Simulation Modelling Practice and Theory 18 (2010) 177–189

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

journal homepage: www.elsevier.com/locate/simpat

Some effects of transient deletion on sequential steady-state simulation

Don McNickle [a,*], Gregory C. Ewing [b], Krzysztof Pawlikowski [b]

[a] Department of Management, University of Canterbury, Christchurch, New Zealand
[b] Department of Computer Science, University of Canterbury, Christchurch, New Zealand

ARTICLE INFO

Article history:
Received 4 February 2009
Received in revised form 5 October 2009
Accepted 7 October 2009
Available online 22 October 2009

Keywords:
Discrete event simulation
Steady-state simulation
Sequential simulation
Transient deletion

ABSTRACT

In discrete event steady-state simulation, deleting the initial transient phase of the simulation is usually recommended in order to reduce bias in the results. Various heuristics and tests have been proposed to determine how many observations to delete. The plummeting cost of simulation, combined with uncertainties about the overall reliability of transient methods, suggests revisiting the notion that deletion is essential. We consider this in a framework of sequential simulation, where the simulation is run until a pre-specified accuracy of the results is reached. Our results show that for run lengths required for commonly used levels of accuracy, there is no substantial difference in point or interval estimates of means due to deleting the initial transient for the models we consider. However, in sequential simulation, deleting the initial transient turns out to have considerable value in reducing the risk that the simulation stops too early, thus ensuring that the accuracy of the final results is closer to that specified by the decision-maker.

1. Introduction

A standard part of simulation methodology for discrete event non-regenerative steady-state simulation is that data from the initial transient phase of simulation should be deleted in order to reduce the bias in the final estimates. The underlying assumption is that the distribution of the process being simulated may be changing over the transient phase (see, for example [13, p. 488]), and thus including data from the transient phase would introduce bias in the results.

A large number of methods for selecting the number of observations to delete, and for testing if the system is adequately close to "steady state", have been proposed. Hoad et al. [11] list 42 methods, which they broadly classify as graphical methods, where the convergence of the process towards steady state is examined visually; heuristic methods; and those based on statistical testing. Some of these proposals appear to have had limited testing, so their validity remains in question. It is also noticeable that these methods have had little impact on commercial simulation packages, which usually only offer a user-specified (fixed) deletion period.

With the steep decline in the cost of computing, the availability of large-scale computing resources via networks and the web, and simulation software that can carry out multiple replications in parallel, such as Akaroa2, [3] (available at http://www.akaroa2.canterbury.ac.nz/) it is now possible to collect large amounts of simulation output data in acceptable time and at acceptable cost. Since the initial transient may now form a very small fraction of the total run, is it true that the influence of the initial state of the simulated system is quite limited? Given the uncertainty about the overall performance of some of the deletion methods, has the balance shifted back in favour of not deleting observations i.e. a "brute force" approach? Or are there other problems which removal of the transient helps to control?

* Corresponding author. Tel.: +64 3 3642666; fax: +64 3 3642020.
E-mail address: Don.McNickle@canterbury.ac.nz (D. McNickle).

1569-190X/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.simpat.2009.10.004

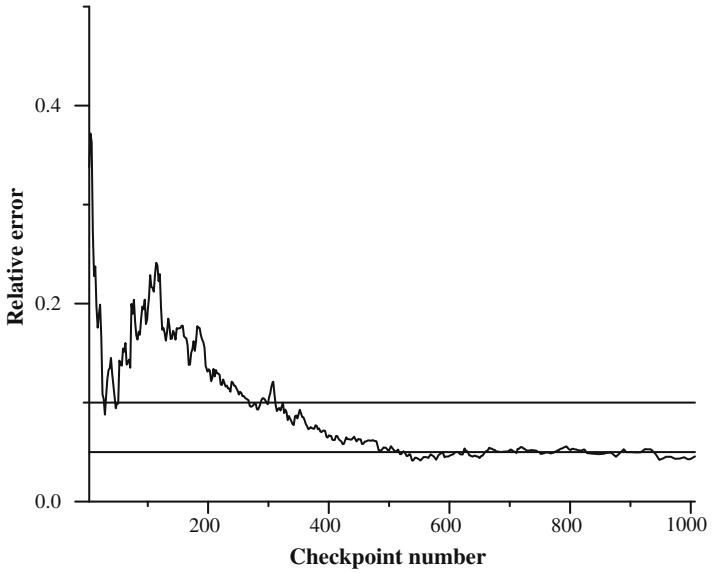

Fig. 1. Convergence of estimated relative error.

Sequential analysis of steady-state simulation output provides an attractive experimental framework for decision-makers [13,18]. Rather than specifying a simulation run length in advance, only the statistical accuracy of the required results needs to be specified. The simulation runs until this criterion is apparently satisfied.

Sequential simulation has the problem that some of the simulation experiments may stop with an insufficient number of observations because, by chance, the required accuracy is apparently temporarily attained. Fig. 1 shows a plot of the convergence of the estimated relative error (relative precision) for a single simulation experiment on an M/M/1 queue with a load of $\rho = 0.9$. The 10% relative error criterion is temporarily satisfied twice before the 100th checkpoint, whereas in fact about 250 checkpoints must elapse before that degree of accuracy is obtained.

As we shall show, the major effect of not deleting the initial transient is that, rather than having a substantial effect on the bias of the point estimate, it reduces the average run lengths at which the stopping criterion is apparently satisfied. Given that sequential simulation already has this problem of premature stopping, this argues strongly for the use of a simple and reliable transient deletion method. As an aside, we note that Lee et al. [14] give some practical heuristics that can guard against runs that are too short. However, in this paper the effect of discarding the initial transient data is considered without applying these heuristics. Thus the coverage results in this paper appear worse than we can achieve in practice.

We set out to answer a simple question: at the levels of accuracy that are usually specified in sequential simulation, can the effects of deleting or not deleting the initial transient data be detected, and what are these effects?

Heidelberger and Welch [10] addressed a similar question when they proposed a unified approach for transient detection and run length control, based on spectral analysis. The differences here are: we are using a simple method for detecting the initial transient, thus avoiding possible correlation between the transient method and run length control; we concentrate on coverage obtained in a practical experimental framework, rather than having run-length/coverage as just one of the outputs; and we are using more replications for our coverage estimates (typically about 10,000 rather than 100). While we come to the same conclusion, that a transient deletion method is needed, we come up with different, and we hope clearer, reasons why it is needed in sequential simulation, and what the effects of it are.

