Quantile Estimation in Sequential Steady-State Simulation

J.-S. R. Lee[†], D. McNickle[‡] and K. Pawlikowski[†]

Department of [†]Computer Science and [‡]Management
University of Canterbury, Christchurch, New Zealand
e-mail: {ruth, krys}@cosc.canterbury.ac.nz
Ph.: +(64) 3 364 2362, Fax.: +(64) 3 364 2569

Abstract

Performance of such stochastic dynamic systems as computer and telecommunication networks is often assessed by quantiles of, for example, system's response times or experienced delays. Thus, ability of estimation of quantiles during stochastic simulation of computer and telecommunication networks is an important practical issue. However, unlike estimation of mean values, estimation of quantiles requires that simulation output data (observations) have to be stored for processing in several steps. Thus, QE (quantile estimation) requires large amounts of computer storage as well as longer computation time than for example, estimation of mean values. Several approaches for estimating quantiles in RS (regenerative simulation) and non-RS have been proposed to weaken this problem. In this paper, we investigate three approaches, known as: *linear* QE, *batching* QE and *spectral* P^2 QE, and study their quality in the context of sequential steady-state simulation. Numerical results of coverage analysis of quantile estimators used in these three approaches to QE are presented.

Keywords : sequential steady-state simulation, quantile estimations, coverage analysis

1 Introduction

When simulating a dynamic stochastic system, such as a computer or telecommunication network, the simulator is frequently more concerned with the extreme performance of the simulated system characterized by quantiles, than with its average behaviour. Quantiles are particularly useful for planning necessary capacities for various resources, comparing the overall performance of alternative designs or establishing minimum standards of performance. Therefore, from a practical point of view, the problem of reliable estimators of quantiles during stochastic simulation is quite important.

The pth quantile of a random variable X equals Q_p if $Pr[X \leq Q_p] = p$. Let $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, $x_{(i)} \geq 0$, be the ordered sequence of n observations of a random variable X, collected during simulation. The usual point estimator of the pth quantile is given by

$$\hat{Q}_p(n) = x_{(\lfloor n*p+1 \rfloor)}, \quad 0 < p < 1 \qquad (1)$$

where $\lfloor z \rfloor$ denotes the integral part of z.

The problem with using this formula is that, especially in the case of correlated sequences of observations, the length of sample sequence required for achieving adequate precision of $\hat{Q}_p(n)$ can be very large and impossible to predict in advance. Direct application of Equation (1) in QE (quantile estimation) requires large amounts of computer storage for storing the entire sequence of observations since sorting through the entire long sequence of output data is required whenever a new observation is recorded. For example, in steady-state simulation of M/M/1 queueing system, with traffic intensity ρ = 0.9, the estimation of 0.99 quantile of the waiting times requires roughly 500,000 observations, to achieve an estimate with the relative precision of at least 10%. For the 0.999 quantile, the required sample size is approximately 2,300,000 [17].

Clearly, storing and sorting the entire sequence

is impractical in such long runs. To produce a point estimate of Q_p requires storing the sequence and its ordering must be dynamically maintained as the sequence is expanding. Additional storage is also required to estimate the variance of $\hat{Q}_p(n)$ for determining the confidence interval for Q_p. Several approaches for avoiding these storing and sorting difficulties when estimating quantiles and their precision in RS (regenerative simulation) and non-RS have been proposed in [1], [5], [17], and [19]. However, these approaches were originally developed for traditional (non-sequential) procedures, even though sequential analysis of simulation output data is generally accepted as the only efficient way for achieving acceptable level of precision of the final results [3].

The most commonly used stopping rule of sequential stochastic simulation is based on the relative half-width of the confidence interval at a given confidence level, defined as the ratio

$$\epsilon = \frac{\Delta_{1-\alpha}}{\hat{\Theta}(n)}; \qquad 0 < \epsilon < 1,$$

and called the *relative precision*. Here $\hat{\Theta}(n)$ is the point estimate of unknown Θ from the sequence of n observations and $\Delta_{1-\alpha}$ is the half-width of the confidence interval for Θ at (1-α) confidence level, $0 < \alpha < 1$. In sequential estimation , the simulation experiment is stopped at the first checkpoint at which $\epsilon \leq \epsilon_{max}$, where ϵ_{max} is the required upper limit of relative precision of the final results at the 100(1-α)% confidence level, $0 < \epsilon_{max} < 1$.

