyses of coverage cannot be considered as reliable.

Coverage analysis of different methods proposed for estimating confidence intervals of mean values in sequential steady-state stochastic simulation is illustrated here by the results obtained for the method of non-overlapping batch means (BM); see Fig.4 (a) and (b). The method was implemented following procedures specified in Pawlikowski (1990), including the procedure described there for detecting the length of the initial transient period. In the case of simulations executed on a single processor, SA/HW and BM offer similar (bad) coverage. When more processors are used under MRIP,

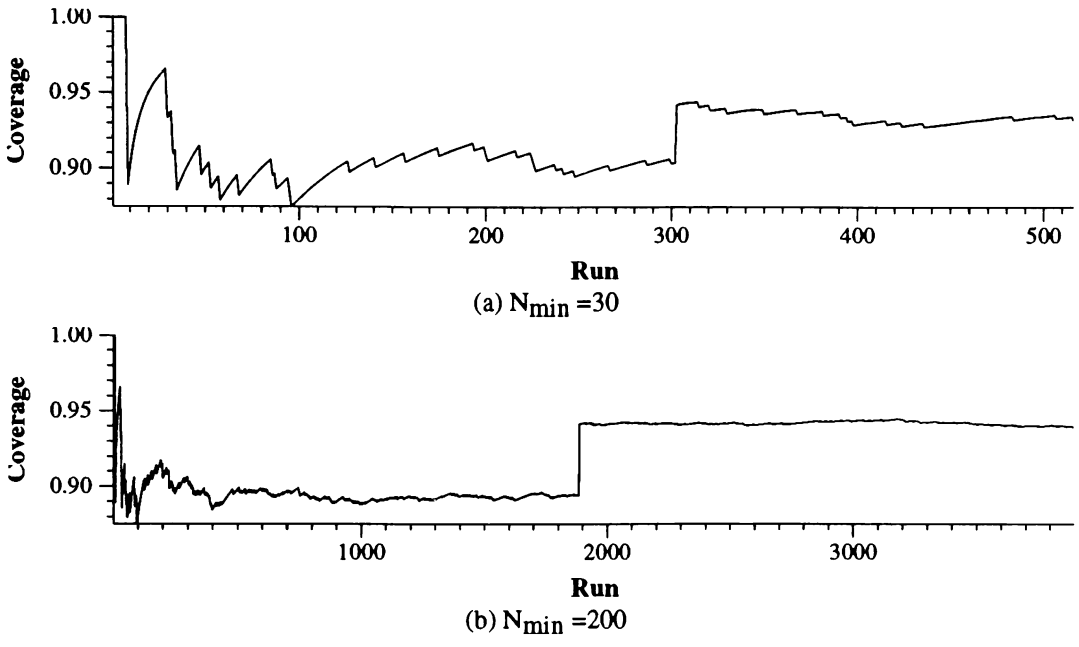


Figure 1: Coverage as a Function of Sample Size for SA/HW in Steady-State Simulation of an M/D/1/∞ Queuing System for ρ=0.5, P=1 processor