2. Methodology

In sequential simulation the simulation stops when a pre-specified level of statistical accuracy of the results is apparently reached. A common stopping method is to specify the *relative precision* of the estimate. This uses the ratio:

$$\varepsilon(N) = \frac{\Delta_{1-\alpha}(N)}{\hat{\theta}(N)},$$

where $\Delta_{1-\alpha}(N)$ is the half-width of the confidence interval at the $1-\alpha$ confidence level for the estimate $\hat{\theta}(N)$ of the required parameter θ after N observations, i.e.:

$$\Delta_{1-\alpha}(N) = t_{df,1-\alpha/2}\sigma(\hat{\theta}(N)),$$

where $t_{df,1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the t-distribution, and df is the degrees of freedom implied by the method used to estimate the variance of the estimate $\hat{\theta}(N), \sigma^2(\hat{\theta}(N))$. If, for example, the stopping criterion is that the simulation stops when

a relative precision of 10% (0.1) has been reached, the simulation will stop when $\varepsilon(N) \leqslant 0.1$ for the first time. An alternative stoping method is that of *absolute precision* which looks at the half-width of the confidence interval, $\Delta_{1-\alpha}(N)$, only. Relative precision is usually preferred as the value, or even order of magnitude, of the parameter θ is usually not known before the simulation starts. To save computational effort and time in sequential simulation the precision is usually calculated only at specified intervals ("checkpoints") as the sequential simulation progresses.

We will concentrate on coverage analysis of the estimated confidence intervals. Coverage analysis picks up biases in both point and interval estimates and goes directly to measuring the quality of the results that the decision-maker can expect with simulation. For example, if supposedly 95% confidence intervals of a specified relative precision (say 10%) are being used as the stopping criterion, then what size confidence intervals are we actually getting? Are they actually 90% confidence intervals? We use independent replications to measure the fraction of estimated confidence intervals that actually contain the true value of the parameter of interest. While we will comment briefly on bias of the final point estimates, any bias due to the initial transient could also be expected to be observed in reduced actual coverage.

The experiments were run using the Akaroa2 simulation package, using it in its single-processor mode. Thus each of the approximately half a million simulation experiments required for this paper can be considered as being carried out on a separate processor, with an independent stream of random numbers. An automated method was used to determine the length of the initial transient period. This first uses a heuristic proposed by Gafarian et al. [5] to decide when to start testing for stationarity. Its use in a sequential context is described in detail in Pawlikowski [18]. In this heuristic, the length of initial transient period is first taken to be over when the sequence has crossed its running mean 25 times. Then a sequential version of Schruben's test [20,21,7] is used to test for stationarity. If the null hypothesis of stationarity is rejected, the length of the potential transient period is doubled and the test repeated [18]. Comparisons with a limited range of other transient deletion methods can be found in [6,17], which show that this method, although simple, does appear to work well, at least for basic queueing models. The method usually picks short transient periods. The use of a conservative, simple method turns out to suit our conclusions well.

Sequential spectral analysis, a modification of the method proposed by Heidelberger and Welch [9] and specified in [18], was used to estimate the confidence interval width. We have found that this method gives accurate confidence intervals, especially for highly correlated data, such as waiting times in highly loaded queues [4,16]. Since spectral analysis is not as well known a confidence interval estimation technique as, for example, Batch Means, all the models below were also run using the automated version of the Batch Means technique in Akaroa2 [18]. These results were entirely compatible with our conclusions.

For this study a further automated sequential framework, for estimating coverage, was essential as producing just one of the estimates of coverage involved up to 30,000 independent replications, each using thousands of observations. This framework used the following two rules: Since we are estimating large binomial probabilities, coverage estimation starts only after a minimum number (say 200) of "bad" confidence intervals (confidence intervals not containing the actual value of the parameter θ) have been recorded. Then analysis of coverage continues until the absolute precision (half-width of the confidence interval) of the estimated coverage reaches a specified level, say 0.005. Further details of this coverage estimation methodology are described in [19].

Experiments were conducted for waiting times in a representative range of queueing models: M/M/1, M/D/1, and $M/H_2/1$ with a coefficient of variation of the service times set to $\sqrt{5}$. We start the experiments from the empty and idle state, since that is the setting most often used in practical simulation.

3. Results

The experiments were replicated until a 95% confidence interval for the estimated coverage was reduced to a half-width of 0.005. For the worst case ($M/H_2/1$, with a load of 0.9 and relative precision of 5%, and the transient not deleted) this involved 14,464 replications, each involving about 2,000,000 observations. Separate experiments were run for the case of deletion, and of no deletion of the transient data. The experiments were conducted as paired comparisons, with the same starting position for the random number stream used for both a deletion and no deletion replication. This can be seen in the way the deletion and no deletion lines mimic each other in each of the graphs in Figs. 2 and 7.

3.1. Bias

We first wish to show that for the run lengths required for relative precisions of 5% or 10%, the effect of deleting or not deleting the initial transient on the estimated mean waiting time is very small indeed. For one of the models, M/M/1, this can be done theoretically. We can calculate the expected waiting time of the Nth arriving customer, for $N = 1, 2, \ldots, T$, and hence the average waiting time over all the observations in the transient period, using the Markov chain technique described in Kelton and Law [12]. Table 1 shows the results of including these observations in the estimated mean waiting time – i.e. not deleting them.

The values in Table 1 were calculated as follows. We take the length of the initial transient period to be T, the average number of observations deleted in the experiments. For a queue with a load of 0.9 this was observed to be 726 arrivals. By Kelton and Law's method, the average mean waiting time of the first 726 arrivals is 7.8132. After the

Table 1
Theoretical effect of not deleting the initial transient for the M/M/1 model.

ρ	Average transient period (T)	Average number of observations (N_{Del})	Average mean waiting time of the first T arrivals	Average mean waiting time of the first T plus N_{Del} arrivals	Fraction of the steady-state mean waiting time (%)
0.1	367	50,050	.1110	.1111	100.0000
0.2	321	37,231	.2496	.25	100.0000
0.3	309	35,936	.4264	.4286	100.0000
0.4	311	39,309	.6629	.6666	100.0000
0.5	321	47,281	.9907	.9999	99.9999
0.6	344	62,596	1.4771	1.4999	99.9999
0.7	386	95,886	2.2722	2.3330	99.9999
0.8	475	187,359	3.7979	3.9995	99.9872
0.9	726	654,449	7.8132	8.9987	99.9854

Table 2
Bias due to not deleting the initial transient. M/D/1 with 10% relative precision.

