DO NOT TRUST TOO SHORT SEQUENTIAL SIMULATION

Jong-Suk R. Lee, Krzysztof Pawlikowski and Don McNickle[†]
Department of Computer Science and [†]Management
University of Canterbury, Christchurch, New Zealand
e-mail: {ruth, krys}@cosc.canterbury.ac.nz

KEYWORDS

Stochastic simulation, steady-state simulation, sequential output data analysis, sequential stopping rules, coverage of confidence intervals

ABSTRACT

Sequential stochastic simulation has been widely accepted as the only practical way of controlling statistical errors of the final simulation results. Such simulation evolves along a sequence of consecutive checkpoints at which the accuracy of estimates, conveniently measured by the relative statistical error (defined as the ratio of the half-width of a given CI, at an assumed confidence level, and the point estimate) is assessed.

Inherently random nature of output data collected during stochastic simulation can cause an accidental, temporal satisfaction of the stopping rule of such sequential estimation.

In this paper, having given an experimental evidence of frequent occurrence of this phenomenon, and the resulted significant degradation of the coverage of the final results in such cases, we propose a simple heuristic rule which helps to solve this problem. The effectiveness of this rule of thumb is quantitatively assessed on the basis of the results of coverage analysis of a few methods of sequential output data analysis in the context of steady-state simulation.

INTRODUCTION

Any stochastic discrete-event simulation has to be regarded as a (simulated) statistical experiment. Hence, statistical analysis of simulation output is mandatory. Otherwise, *"... computer runs yield a mass of data but this mass may turn into a mess* <if the random nature of such output data is ignored, and then> *... instead of an expensive simulation model, a toss of the coin had better be used"* (Kleijnen 1979).

Two different scenarios for determining the duration of stochastic simulation exist. Traditionally, the length of simulation experiment was set as an input to simulation programs. In such a *fixed-sample-size scenario*, the final statistical error of the results is a matter of luck. This is no longer an acceptable approach. Modern simulation methodology offers an attractive alternative, known as the *sequential scenario* of simulation or, simply, *sequential simulation.* Today, the sequential scenario is recognised as the only practical approach allowing control of the error of the final results of stochastic simulation, since *"... no procedure in which the run length is fixed before the simulation begins can be relied upon to produce a confidence interval that covers the true mean with the desired probability level"* (Law and Kelton 1982; Law and Kelton 1991).

Statistical errors associated with the final results of such simulation are commonly measured by the half-widths of their confidence intervals (CI), at a given confidence level. In any correctly implemented simulation, the width of a CI will tend to shrink with the number of collected simulation output data, i.e. with the duration of simulation.

Sequential simulation follows a sequence of consecutive checkpoints at which the accuracy of estimates, conveniently measured by the *relative statistical error* (defined as the ratio of the half-width of a given CI, at an assumed confidence level, and the point estimate) is assessed. Thus, in the case of simulation during which, for example, a mean value μ is estimated, when n observations (or output data items) are available at a given checkpoint and the estimate of μ equals $\bar{X}(n)$, the relative statistical error of the estimate is measured by $\epsilon(n) = \frac{\Delta(n)}{\bar{X}(n)}$, where $\Delta(n)$ is the current half-width of the confidence interval for μ at $(1-\alpha)$ confidence level;