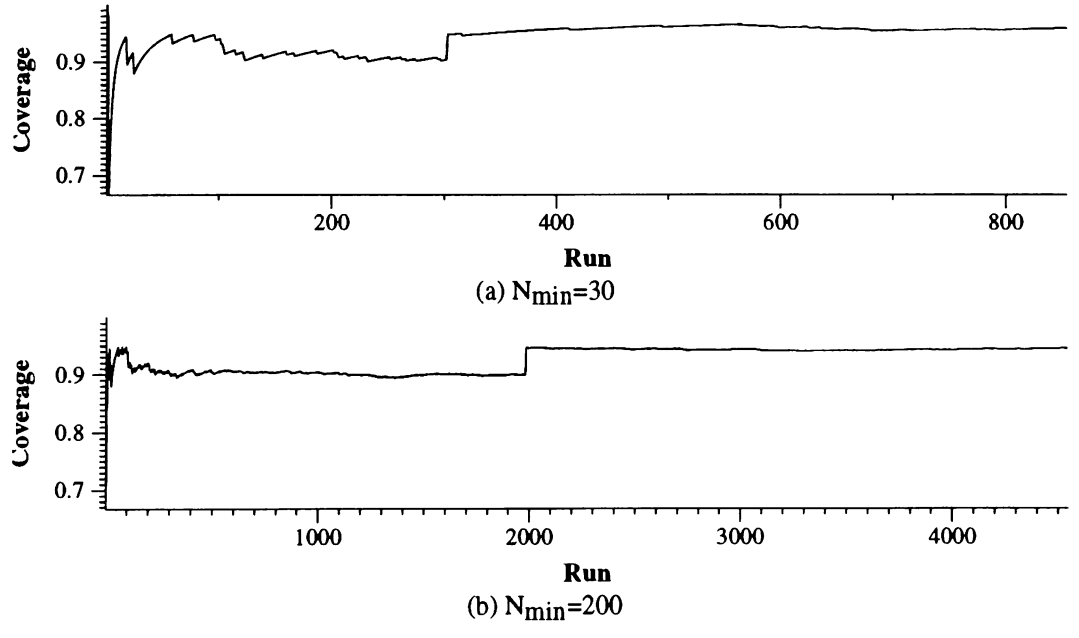
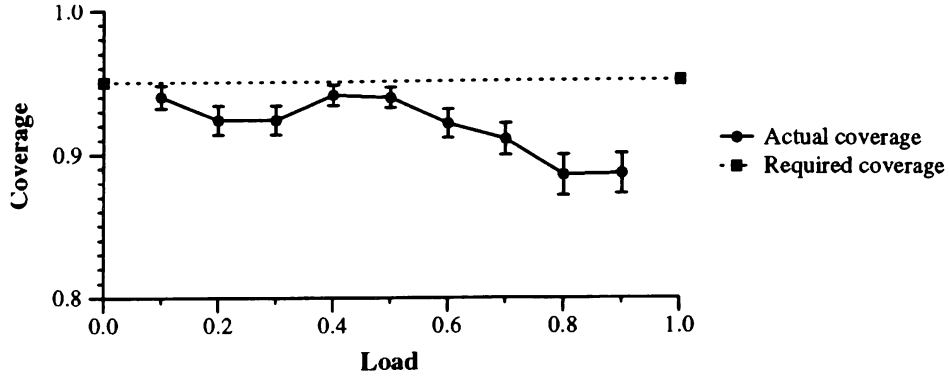
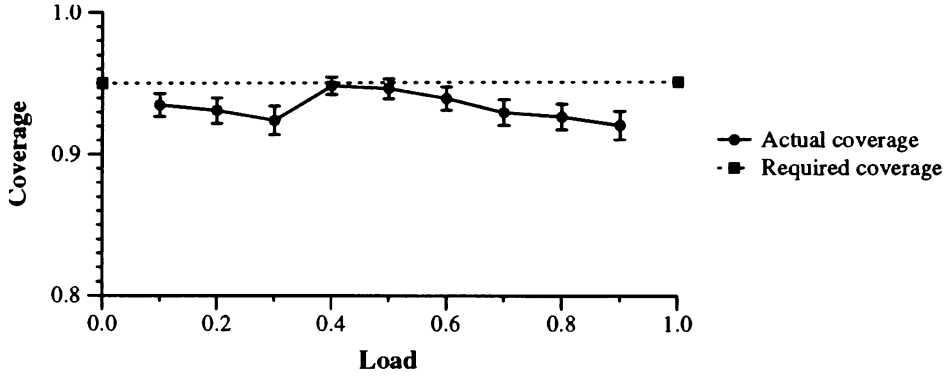


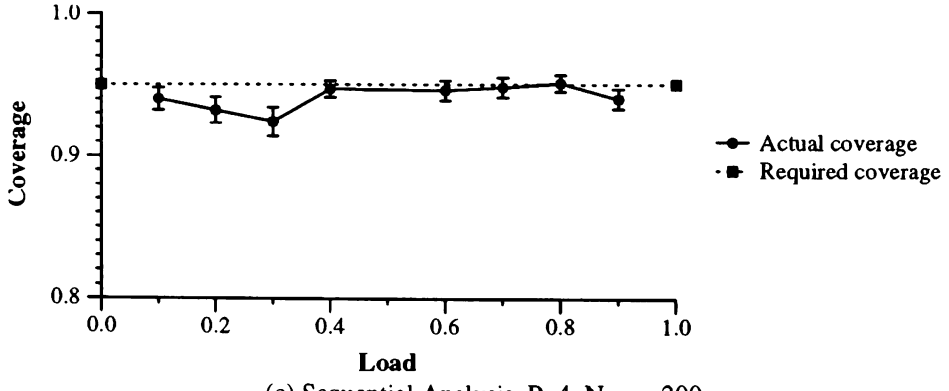
Figure 2: Coverage as a Function of Sample Size for SA/HW in Steady-State Simulation of an M/D/1/∞ Queueing System for ρ=0.5, P=4 processors