ρ	Number of replications	Mean waiting time with deletion	Mean waiting time without deletion	Difference	Relative difference	t-Value	P
0.1	9636	.05582	.05573	.00009	.0016	2.1	.036
0.2	8612	.12479	.12473	.00006	.0004	.89	.376
0.3	8834	.21361	.21325	.00036	.0017	3.94	$<10^{-3}$
0.4	9253	.33216	.33144	.00072	.0022	5.51	$<10^{-3}$
0.5	9491	.49795	.49663	.00132	.0027	6.88	$<10^{-3}$
0.6	10,010	.74586	.74347	.00239	.0032	8.64	$<10^{-3}$
0.7	10,585	1.15768	1.15390	.00378	.0032	9.36	$<10^{-3}$
0.8	10,979	1.98079	1.97449	.00299	.0015	9.68	$<10^{-3}$
0.9	12,716	4.44501	4.42805	.00630	.0014	13.38	$<10^{-3}$

Table 3
Bias due to not deleting the initial transient. M/H_2/1 with 5% relative precision.

ρ	Number of replications	Mean waiting time with deletion	Mean waiting time without deletion	Difference	Relative difference	t-Value	P
0.1	10,926	.333195	.333211	−.000016	−.00005	−.42	.676
0.2	11,389	.749427	.749290	.000137	.00018	1.70	.089
0.3	10,979	1.28471	1.28459	.000128	.0001	.85	.395
0.4	11,270	1.99699	1.99674	.000245	.00012	1.09	.277
0.5	11,110	2.99509	2.99435	.000735	.00025	2.21	.027
0.6	11,038	4.49255	4.49146	.001096	.00024	2.19	.028
0.7	11,860	6.98251	6.98161	.000902	.00013	1.16	.247
0.8	11,348	11.9678	11.0659	.00193	.00016	1.45	.147
0.9	11,171	26.9159	26.9084	.00745	.00028	2.66	.008

transient period, we collect an average of $N_{Del} = 65449$ observations. The subscript *Del* indicates that this value comes from experiments where data from the transient period has been deleted. If we assume that these are steady-state observations, the average waiting time estimated over all 726 + 654,449 observations will be 8.9987, resulting in a difference of less than .015%.

So the theoretical analysis suggests that there is little bias in the mean waiting time detectable at these run lengths.

Now from the experimental data we compare the observed mean waiting times with deletion of transient data, with those where the transient results have been included. To save space we give two sets of results only, from opposite ends of the range of models considered, for M/D/1 with 10% relative precision and M/H_2/1 with 5% relative precision. When analysed with the usual t-test for differences of means, there were no significant differences between the mean waiting times at all, in spite of the large numbers of replications. Some significant differences were obtained when the paired comparisons approach was exploited, and the t- and P-values reported in Tables 2 and 3 are for paired-comparisons t-tests. For these, replications had to be discarded from the larger of the two data sets. (We note that, for all cases studied, this was the set without deletion. So it takes more replications to estimate the relative precision to a given accuracy if the initial transient is not discarded.)

Tables 2 and 3, and the other results not listed here, show that there is a small bias when the data from the initial transient is included, (columns 5 and 6). Not surprisingly the bias tends to decrease with longer runs (either resulting from more variable models, or smaller required relative precision.) The mean waiting time estimated without removing the initial

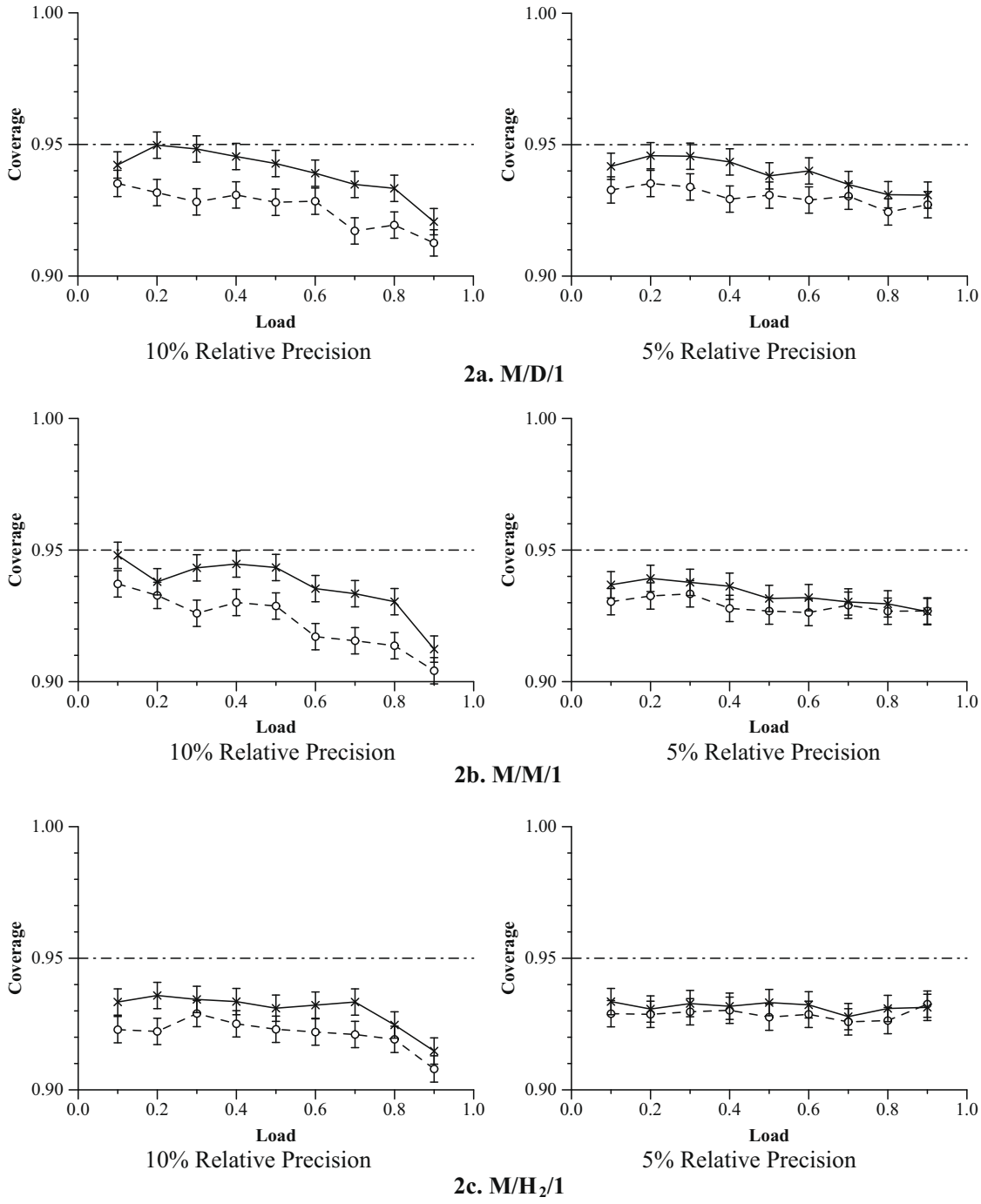

Fig. 2. Achieved coverage with sequential stopping times.

transient is slightly smaller than that produced if a transient period is deleted, which is not surprising considering that the simulations start from an empty and idle state. For 5% relative precision, the bias is very small and when measured in terms of a *t*-test the results are almost all not statistically significant.