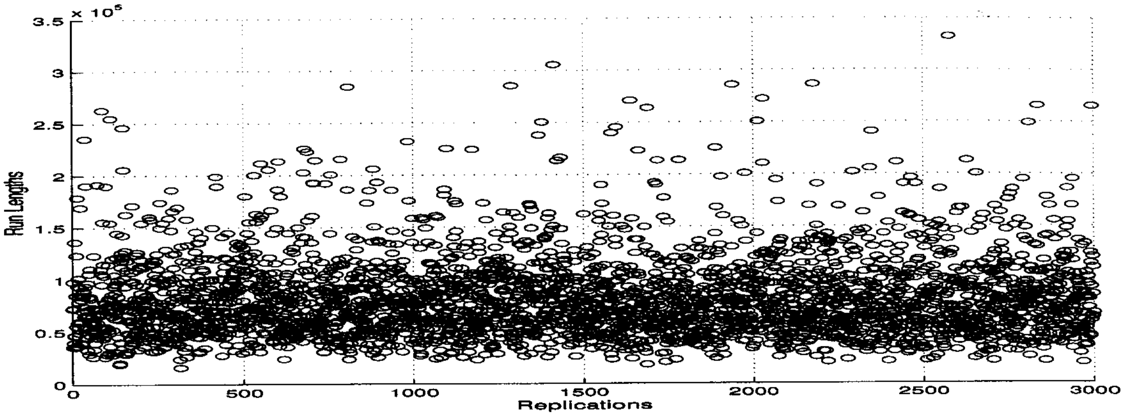

Figure 1: Run lengths for sequential NOBM (M/M/1/∞, load = 0.9)

$0 < \alpha < 1$.

If the acceptable upper level of relative statistical error of the results equals ϵ_{max}, then the simulation can be stopped at a given checkpoint, iff $\epsilon(n) \leq \epsilon_{max}$. Otherwise, the simulation has to be continued. Note, that, to apply this criterion, a simulator does not even need to know the order of magnitude of the estimated parameter(s).

Inherently random nature of output data collected during stochastic simulation can cause an accidental, temporal satisfaction of the stopping rule of such sequential estimation.

An experimental evidence of this phenomenon, and the resulted significant degradation of the coverage of the final results, is documented in Section 2. In Section 3 we propose a simple heuristic rule which offers a solution of the problem. Its effectiveness is quantitatively assessed on the basis of the results of coverage analysis of a few methods of sequential output data analysis in the context of steady-state simulation.

EXPERIMENTAL EVIDENCE

One problem faced in practical applications of sequential stochastic simulation is that a stopping RULE based on relative statistical error can be accidently satisfied too early, giving very inaccurate estimates of the analysed parameters. This happens due to random fluctuations in the estimated relative error occurring during sequential simulation; see for example (Pawlikowski and de Vere 1993).

This phenomenon of prematurely stopped simulation runs can be documented experimentally using results of our exhaustive studies of coverage of various methods of output data analysis proposed for sequential steady-state analysis, in which we followed the methodology proposed in (Pawlikowski et al. 1998) and (Lee et al. 1999). In this paper we restrict our discussion to three methods of sequential mean value analysis: Non-overlapping Batch Means (NOBM), Spectral Analysis in its version proposed by Heidelberger and Welch (SA/HW), and Regenerative Cycles (RC), also known simply as regenerative simulation. The theoretical bases of these three methods of simulation output data analysis, as well sequential implementations of two first methods, are given for example in (Pawlikowski 1990). Our sequential implementation of the RC method is described in (Lee 1999).

Since experimental investigation of consequences of too short simulation runs requires that the exact values of analysed parameters are known, we use results obtained from steady-state simulation of the $M/M/1/\infty$ queueing system. This queueing system is notorious for strong (auto)correlations of data in output sequences and long simulation runs required for achieving satisfactorily low level of statistical errors, and, because of this, it has been proposed as the reference model in research on methods of simulation output data analysis (Schriber and Andrews 1981).

All the results were obtained from 3,000 independent replications of steady-state simulations of an $M/M/1/\infty$ queueing system, estimating the mean response time in the system, with $\epsilon_{max} \cdot 100\% = 10\%$ as the upper level of the acceptable relative error of the final results, at a confidence level of 0.95.

Figures 1-3 show the recorded run-lengths of 3,000 simulation runs of an $M/M/1/\infty$ queueing system at a load of 0.9, and figures 4-6 depict the same data by means of the corresponding histograms. The run-lengths were measured by the number of collected observations.