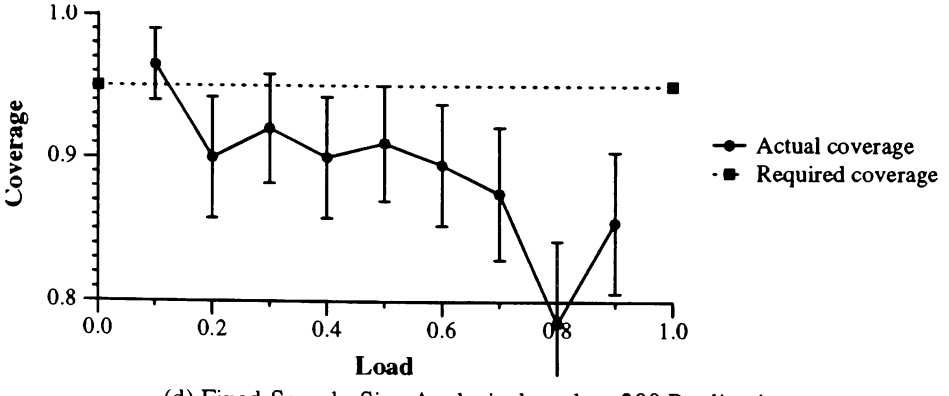
(a) Sequential Analysis, P=1, N_{min} =200