It would be no problem with assessing statistical accuracy of the final results, if the sequence of observations $x_1, x_2, \ldots, x_n$ collected during simulation were realisations of identically and normally distributed random variables $X_1, X_2, \ldots, X_n$. However, in practical applications of simulation, collected observations are always (strongly) correlated. Thus, calculation of ϵ always involves different approximations or transformations of original data into secondary data, of easier analytically tractable properties.

Our motivation is to find the robust estimator of quantiles which could be used in practical applications of sequential steady-state simulation. In this paper we have investigated three approaches: *linear* and *batching* QE for RS, and *spectral* P^2 QE for non-RS, which do not require storing and sorting of the entire sequence of collected observations. The first two approaches, applicable in RS, are discussed in Section 2. In Section 3 *spectral* P^2 QE, a sequential QE for non-RS, is discussed. The numerical results of coverage analysis of these three sequential quantile estimators are presented in Section 4, while conclusions are in Section 5.

2 Sequential QE in RS

The RM (regenerative method) of simulation for analysis of regenerative processes, was first suggested by Cox and Smith [6]. This is motivated by the fact that many stochastic systems are regenerative, in the sense that they start afresh (in probabilistic sense) from time to time. The central idea of the RM is that, when $\{X(t) : t \geq 0\}$ is a regenerative process, random variables determined within successive regeneration cycles (RCs) are independent and identically distributed(i.i.d.). Thus, this can circumvent the problem of analysis of correlated observations.

Iglehart, Moore, and Seila ([2], [5]) have developed special methods of QE for processes with regenerative structures, i.e., processes for which there exist random time points at which they restart probabilistically afresh. An example is the waiting time process in M/M/1/∞ queueing system which regenerates every time a customer arrives to find the queue empty. Detailed comparisons of their approaches to QE in fixed-sample size stochastic simulations, i.e. simulations whose lengths have to be determined in advance, are given in [2].

Here, we consider two sequential quantile estimators for RS, based on Iglehart's and Seila's non-sequential proposals, assuming the sequential stopping rule based on the *relative precision*. We call them the *linear* and the *batching* method of QE, respectively. We do not adopt *Moore*'s estimator, because it has low memory and computing time efficiency. It requires $O(n\log n)$ computation time because all the observations have to be sorted, while the *linear* and the *batching* approaches to QE re-

quire $O(n)$ computation time as they do not need sorting. *Moore*'s approach also requires that all the observations are stored. This is in contrast to the *linear* and the *batching* approaches which require storing only aggregated data for four and two summary statistics, respectively.

2.1 Sequential QE using the *Linear* Approach

The *linear* approach has been originally developed for fixed-sample size regenerative simulation by Iglehart [5]. Here, we modify that approach for sequential QE. First, sequential QE requires specification of a grid of h+1 points $a_0, a_1, ..., a_h$, $a_0 < a_1 < ... < a_h$, so that all observations lie between a_0 and a_h. Next, to find a given quantile estimate, this method estimates the cumulative distribution function only at grid points, and use linear interpolation between the estimates. Simulation continues until the steady-state quantile has been estimated with the required relative precision, at the given confidence level.

Let us consider how the quantile Q_p of the random variable X would be estimated in the course of a simulation experiment if the simulated regenerative process has a discrete state space $\{0, 1, \cdots, M\}$. First, having simulated m RCs, we would accumulate the time spent in each of the states. If $w_m(i), i = 0, 1, \cdots, M$, is the total time spent in state i during m RCs, then the empirical distribution function of random variable X estimated after m RCs, would jump by $w_m(i)/\beta_m$ at state i, where β_m is the mth regeneration time. The sample quantiles after m RCs would then be estimated by

$$\hat{Q}_p(m) = \min\{k \geq 0 : w_m(0) + \cdots + w_m(k) \geq p\beta_m\}.$$

Let Y_i and a_i be the sum and the number of observations collected during the ith RC. Variance of $\hat{Q}_p(m)$ is then estimated as

$$\hat{\sigma}^2(\hat{Q}_p(m)) = \hat{\sigma}^2(Y_i) -$$

$$2F(\hat{Q}_p(m))cov(Y_i, a_i) + F^2(\hat{Q}_p(m))\hat{\sigma}^2(a_i),$$

and then linearly interpolated.