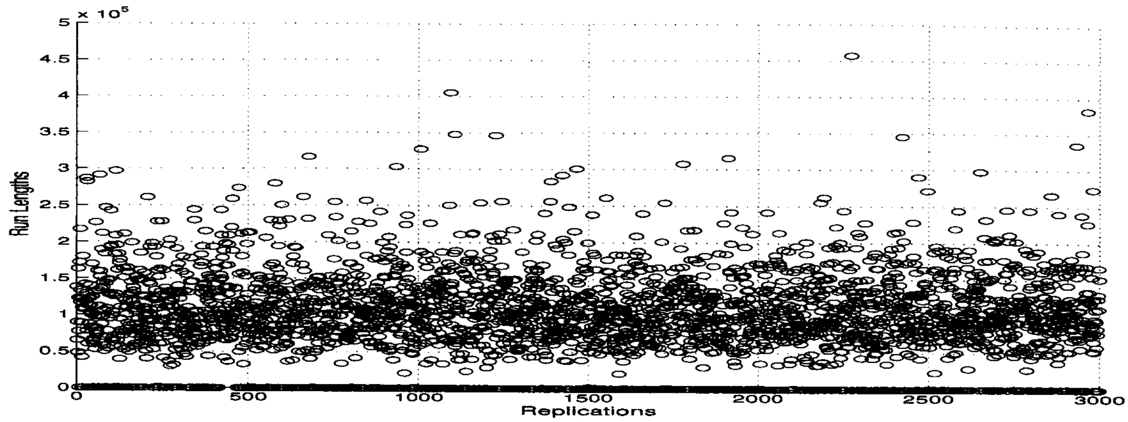

Figure 2: Run lengths for sequential RS (M/M/1/∞, load = 0.9)

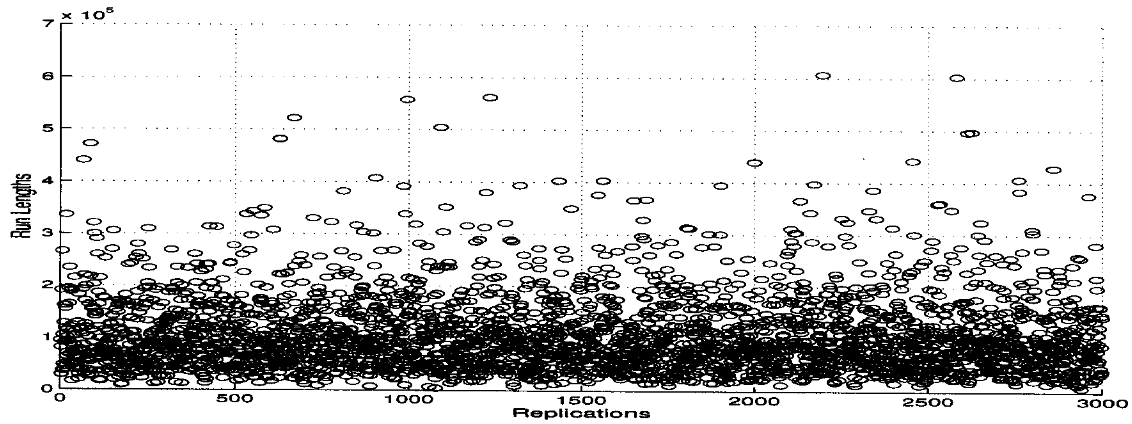

Figure 3: Run lengths for sequential SA (M/M/1/∞, load = 0.9)

Statistics describing the sets of recorded run-lengths are presented in Tables 1-3. Following (Pawlikowski et al. 1998), we have classified a simulation as "too short" if its run-length was shorter than the mean run-length by more than one standard deviation. The overall mean run-lengths and threshold values for sufficiently long simulations are given in the last two columns of the tables, together with the number of replications classified as "too short" (in the second column).

The quality of the results produced by "too short" simulations can be assessed by their coverage, i.e. by the experimental frequency with which the final confidence intervals contain the true (estimated) value. In the ideal situation, the coverage should be equal to the assumed confidence level. The results of our analysis reveal that the coverage of simulation results from "too short" runs can be very poor indeed. While this should be of concern in the case of all three methods considered, the coverage of results associated with RC is really appallingly low; see the third column in Table 2.

On the other hand, other results of our coverage analysis of NOBM, SA and RC, obtained by applying the principles of sequential coverage analysis formulated in (Pawlikowski et al. 1998), show that all three methods are able to offer the final results of similar (acceptable) quality (in the sense of coverage) if the "too short" runs are eliminated; see figure 7-9. A jump in the current value of coverage, clearly seen in each of these figures is associated with discarding of all results taken from "too short" simulations. In all these cases, the lengths of simulation runs were classified using the "mean run-length minus one standard deviation" threshold.

The results show how much wrong the results obtained from too short simulation runs can be. The question is how one can recognise that a given simulation has lasted too short in practical applications of sequential simulation.