3.2. Coverage analysis

We now turn to comparing the coverages when data from the transient period is deleted, to those obtained when it is not. The graphs in Fig. 2 plot the average coverage (together with 95% confidence intervals) from simulating the waiting times in the models specified, with deletion of transient data (solid lines) or no deletion (dashed lines). Note that the confidence intervals for average coverage all have the same half-width, 0.005. So the graphs show the actual coverage that was achieved when the required coverage was set to that of a 95% (0.95) confidence interval having a relative half-width (relative precision) of either 10% (left-hand graphs) or 5% (right-hand graphs).

From the graphs in Fig. 2 it can be seen that deletion of the initial transient does appear to produce some effect on coverage, especially for M/D/1, where the run lengths are short. The effect reduces with: the variability of the model, and the accuracy of the precision, to the point where in a single experiment the reduction in coverage would no longer be deemed to be large.

Thus for $M/H_2/1$ at 5% precision (the longest runs) the effects are very small. In all cases the no-deletion coverage increases towards that produced when initial transient data are deleted, as the run length increases.

However, the important point is that almost all the differences in Fig. 2 can be explained by the fact that the average run lengths without deletion are uniformly shorter than those with deletion. It might be thought that unless the transient period is deleted, a sequential simulation will run substantially longer on the average than a simulation with deletion, due to the bias produced by inclusion of the transient results. But for all the models in this study it turns out that the opposite of this is true.

Tables 4–6 give the average run characteristics over the 10,000 to 14,000 replications, for 5% relative precision. Those for 10% relative precision (not included here) show similar effects. The subscripts *Del* and *NoDel* refer to the measurements

Table 4
Run lengths and coverages for M/D/1 (5% relative precision).

	Deletion		No deletion				
ρ	Average number of observations (N_{Del})	Coverage (p_{Del})	Average number of observations (N_{NoDel})	Coverage (p_{NoDel})	Difference in average no. of observations	Coverage (p'_{NoDel}) corrected for difference in no. of observations	Difference $p'_{NoDel} - p_{Del}$
0.1	30051	.942	26,187	.933	3864	.950	.009
0.2	20363	.946	17,483	.935	2880	.954	.011
0.3	18539	.946	15,948	.934	2591	.952	.012
0.4	19507	.943	16,808	.929	2699	.948	.013
0.5	23061	.938	20,123	.931	2938	.948	.007
0.6	30337	.940	27,036	.929	3301	.944	.011
0.7	46368	.935	42,360	.930	4008	.942	.005
0.8	89971	.931	85,376	.924	4595	.932	.017
0.9	318054	.931	313,712	.927	4342	.929	.004

Table 5
Run lengths and coverages for M/M/1 (5% relative precision).

	Deletion		No deletion				
ρ	Average number of observations (N_{Del})	Coverage (p_{Del})	Average number of observations (N_{NoDel})	Coverage (p_{NoDel})	Difference in average no. of observations	Coverage (p'_{NoDel}) corrected for difference in no. of observations	Difference $p'_{NoDel} - p_{Del}$
0.1	50050	.937	45,667	.930	4383	.942	.005
0.2	37231	.939	33,502	.933	3729	.946	.006
0.3	35936	.938	32,385	.933	3551	.946	.008
0.4	39309	.936	35,623	.928	3686	.942	.006
0.5	47281	.932	43,529	.927	4022	.938	.006
0.6	62596	.932	58,372	.926	4224	.936	.006
0.7	95886	.930	90,900	.929	4896	.936	.006
0.8	187359	.930	181,939	.927	5420	.931	.001
0.9	654449	.927	644,927	.927	8522	.929	.002

Table 6
Run lengths and coverages for $M/H_2/1$ (5% relative precision).

	Deletion		No deletion				
ρ	Average number of observations (N_{Del})	Coverage (p_{Del})	Average number of observations (N_{NoDel})	Coverage (p_{NoDel})	Difference in average no. of observations	Coverage (p'_{NoDel}) corrected for difference in no. of observations	Difference $p'_{NoDel} - p_{Del}$
0.1	152500	.934	147,897	.929	4603	.934	0
0.2	138604	.931	134,369	.929	4235	.933	.002
0.3	146828	.933	142,328	.930	4500	.934	.001
0.4	164688	.932	160,280	.930	4408	.934	.002
0.5	198206	.933	192,687	.928	5519	.932	−.001
0.6	258459	.932	252,428	.929	6031	.932	0
0.7	379124	.928	370,704	.926	8510	.929	.001
0.8	692801	.931	681,722	.926	11,079	.928	−.003
0.9	2223173	.931	2,206,309	.933	16,864	.934	.003

where data from the transient period has been deleted, and those where it has not, respectively. N_{Del} is the average number of observations collected after those in the transient period have been discarded.

We note from Tables 4–6 that the average run lengths without deletion (column 4) are always shorter than those where the initial transient data are deleted (column 2), by a reasonably fixed amount (column 6) that tends to increase with load and variability of the model.