(b) Sequential Analysis, P=2 N_{min}=200

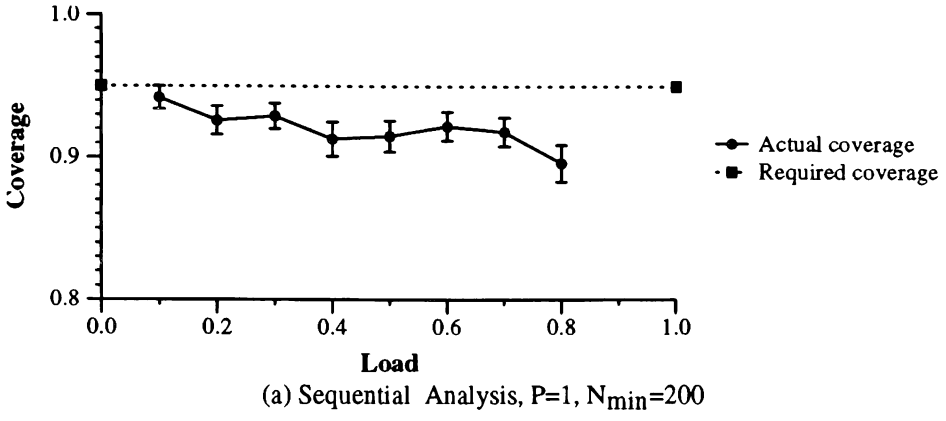


(c) Sequential Analysis, P=4, N_{min}=200

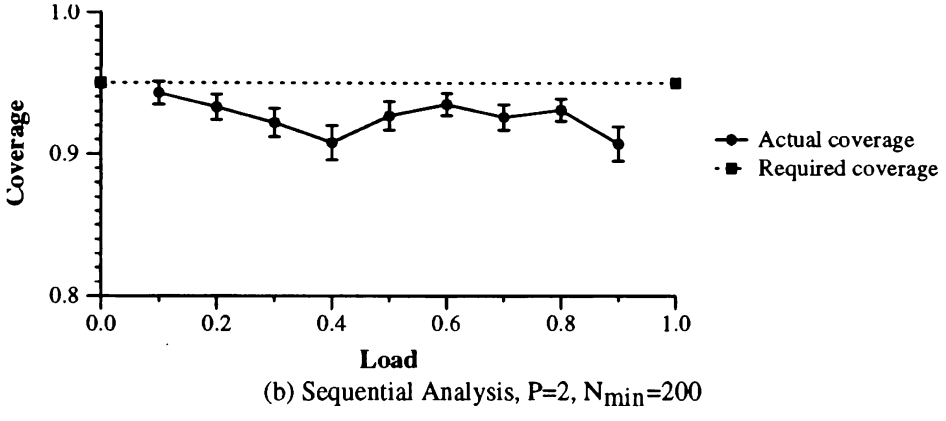


(d) Fixed-Sample-Size Analysis based on 200 Replications

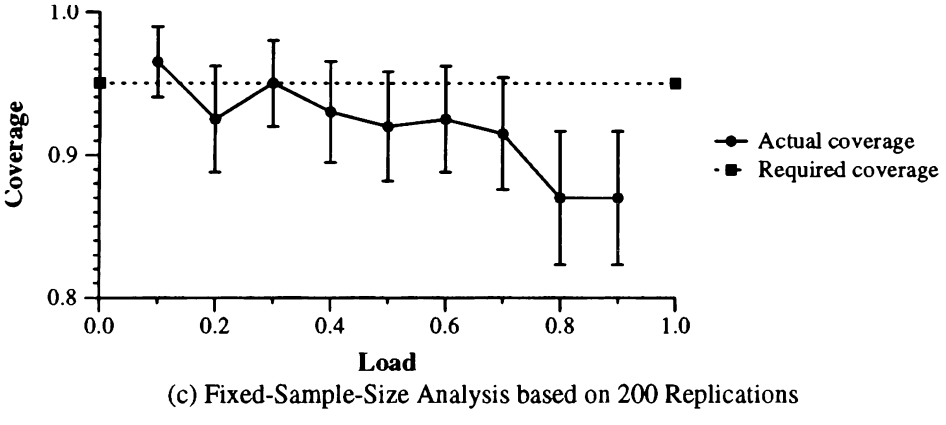
Figure 3: Coverage of SA/HW in Steady-State Simulation of an M/D/1/∞ Queueing System



(a) Sequential Analysis, P=1, N_{min}=200



(b) Sequential Analysis, P=2, N_{min}=200



(c) Fixed-Sample-Size Analysis based on 200 Replications

Figure 4: Coverage of BM in Steady-State Simulation of an M/D/1/∞ Queuing System

using SA/HW one can obtain much better results. This has been consistently observed both when simulating M/M/1/∞ and M/D/1/∞ buffers. Fig. 4(c) again shows results one could obtain applying the traditional fixed-sample-size approach, estimating the coverage on the basis of first 200 replications. These results are much too optimistic if compared with the ones obtained on the basis of representative samples from sequential analysis of coverage. Our experimental comparative studies of different methods proposed for estimating confidence intervals of mean value in sequential steady-state stochastic simulation are continued.

4 CONCLUSIONS

We have formulated basic rules that should be followed in proper experimental analysis of coverage of different steady-state interval estimators. Our main argument is that such analysis should be done sequentially. The numerical results of our preliminary coverage analysis of the method of Batch Means and Spectral Analysis (SA/HW) have been also presented and compared with those obtained by traditional, non-sequential approach. As advocated in Law (1983), to draw more general conclusions about performance of interval estimators used in various methods of sequential steady-state simulation one needs to consider a number of different simulation models. Unfortunately until now no such standard set of reference models for coverage analysis has been wider adopted, in spite of that the issue being raised already in 1981 in Schriber and Andrews (1981).

REFERENCES

Adam, N. R. 1983. Achieving a Confidence Interval for Parameters Estimated by Simulation. *Management Science* 29: 856-866.

Fishman, G. S. 1978. Grouping Observations in Digital Simulation. *Management Science* 24: 510-521.

Fox, B. L., D. Goldsman and J. Swain. 1991. Spaced Batch Means. *Operations Research Letters* 10: 255-263.

Glynn, P. W. 1982. Coverage Error for Confidence Intervals Arising in Simulation Output Analysis. In *Proceedings of the 1982 Winter Simulation Conference*, IEEE Press, 369-375.

Goldsman, D., and K. Kang. 1991. Cramer-von Mises Variance Estimators for Simulations. In *Proceedings of the 1991 Winter Simulation Conference*, IEEE Press, 916-920.