Table 1: Run-length statistics from 3,000 simulation replications: NOBM, $M/M/1/\infty$, theoretical confidence level = 0.95.

Load	*Num. of short runs*	*Coverage*	*Prob(short)*	*Threshold*	*Mean of lengths*
0.1	0	N/A	N/A	8603	11897
0.2	0	N/A	N/A	8594	11908
0.3	0	N/A	N/A	8481	11998
0.4	0	N/A	N/A	8484	12167
0.5	0	N/A	N/A	8446	12403
0.6	0	N/A	N/A	8410	13246
0.7	0	N/A	N/A	8892	15539
0.8	154	42.2%	5.1%	12493	24942
0.9	331	36.3%	11.0%	43081	82083

PROPOSED SOLUTION

On the basis of the above reported results we propose the following, simple rule of thumb which should help to eliminate results obtained from too short simulation runs in practical simulation studies. Namely, one should:

- execute R independent replications of a given sequential simulation and record its length (measured by the size of the sample of simulation output data),
- accept the results produced by the longest simulation only.

One can assess the probability of the error associated with such a decision rule for a given R as equal P_{short}^{R}, with P_{short} being the probability that a simulation run is "too short". This is the probability of all R replications belonging to the class of "too short" simulations.

Our experimental data allow us to assume that a "too short" simulation run can occur with the probability $P_{short} = 0.18$ or less (see table 2). Thus, the probability of using the final results originated from still "too short" simulation, when applying our rule of thumb for $R = 2$ is not larger than 0.032. It becomes not larger than 0.006 for $R = 3$, and drops to 0.0002 if one repeats sequential simulation $R = 5$ times.

A similar approach, although in a different context, was proposed by D. Knuth in 1969, when he wrote that "*... the most prudent policy for a person to follow is to run each Monte Carlo program at least twice, using quite different sources of pseudo-random numbers, before taking the answers of the program seriously.*"(Knuth 1969).

Such a rule of thumb could be easy implemented in commercial simulation packages offering automated control of sequential simulation. These include, for example, QNAP2*, Prophesy** and whole family of simulation packages based on SIMSCRIPT II.5†. There exist also such packages offered as free-ware for non-profit research organisations. One of them is Akaroa-2 (Ewing et al. 1999), designed at the University of Canterbury, in Christchurch, New Zealand.

CONCLUSIONS

No rule of thumb can ensure that the final confidence interval from a sequential stochastic simulation will contain the theoretical value with the probability equal to the assumed confidence level. One of the ongoing problems of the research in the area of sequential steady-state simulation is to find a method of simulation output data analysis which would be valid (in the sense of coverage) also when one applies it in simulation of highly dynamic stochastic processes. All methods whose coverage has been so far analysed evidently perform worse when they are applied in simulation of heavier loaded queueing systems and networks; see for example (Pawlikowski et al. 1998). Nevertheless, lowering the probability of using results from too short simulation runs by applying such a rule of thumb as the one formulated in this paper is one of very few possible practical ways available for simulation practitioners to improve the quality of results from their simulation experiments.

ACKNOWLEDGEMENTS

The work reported in this paper was partially supported by a grant from the University of Canterbury, Christchurch, New Zeland (Grant U6301).

*see http://www.simulog.fr
**see http://www.abstraction.com
†see http://www.caciasl.com

Table 2: Run-length statistics from 3,000 simulation replications: RS, $M/M/1/\infty$, theoretical confidence level = 0.95.

Load	*Num. of short runs*	*Coverage*	*Prob(short)*	*Threshold*	*Mean of lengths*
0.1	199	12.1%	6.6%	306	458
0.2	239	19.7%	8.0%	364	588
0.3	258	18.6%	8.6%	440	769
0.4	264	20.5%	8.8%	553	1021
0.5	307	21.5%	10.2%	735	1362
0.6	314	15.0%	10.5%	985	1891
0.7	330	15.2%	11.0%	1394	2836
0.8	379	5.3%	12.6%	1962	4568
0.9	539	5.6%	18.0%	3233	9378

Table 3: Run-length statistics from 3,000 simulation replications: SA, $M/M/1/\infty$, theoretical confidence level = 0.95.