We wish to show that almost all the differences between the pairs of graphs in Fig. 2 can be explained by the shorter run lengths which occur when the initial transient data are not deleted. We do this first theoretically by approximately correcting the coverage (p_{NoDel}) in Tables 4–6 to account for the shorter run lengths. As with any correlated sample, the variance of the sample mean waiting time is given by (see, for example, [13, p. 230]):

$$\sigma^2(\hat{W}) = \frac{\sigma_W^2}{N}\left(1 + 2\sum_{j=1}^{N-1}\rho_j(1 - j/N)\right), \tag{1}$$

where σ_W^2 is the steady-state waiting time variance, N is the number of observations, and ρ_j is the lag-j autocorrelation. For large values of N, the term in parentheses only depends weakly on N, as the autocorrelations ρ_j usually decay away exponentially. In fact for large N, (1) is often written as [1]:

$$\sigma^2(\hat{W}) \approx \frac{\sigma_W^2}{N}\left(1 + 2\sum_{j=1}^{\infty}\rho_j\right),$$

Thus if we use the subscripts *Del* and *NoDel* to indicate the variances of the sample mean waiting time and the number of observations with and without deletion, then:

$$\sigma_{Del}^2(\hat{W})/\sigma_{NoDel}^2(\hat{W}) \approx N_{NoDel}/N_{Del} \tag{2}$$

Since the estimated coverages are the results of thousands of independent replications, we use a normal approximation. The first step is to convert the coverage into points on a normal distribution. For a coverage of p_{NoDel}, the area to the left of the right hand tail on a standard normal distribution is $(1 + p_{NoDel})/2$. Thus the corresponding point on the normal distribution (z score) is $F^{-1}((1 + p_{NoDel})/2)$, where F is the standard normal cumulative distribution function. From (2) this is scaled by $(N_{Del}/N_{NoDel})^{1/2}$, compensating for the shorter run length without deletion, to produce a wider confidence interval, from which in turn we can re-estimate the coverage from:

$$p'_{NoDel} = 2F\left((N_{Del}/N_{NoDel})^{1/2}F^{-1}((1 + p_{NoDel})/2)\right) - 1$$

In the last two columns of Tables 4–6 this correction is applied to the coverage without deletion, p_{NoDel}, to produce an estimate of what the coverage would have been, had the run lengths been longer. It can be seen that except for two cases (out of 27), both within the margin of error, the *corrected* coverage without deletion (p'_{NoDel}), is greater than the coverage obtained with deletion of the transient period, (p_{Del}). If a relative precision of 10% is used, all the corrected coverages are greater than the coverage with deletion. (The results for these are not shown to save space.) While it would be unwise to put too much emphasis on the magnitude of the changes, since these are based on average run lengths, it appears possible that the differences in run lengths account for the differences in coverage. We next give our explanation why the run lengths without deletion are shorter.

We have noted that the average run lengths without deletion of transient data are uniformly shorter than those when deletion is used. Our conjecture is that this is because, when the system starts from the idle state, the results measured during the transient period are not only biased but also have low variance. This leads to optimistic estimates of the relative precision and hence shorter runs.

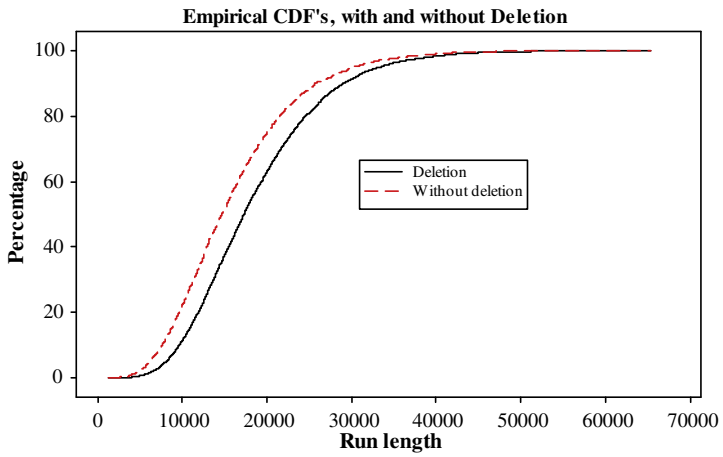

Fig. 3. Run length CDF's for M/D/1, $\rho = 0.3$, 5% relative precision.

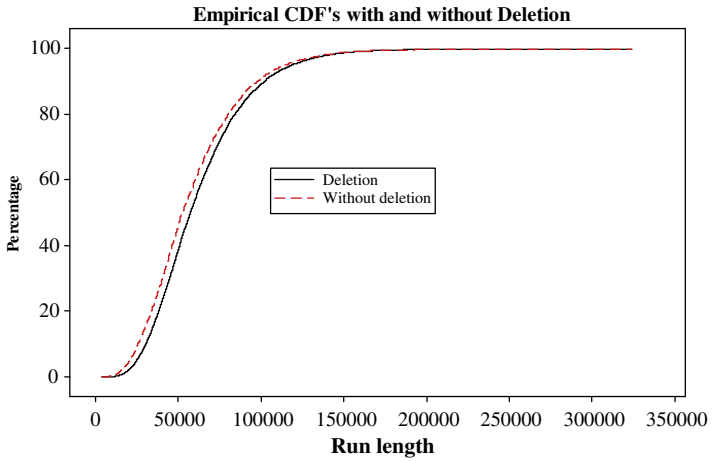

Fig. 4. Run length CDF's for M/M/1, $\rho = 0.6$, 5% relative precision.

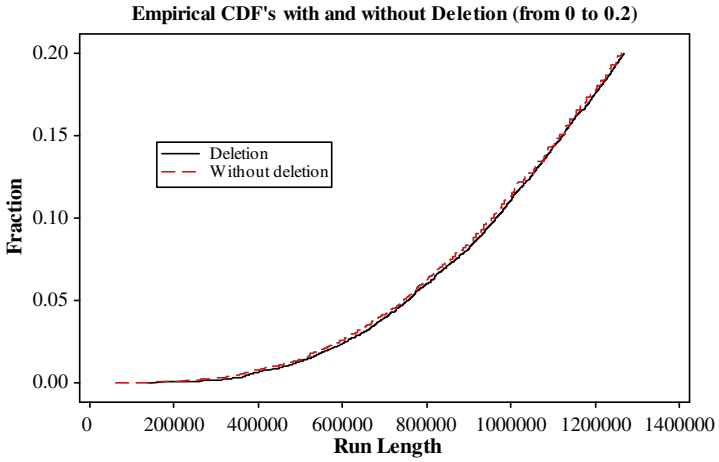

Fig. 5. Run length CDF's for M/H_2/1, $\rho = 0.9$, 5% relative precision.

Figs. 3–5 show the cumulative distribution of run lengths for representative models, covering the range from the shortest average run length, M/D/1, with a load of 0.3, to the longest run length, M/H_2/1, with a load of 0.9. (For clarity only the portion up to the 20th percentile of this graph is shown.)

What is noticeable is that in all the graphs the difference in distribution occurs very early on, with the 10th percentile of no-deletion runs occurring significantly earlier than that for the runs with transient data deletion, and that there is very little additional difference in the distributions thereafter (the difference in height between the two distributions remains constant up to about the 95th percentile.)