Here, $\hat{\sigma}^2(\cdot)$ and $cov(\cdot, \cdot)$ mean estimates of the variance and covariance, and $F(\cdot)$ is the empirical distribution function of random variable X after m RCs. $F(\hat{Q}_p(m))$ is estimated by $\sum_{k=0}^{i} w_m(k)/\beta_m$.

A $100(1-\alpha)\%$ confidence interval for the quantile Q_p is given by

$$\hat{Q}_p(m) \pm z_{1-\alpha/2}\hat{\sigma}(\hat{Q}_p(m))/(\bar{a}F'(\hat{Q}_p(m))m^{\frac{1}{2}})$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution, $F'(\hat{Q}_p(m))$ is estimated by $w_m(\hat{Q}_p(m)+1)/\beta_m$, and $\bar{a} = (1/m)\sum_{i=1}^{m} a_i$; see [5] for more details.

2.2 Sequential QE using the *Batching* Approach

The *batching* approach was also originally developed for fixed-sample size stochastic simulation only [1]. First, to adopt it to sequential QE, one needs to group observations from a number of RCs into batches and consider quantile estimates computed for the batches as independent and identically distributed secondary observations (secondary data). Before applying this method, we must select the batch size b (the number of RCs in a batch) to achieve a satisfactory data reduction. The *batching* method groups each b RCs in a batch and a quantile estimate is computed. Simulation stops when the relative precision of the analysed steady-state quantile reaches the required level of relative precision, at the given confidence level.

The *batching* method incorporates a two-fold jackknife procedure, in order to reduce bias of the quantile estimators. Assume that b is even, and let $\hat{Q}_p(b/2, 2i-1)$ and $\hat{Q}_p(b/2, 2i)$ be the estimates of Q_p computed from the first and second $b/2$ cycles in the ith batch using the ordinary quantile estimator. The jackknifed batch quantile is

$$J(\hat{Q}_p(b,i)) =$$

$$2\hat{Q}_p(b,i) - \frac{1}{2}(\hat{Q}_p(b/2, 2i-1) + \hat{Q}_p(b/2, 2i)).$$

The sequence $\{J(\hat{Q}_p(b,i)), i = 1, 2, ..., r\}$ over r batches consists of r i.i.d. random variables. Let $J(\bar{Q}_p(b,r))$ and $\hat{\sigma}^2(J(\bar{Q}_p(b,r)))$ denote the mean and variance of such jackknifed quantile estimators, i.e.,

$$J(\bar{Q}_p(b,r)) = \frac{1}{r}\sum_{i=1}^{r} J(\hat{Q}_p(b,i));$$

and

$$\hat{\sigma}^2(J(\bar{Q}_p(b,r))) =$$

$$\frac{1}{r-1}\sum_{i=1}^{r}(J(\hat{Q}_p(b,i)) - J(\bar{Q}_p(b,r)))^2.$$

Then, a $100(1-\alpha)\%$ confidence interval for Q_p is given by

$$J(\bar{Q}_p(b,r)) \pm t_{1-\alpha/2,r-1}\hat{\sigma}(J(\bar{Q}_p(b,r)))/\sqrt{r}$$

where $t_{1-\alpha/2,r-1}$ is the $(1-\alpha/2)$ quantile of the t distribution with $r-1$ degrees of freedom.