Load	*Num. of short runs*	*Coverage*	*Prob(short)*	*Threshold*	*Mean of lengths*
0.1	0	N/A	N/A	1341	1723
0.2	0	N/A	N/A	1377	2002
0.3	133	83.5%	4.4%	1538	2475
0.4	500	80.2%	16.7%	1813	3278
0.5	297	69.7%	9.9%	2383	4670
0.6	364	63.7%	12.1%	3415	7247
0.7	307	53.1%	10.2%	5214	12727
0.8	236	46.2%	7.9%	9743	27906
0.9	263	38.0%	8.8%	33461	107049

REFERENCES

Ewing, G.; K. Pawlikowski; and D. McNickle. 1999. "Akaroa 2: Exploiting Network Computing by Distributed Stochastic Simulation." *13th European Simulation Multiconference* (Warsaw, Poland, June 1999). SCS.

Kleijnen, J. P. C. 1979. "The Role of Statistical Methodology in Simulation." In *Methodology in Systems Modelling and Simulation*, B.P.Zeigler et al. (eds), North-Holland, Amsterdam

Knuth, D. E. 1969. *Art of Programming*, vol. 2, Addison-Wesley.

Law, A. M. and W. D. Kelton. 1982. "Confidence Intervals for Steady-State Simulations, II: A Survey of Sequential Procedures." *Management Science*, 28, no.5 (May): 550-562.

Law, A. M. and W. D. Kelton. 1991. *Simulation Modelling and Analysis.*, McGraw-Hill

Lee, J.-S. R. 1999. "Coverage Analysis of Sequential Regenerative Simulation." Technical Report TR-COSC 02/99. Department of Computer Science, University of Canterbury, Christchurch, New Zealand.

Lee, J.-S. R.; D. McNickle; and K.Pawlikowski. 1999. "Quality of Sequential Regenerative Simulation." *13th European Simulation Multiconference* (Warsaw, Poland, June 1999). SCS.

Pawlikowski, K. 1990. "Steady-State Simulation of Queueing Processes: A Survey of Problems and Solutions." *ACM Computing Surveys*, 2: 123-170.

Pawlikowski, K. and L. de Vere. 1993. "Speeding up Sequential Simulation by Relative Variance Reduction." In *Proc. 8th Australian Teletraffic Research Seminar, ATRS'93*, RMIT, Telstra, 203-212.

Pawlikowski, K.; D. McNickle; and G. Ewing. 1998. "Coverage of Confidence Intervals in Sequential Steady-State Simulation." *Simulation Practice and Theory*, 6, no.2: 255-267.

Schriber, T. J. and R. W. Andrews. 1981. "A Conceptual Framework for Research in the Analysis of Simulation Output." *Communications of the ACM*, 24, no.4 (April): 218-232.

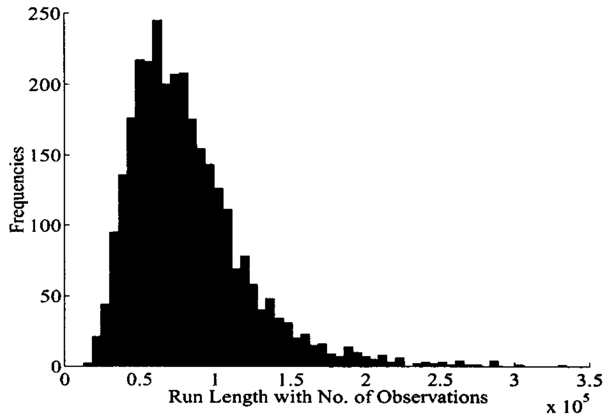

Figure 4: Histogram of run lengths for sequential NOBM (M/M/1/∞, load = 0.9)

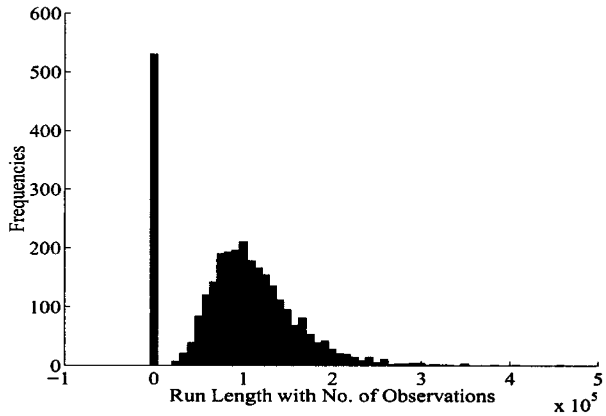

Figure 5: Histogram of run lengths for sequential RS (M/M/1/∞, load = 0.9)