So other things being equal there is a close-to-constant reduction in the run length due to not deleting the data from the initial transient. Our explanation for this is that when starting from empty and idle, the observations collected during the transient period have lower variance than those from the "steady state" portion of the simulation. This produces an underestimate of the variance of the sample mean waiting time, which in turn causes premature stopping. For the M/M/1 model we can support this explanation by calculating the variance of the waiting time of the Nth arriving customer, this time using an extension of the Markov chain technique of Kelton and Law.

Fig. 6 shows the variance of the waiting time of the Nth arriving customer at an M/M/1 queue with a load of 0.9, converging towards its steady state value of 99.0 (using an arrival rate of $\lambda = 0.9$ and a service rate of $\mu = 1$.) The formula for this: $1/(\mu-\lambda)^2 - 1/\mu^2$, can be derived from the exponential distribution for the time in system (see, for example [8, p. 65]). This is plotted out as far as the 726th arrival, since that was the average number of observations deleted (Table 1).

Thus during the transient period, the contribution towards the long-run estimate of the variance of the waiting time is biased downwards. From the theoretical formula (1) for the variance of the sample mean waiting time, it follows that since the estimate of the variance of the waiting time, σ_W^2 is too small, the confidence interval for the sample mean waiting time will appear smaller than it actually should be, and the run will finish early. Thus while the bias in the mean due to the initial waiting times disappears in the long run length, the sequential run length is reduced if the initial transient is not deleted. Since the effect on the variance of the sample mean is a scaling rather than an additive effect, it is only reduced proportionately, but not eliminated, by setting the relative precision to smaller values.

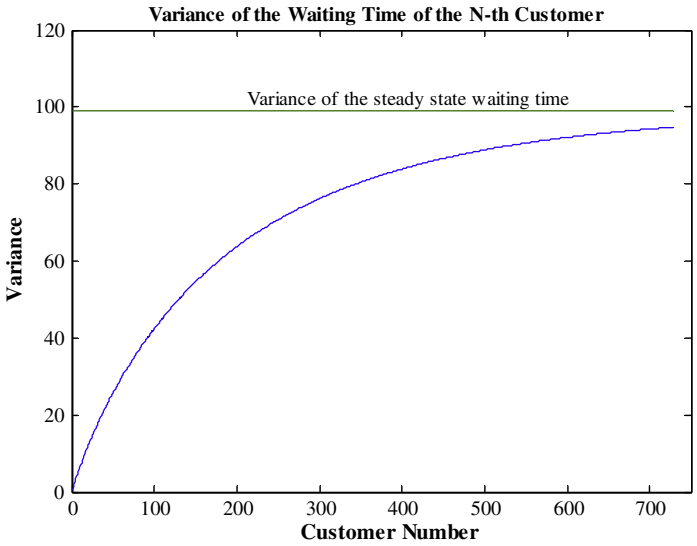

Fig. 6. Waiting time variances during the transient period.

To remove the effect of reduced run lengths when the transient data are not deleted, and the general variability produced by sequential stopping, we ran the same models with fixed run lengths for each case. The run lengths have been set at the theoretical number of observations required to reach the required relative precision. The number of observations can be calculated, for M/G/1 queues, from the equations in Daley [1]. From Eq. (28) there, the variance of the sample mean waiting time, for a large sample, is also given by:

$$\sigma^2(\hat{W}) = \frac{\sigma^2}{N}\left(\frac{1+\rho}{1-\rho} + \frac{\lambda(E[W^3] - E[W]E[W^2])}{(1-\rho)(E[W^2] - E[W]E[W])}\right),$$

Then for a 95% confidence interval to have a relative precision of say 0.05, we need $P(|\hat{W} - E[W]| \leq 0.05E[W]) = 0.95$ or $0.05E[W]/\sigma(\hat{W}) = 1.96$, from which we can solve for the number of observations required, as listed in Table 7. The moments of the waiting time, $E[W]$, $E[W^2]$ and $E[W^3]$ were found by differentiating the Pollaczek–Khinchine transform expression (see, for example [8, p. 237]) using Maple. For M/D/1 the values of N are the limits of those for M/E$_k$/1 as $k \to \infty$.

The coverages for the same set of models as before, but with fixed run lengths from Table 7, are plotted in Fig. 7. As in Fig. 2, solid lines connect the coverages where transient deletion was used, and dashed lines the coverages where there was no deletion. Note that as the differences are now much smaller, the vertical scales have been magnified by a factor of 2.5 over those used in Fig. 2.

From Fig. 7, now that we have removed the effect of premature stopping, we find there are no statistically significant differences in coverage for the 54 comparison experiments (nine traffic intensities, three models, and two levels of relative precision) as can be seen from the way in which the experimental confidence bands overlap at each point in the graphs. In 26 cases the coverage with deletion of initial transient data was greater than that for no deletion, and in 28 cases it was less. We could not find any significant differences between the coverage at 10% relative precision, and that at 5%. Thus fixing the run lengths at appropriate numbers removes all detectable effects of deleting or not deleting initial transient data. We can conclude that the "brute force" approach to the transient problem works at these run lengths, and has in fact worked by the run length required for an accuracy of 10% relative precision.

Table 7
Number of observations for a specified relative precision.

ρ	M/D/1		M/M/1		M/H$_2$/1	
	10%	5%	10%	5%	10%	5%
0.1	6639	26,559	11,671	46,687	37,491	149,964
0.2	4401	17,607	8547	34,190	34,094	135,376
0.3	4007	16,028	8276	33,105	36,085	144,343
0.4	4268	17,073	9134	36,537	40,977	163,908
0.5	5122	20,488	11,140	44,562	49,556	198,226
0.6	6936	27,744	15,110	60,441	64,880	259,521
0.7	10,976	43,904	23,677	94,710	95,948	383,794
0.8	22,409	89,637	47,443	189,775	177,674	710,696
0.9	82,523	330,093	170,268	681,072	577,022	2,308,090