3 Sequential QE in Non-RS

Methods of QE for non-RS have been proposed by Heidelberger and Lewis [17] and Jain and Chlamtac [19]. Heidelberger and Lewis' QE method is based on an aggregation of data sequences, allowing to work with shorter sequences of (secondary) data. There is still the need for storing and sorting of the reduced sequences. Jain and Chlamtac's QE method is based on a P^2 (*Piecewise-Parabolic*) formula. Detailed algorithm and pseudocodes of P^2 algorithm are given in [19]. The P^2 algorithm solves the storage problem by allowing calculations of quantiles dynamically, as the observations are generated. The sequence of observations does not need to be stored. Instead, a few statistical counters are maintained which help refine the estimate. Therefore, QE using the P^2 algorithm has a very small storage requirement, regardless of the number of observations collected, and a small computing time because no sorting is required.

An *extended* P^2 method has been proposed by Raatikainen [12]. This *extended* P^2 method simultaneously estimates several quantiles without storing and sorting the observations, although it was still proposed for a fixed-sample size simulation. A sequential procedure for simultaneous estimation of several quantiles in non-RS has been proposed in [13]. This sequential version of the *extended* P^2 algorithm for estimating steady-state quantiles uses the *spectral analysis* of variance [18] of the quantile estimates. The procedure has not been fully automated since the length of initial transient stage of steady-state estimation had to be decided beforehand.

Here, we consider a fully sequential procedure for QE, which we will call *spectral* P^2. The length of initial transient stage is automatically determined, then steady-state estimation begins, and QE stops when the relative precision of the steady-state quantile reaches the required level of relative precision. It is based on the P^2 algorithm of Jain & Chlamtac [19] and on the SA/HW (Spectral Analysis in the version proposed by Heidelberger and Welch [18]) for estimating the variances of the quantile estimates. However, the length of the initial transient period is estimated sequentially, adopting one of the procedures proposed in [14].

Having collected n observations, the pth quantile Q_p estimated by the P^2 algorithm is actually calculated from the empirical distribution function approximated by $F(q) = cq^2 + dq + e$, at $q = \hat{Q}_p(n)$. As the number of observations become large, the variance of $\hat{Q}_p(n)$ can be approximated by

$$\sigma^2(\hat{Q}_p(n)) = S(0;\hat{Q}_p(n))/n\hat{f}(\hat{Q}_p(n))^2,$$

where $S(0;\hat{Q}_p(n))$ is the stationary spectral density at frequency 0, estimated using the SA/HW method applied to the sequence $\{I_i(\hat{Q}_p(n)), i = 1, \cdots, n\}$. Here, $I_i(\hat{Q}_p(n))$ is

$$I_i(\hat{Q}_p(n)) = \begin{cases} 1 & if \quad x_i \leq \hat{Q}_p(n), \\ 0 & otherwise. \end{cases}$$

$\hat{f}(\cdot)$ is the empirical density function which, in the neighbourhood of $\hat{Q}_p(n)$, is approximated by $\hat{f}(\hat{Q}_p(n)) = (b + 2a\hat{F}(\hat{Q}_p(n)))^{-1}$ [17].

A $100(1-\alpha)\%$ confidence interval for the quantile Q_p is given by

$$\hat{Q}_p(n) \pm t_{1-\alpha/2,n-1}\hat{\sigma}(\hat{Q}_p(n))/\sqrt{n}$$

where $t_{1-\alpha/2,n-1}$ is the $(1-\alpha/2)$ quantile of the t distribution with $n-1$ degrees of freedom.

4 Numerical Results

All three approaches to QE have been discussed in more detail in [10]. The robustness of any estimator can be measured experimentally by the coverage of its confidence intervals, defined as the proportion of the final confidence intervals containing

the true theoretical value of the estimated parameter. As justified in [4] and [15], coverage analysis of sequential estimators should be done sequentially, i.e. analysis of coverage should be stopped when the relative precision of the estimated coverage falls below an assumed level. Coverage analysis is limited to analytically tractable systems, since the theoretical values of the parameters of interest have to be known. In this paper we report our results obtained while estimating quantiles of response times of M/M/1/∞ queueing systems.