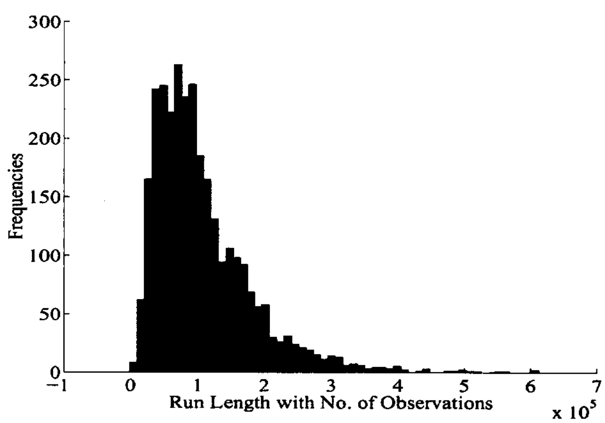

Figure 6: Histogram of run lengths for sequential SA (M/M/1/∞, load = 0.9)

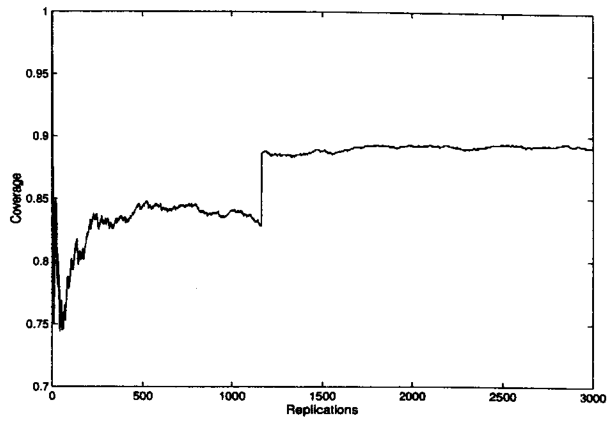

Figure 7: Coverage for sequential NOBM (M/M/1/∞, load = 0.9)

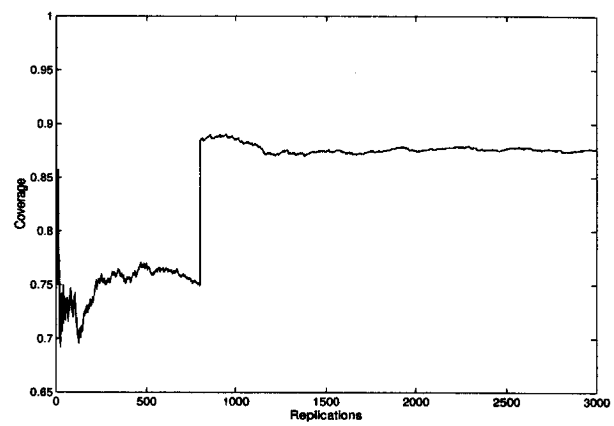

Figure 8: Coverage for sequential RS (M/M/1/∞, load = 0.9)

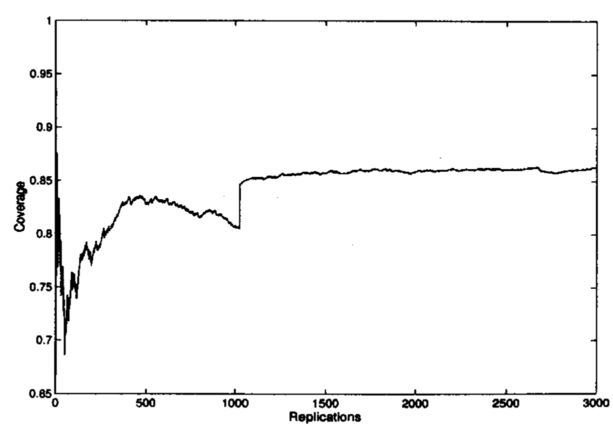

Figure 9: Coverage for sequential SA (M/M/1/∞, load = 0.9)