Goldsman, D., et al. 1986. Large and Small Sample Comparisons of Various Variance Estimators. In *Proceedings of the 1982 Winter Simulation Conference* 278-284. IEEE Press.

Heidelberger, P., and P. D. Welch. 1981. A Spectral Method for Confidence Interval Generation and Run Length Control in Simulation. *Communications of the ACM* 25: 233-245.

Heidelberger, P., and P. D. Welch. 1983. Simulation Run Length Control in the Presence of an Initial Transient. *Operations Research*. 31: 1109-1144.

Howard, R. B. et al. 1992. Confidence Intervals for Univariate Discrete-Event Simulation Output Using the Kalman Filter. In *Proceedings of the 1992 Winter Simulation Conference* 586-593.

Kang, K., and D. Goldsman. 1990. The Correlation Between Mean and Variance Estimators in Computer Simulation. *Transactions of the IIE* 22 (1): 15-23.

Kelton, W. D., and A. M. Law. 1984. An Analytical Evaluation of Alternative Strategies in Steady-State Simulation. *Operations Research* 32 (1): 169-184.

Lavenberg, S. S., and C. H. Sauer. 1977. Sequential Stopping Rules for the Regenerative Method of Simulation. *IBM Journal of Research and Development* 21: 667-678.

Law, A. M., and J. S. Carson. 1979. A Sequential Procedure for Determining the Length of a Steady-State Simulation. *Operations Research* 27: 1011-1025.

Law, A. M., and W. D. Kelton. 1982. Confidence Intervals for Steady-State Simulations, II: A Survey of Sequential Procedures. *Management Science* 28 (5): 550-562.

Law, A. M. 1983. Statistical Analysis of Simulation Output Data. *Operations Research* 31 (6): 983-1029.

Law, A. M., and W. D. Kelton. 1992. *Simulation Modeling and Analysis*. New York: McGraw-Hill.

Pawlikowski, K. 1990. Steady-State Simulation of Queueing Processes: A Survey of Problems and Solutions. *ACM Computing Surveys* 22 (2): 123-170 (plus Corrigendum, *ACM Computing Surveys* 224, 409).

Pawlikowski, K., V. Yau and D. McNickle. 1994. Distributed Stochastic Discrete-Event Simulation in Parallel Time Streams. In *Proceedings of the. 1994 Winter Simulation Conference* 723-730. IEEE Press.

Rego, V. J., and V. S. Sunderam. 1992. Experiments with Concurrent Stochastic Simulation: the EcliPSe Paradigm. *Journal of Parallel and Distributed Computing* 14: 66-84.

Sauer, C. H. 1979. Confidence Intervals for Queueing Simulations of Computer Systems. *ACM Performance Evaluation Review* 8 (1-2): 46-55.

Schriber, T. J. and R. W. Andrews. 1981. A Conceptual Framework for Research in the Analysis of Simulation Output. *Communications of the ACM* 24 (4): 218-232.

Schruben, L. W. 1980. A Coverage Function for Interval Estimators of Simulation Response. *Management Science* 26: 18-27.

Schruben, L. W. 1983. Confidence Interval Estimation Using Standardized Time Series. *Operations Research* 30: 1090-1108.

Yau, V., and K. Pawlikowski. 1993. AKAROA: a Package for Automatic Generation and Process Control of Parallel Stochastic Simulation. In *Proceedings of the. 16th Australian Computer Science Conference* 71-82. Australian Computer Science Communications.

AUTHOR BIOGRAPHIES

DON MCNICKLE is a Senior Lecturer in Management Science in the Deparmtent of Management at the University of Canterbury. His research interests include queueing theory, networks of queues and statistical aspects of stochastic simulation. He is a member of INFORMS.

KRYS PAWLIKOWSKI is an Associate Professor of Computer Science at the University of Canterbury. His research interests include quantitative stochastic

simulation, and performance modelling and evaluation of telecommunication networks. He received a PhD in Computer Engineering from the Technical University of Gdansk, Poland in 1975 He is a Senior Member of IEEE.

GREG EWING is a research associate in the Department of Computer Science at Canterbury. He recently completed his Ph.D. on Georgraphic Information Systems. His research interests include languages, 3D graphics and graphical user interfaces.