All numerical results refer to sequential steady-state simulation of M/M/1/∞ queueing systems stopped when the final steady-state estimate of the 0.9 quantile of the response time reached the required relative precision of 5% or less, at the 0.95 confidence level. We used 21 grid points spaced 0.2 units apart for the *linear* QE and batch sizes of 50 for the *batching* QE. Sequential coverage analysis was stopped with the relative precision of at least 5% at the 0.95 confidence level, and the estimates of coverage were based on the F distribution ([9] and [11]). Rules of sequential coverage analysis for non-RS were proposed in [15]. Adopting these rules to RS, the lengths of simulation runs were measured in the number of RCs. The results for coverage analysis were filtered from unusually short simulation runs after 200 bad confidence intervals (i.e., confidence intervals that do not cover theoretical value) have been collected. This number of observed bad confidence intervals has been recommended in [15], for ensuring representativeness of the analysed data.

To show the relevance of sequential coverage analysis of quantile estimators, we have also obtained results applying traditional fixed-sample size analysis of coverage. The results of fixed-sample size analysis (based on 200 independent replications) are depicted in Figure 1 and 3 for the *linear* and the *batching* QE, and in Figure 5 for the *spectral* P^2 QE. Figures 2, 4, and 6 depict the same results obtained by sequential coverage analysis.

Having compared the results presented in Figures 1 and 2, 3 and 4, 5 and 6, one can see importance of appropriate method of coverage analysis. Note, that sequential analysis of coverage not only leads to more accurate results, but it also allows for full automation of such, very tedious comparative studies of properties of different estimators. The total numbers of independent replications generated in each case of our sequential coverage analysis are given in Table 1.

Comparing the results of Figures 2, 4 and 6, one can see that the best quality of the final results in the sense of coverage can be achieved by applying the batching QE. On the other hand, none of these three approaches seems to lead to satisfactorily accurate results in analysis of very dynamic queueing processes, since the coverage of the final results for M/M/1/∞ significantly drops down for loads exceeding 0.8, even when the batching approach is used. From the practical point of view, it should be desirable to speed up these methods of QE by distributed simulation, for example, by applying MRIP (Multiple Replications In Parallel) scenario of stochastic simulation, suggested in [16] and implemented in a simulation package Akaroa2 [7]. This is the current direction of our research.

5 Conclusions

The direct estimation of quantiles, based on storing and consecutive multiple sorting of entire sequences of observations collected during sequential simulation appears to be impractical in real applications of stochastic simulation. In this paper we have studied properties of three sequential quantile estimators, applicable in RS and non-RS. As our initial results show, only one of these estimators, based on the batching approach in RS, offers a good quality of the final results, providing that the simulated processes are not too strongly autocorrelated (in M/M/1/∞ queueing systems, the degree of autocorrelation in simulation output data increases with the system's load). Currently we are looking at different solutions of this problem, to make the batching approach to QE more robust. This includes an adoption of this approach to the MRIP scenario of stochastic simulation, as implemented in Akaroa2 [8]. Additionally, before any method of QE is implemented in this simulation tool for automated stochastic simulation, its quantity needs to be tested on broader spectrum of basic, analytically tractable simulation models of standard reference systems.

References

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

[2] A. F. Seila. Estimation of Percentiles in Discrete Event Simulation. *SIMULATION*, 39(6):193–200, 1982.

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

[4] D. C. McNickle and K. Pawlikowski and G. Ewing. Experimental Evaluation of Confidence Interval Procedures in Sequential Steady-State Simulation. In *Proceedings of the 1996 Winter Simulation Conference*, pages 382–389, 1996.

[5] D. L. Iglehart. Simulating Stable Stochastic Systems, VI: Quantile Estimation. *Journal of the Association for Computing Machinery*, 23(2):347–360, 1976.

[6] D. R. Cox and W. L. Smith. *Queues*. Methuen, London, 1961.

[7] G. C. Ewing and K. Pawlikowski and D. McNickle. Akaroa 2.5 User's Manual. Technical Report TR-COSC 07/98, Department of Computer Science, University of Canterbury, Christchurch, New Zealand, 1998.

[8] G. C. Ewing and K. Pawlikowski and D. McNickle. Akaroa2: Exploiting Network Computing by Distributing Stochastic Simulation. Technical Report TR-COSC 09/98, Department of Computer Science, University of Canterbury, Christchurch, New Zealand, 1998. (Submitted for 13th ESM'99).