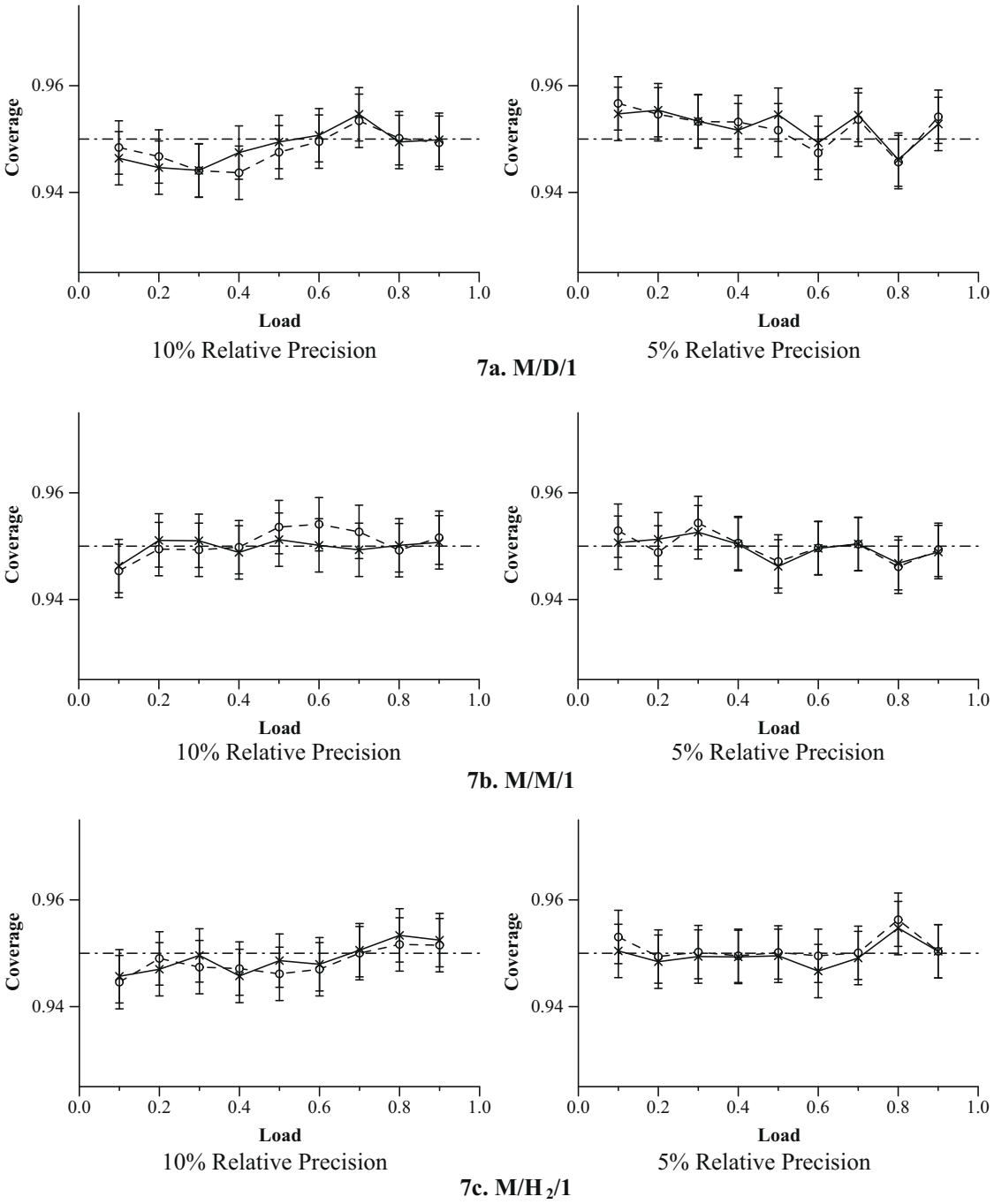

Fig. 7. Achieved coverage with fixed run lengths.

3.3. Does initial loading support our explanation?

Since it appears that collecting observations from when the system starts from empty and idle causes lower variance estimates, shorter runs, and hence poorer coverage in these models, what happens if we change the initial condition? That is, if the system starts from a high initial state, we might expect that the variance contribution during the transient period should be large, and hence run lengths will be longer and coverage better if our explanation is correct. That broadly is true, although the explanation is confused by the fact that for some combinations of load and initial loading the convergence of the variance to its steady state may now not be monotonic, unlike that shown in Fig. 6. High load and high initial loadings is one class where this occurs. So we give the results for only one system, and draw some conservative conclusions. Fig. 8 plots the 5% relative precision coverages with deletion (solid lines) against the coverages produced when the system starts with 10, and 100 customers present (and no deletion), (dashed lines) for an M/M/1 queue. Table 8 gives the average run lengths for these cases, and for the case of no deletion. The run lengths for 10 initial customers are mostly higher than those for no deletion. Transient deletion appears to produce superior coverage to that from 10 initial customers, but most of the

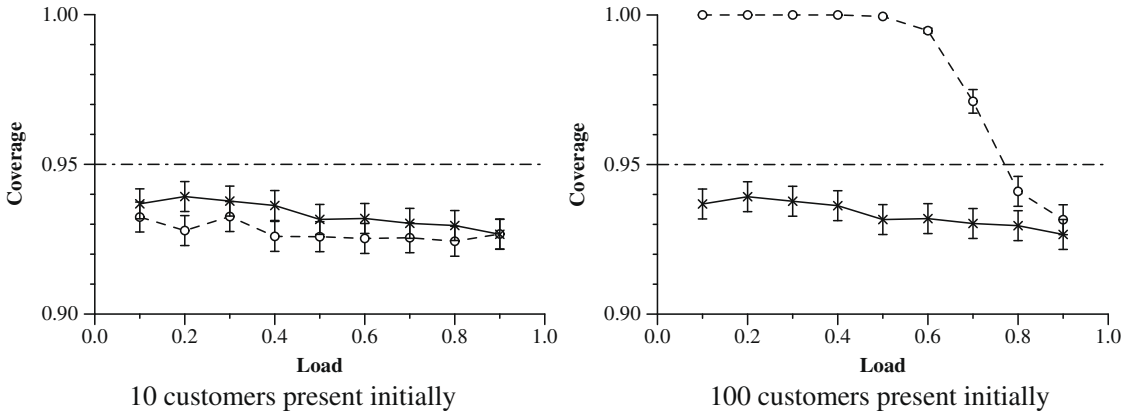

Fig. 8. The effect of initial loading. M/M/1 5% relative precision.

Table 8
Average run lengths M/M/1 model 5% relative precision.

ρ	Idle and no deletion	Idle with deletion	10 Customers initially (no deletion)	100 Customers initially (no deletion)
0.1	45,667	50,050	83,445	2,959,875
0.2	33,502	37,231	40,694	1,346,648
0.3	32,385	35,936	34,494	864,062
0.4	35,623	39,309	36,384	631,357
0.5	43,529	47,281	43,355	505,437
0.6	58,372	62,596	57,884	414,081
0.7	90,900	95,886	90,677	358,764
0.8	181,939	187,359	181,226	357,446
0.9	644,927	654,449	654,335	718,170

confidence intervals overlap. 100 initial customers produces much higher coverages, but this is caused by extraordinary run lengths (up to 64 times longer than those where the system starts from empty).