[9] J. R. Lee and D. McNickle and K. Pawlikowski. A Survey of Confidence Interval Formulae for Coverage Analysis. Technical Report TR-COSC 04/98, Department of Computer Science, University of Canterbury, Christchurch, New Zealand, 1998.

[10] J. R. Lee and D. McNickle and K. Pawlikowski. Sequential Estimation of Quantiles. Technical Report TR-COSC 05/98, Department of Computer Science, University of Canterbury, Christchurch, New Zealand, 1998.

[11] J. R. Lee and D. McNickle and K. Pawlikowski. Confidence Interval Estimators for Coverage Analysis in Sequential Steady-State Simulation. In *Proceedings of the Twenty Second Australasian Computer Science Conference, Auckland*, (In Press), 1999.

[12] K. E. E. Raatikainen. Simultaneous Estimation of Several Percentiles. *SIMULATION*, 49(4):159–164, 1987.

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

[14] K. Pawlikowski. Steady-State Simulation of Queueing Processes: A Survey of Problems and Solutions. *ACM Computing Surveys*, 22(2):122–170, 1990.

[15] K. Pawlikowski and D. C. McNickle and G. Ewing. Coverage of Confidence Intervals in Sequential Steady-State Simulation. *Simulation Practice and Theory*, 6:255–267, 1998.

[16] K. Pawlikowski and V. Yau and D. C. McNickle. Distributed and Stochastic Discrete-event Simulation in Parallel Time Streams. *Proceedings of the 1994 Winter Simulation Conference*, pages 723–730, 1994.

[17] P. Heidelberger and P. A. W. Lewis. Quantile Estimation in Dependent Sequences. *Operations Research*, 32(1):185–209, 1984.

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

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

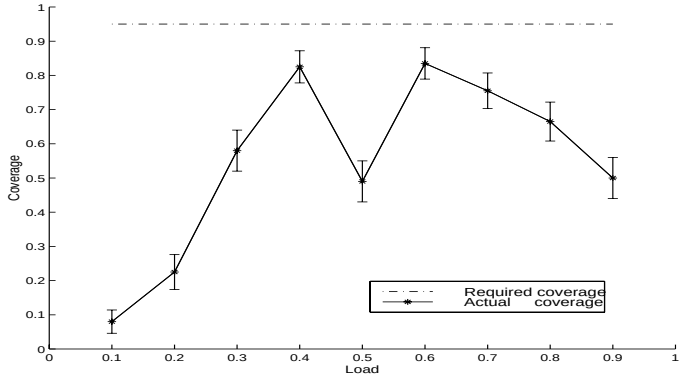

Figure 1: Fixed-sample size coverage analysis of the *linear* QE in RS (M/M/1/∞, 200 replications)

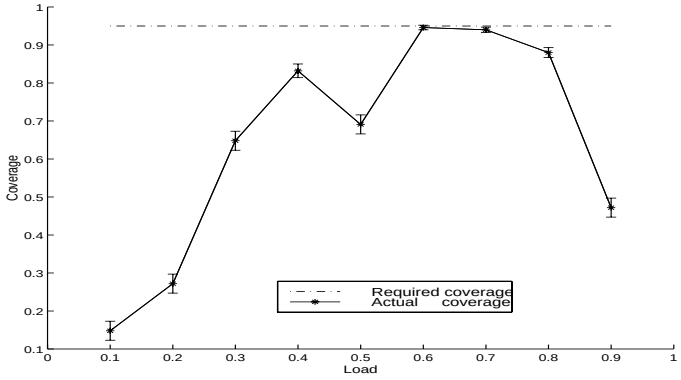

Figure 2: Sequential coverage analysis of the *linear* QE in RS (M/M/1/∞)

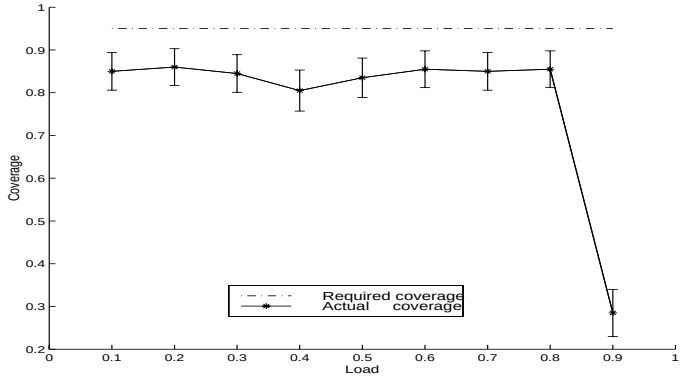

Figure 3: Fixed-sample size coverage analysis of the *batching* QE in RS (M/M/1/∞, 200 replications)

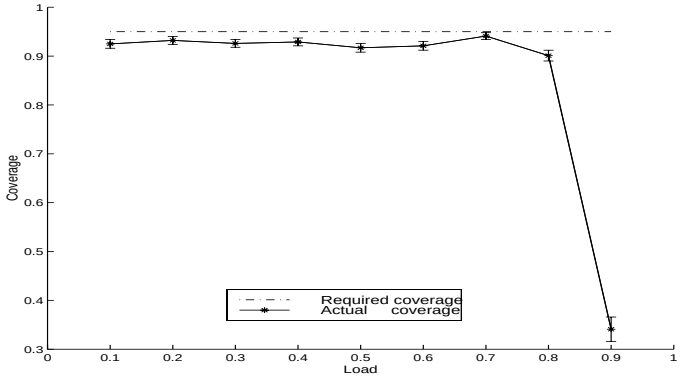

Figure 4: Sequential coverage analysis of the *batching* QE in RS (M/M/1/∞)

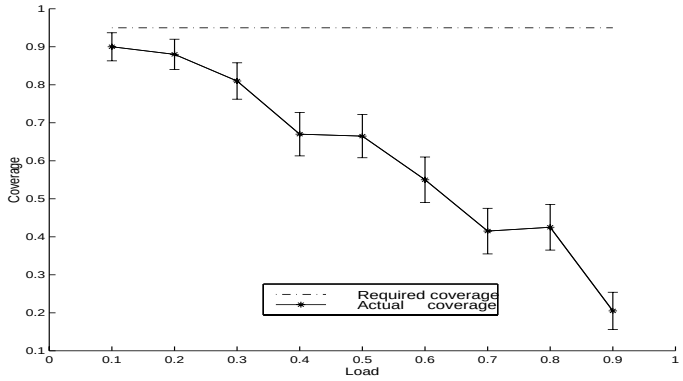

Figure 5: Fixed-sample size coverage analysis of the *spectral* P^2 QE in non-RS (M/M/1/∞, 200 replications)

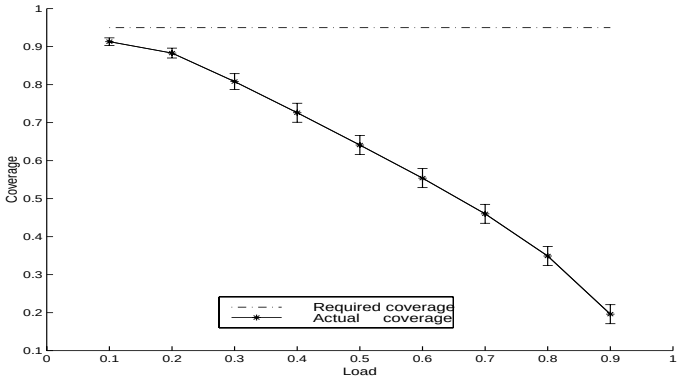

Figure 6: Sequential coverage analysis of the *spectral* P^2 QE in non-RS (M/M/1/∞)

Table 1: Total number of independent simulation replications in sequential coverage analysis of three sequential QE approaches (M/M/1/∞ queueing system)

Load	*Linear*	*Batching*	*Spectral* P^2
0.1	1153	3116	2378
0.2	1660	3477	1711
0.3	1709	3147	1044
0.4	1188	3279	899
0.5	1776	2772	1034
0.6	4458	2884	1108
0.7	4186	3923	1114
0.8	2355	2294	1022
0.9	1527	1527	723