The bias in the mean due to initial loading should be easily wiped out in runs of these lengths. Hence given that, the effects do appear compatible with our conjecture that it is (now over-) estimated confidence intervals, caused by the abnormally high contribution to the variance from the early observations that are causing the higher run lengths and hence the gains in coverage. We would expect the proportional effect of this to be highest when the load is small, as in this case the variance of the remainder of the run should be small, and this is exactly what happens.

Initial loading is often suggested as an alternative method of dealing with the initial transient. The results we have are compatible with there being some benefit in terms of coverage, but the results for initial loadings of 100 customers show that in sequential simulation this may come at the price of excessively long runs. Given the possible non-monotonic convergence of the variance (and hence the difficulty in uniformly supporting our explanation) and the small class of models considered here, we do not draw more extensive conclusions at this time.

For 100 customers present initially there are no confidence bands on the first six points, loads = 0.1, . . . , 0.6 in Fig. 8. This is because we could not obtain 200 "bad" confidence intervals before these runs were stopped after 30,000 independent replications, and thus we could not estimate the coverage, which is close to 100%, to our set standard.

4. Other experiments

Experiments were also conducted for response times (times in system) for the same set of models, and the response times in a simple computer network model as shown in Fig. 9. Here after processing, a random fraction of jobs p_1, p_2 return to Disk 1 or Disk 2, respectively. A fraction p_3 leave the system. For the results in Fig. 10 the mean CPU service time is 6, the mean service time for each disk is 14, $p_1 = p_2 = 0.4$, all distributions are negative exponential, and the source rate is set to give loads at the CPU ranging from 0.1 to 0.9.

For the computer network model the estimated coverage without deletion of transient data (dashed line) is always lower than that with deletion, (solid line) significantly so for 10% relative precision. But for 5% relative precision coverages are very close to those in which data from the transient period has been deleted. This is presumably due to the usually positive autocorrelation in the input process produced by the feedback of jobs resulting in very long run lengths [15]. The results of all of these models were quite consistent with those observed for waiting times: poorer coverage if transient data are not deleted, but with this poorer coverage being entirely explainable in terms of shorter run lengths, and a probable explanation for the

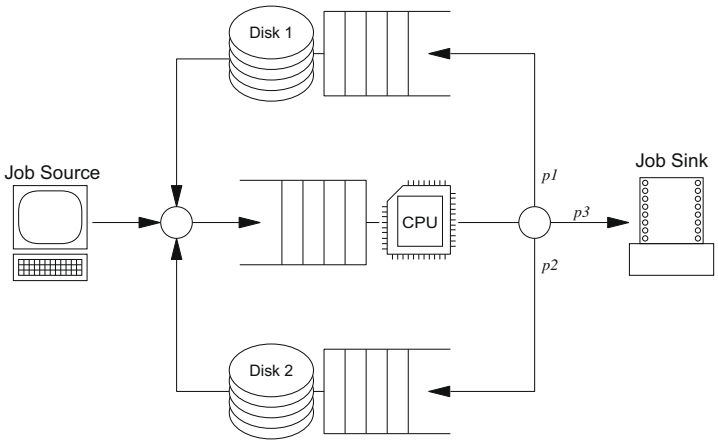

Fig. 9. The computer network model.

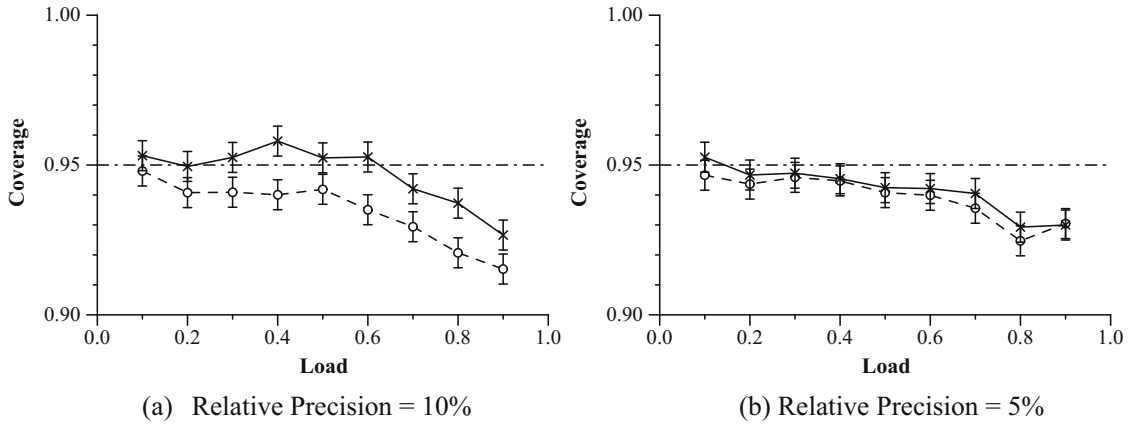

Fig. 10. Effect of transient deletion on the CPU queue response time coverage.

shorter run lengths being lower variance of the observations collected during the transient period. Fixed run-length experiments were not tried as the expected run lengths are unknown.

5. Conclusions

At first sight, from Fig. 2, it appears that, for sequential simulation, deleting the initial transient data does produce a direct improvement in the coverage of the results from sequential simulation. The coverages obtained without deleting the initial transient data are usually worse than those where a relatively small number of initial observations have been deleted. There is some tendency for the coverage to get better as the relative precision is reduced from 10% to 5%, although this is not a uniform effect. So an initial conclusion might have been that in sequential simulation removing the initial transient data directly improves the coverage and hence should be done for that reason, and that improved coverage cannot necessarily be obtained by using a requirement for smaller confidence intervals to increase the run length. In other words it appears that the "brute force" approach does not always work.

However, when we fix the run-lengths to the theoretical numbers required for commonly used levels of relative precision (Fig. 7), the direct gains in reduced bias and improved coverage achieved by deleting the initial transient data are extremely modest. For the range of models considered, there was almost no detectable difference between the means and the coverages of the results provided the same run length was used. So it appears that at these run lengths the effects of the initial transient have indeed been "washed out". So if we can ensure that the runs are long enough it appears to be possible to rely on high-precision (i.e. small) confidence intervals in order to guarantee the accuracy of the final results. This might be helpful in some situations, for example those for which transient deletion methods have not been validated, or where transient deletion methods give highly variable results.

Our experiments show that in sequential steady-state simulation, deleting the initial transient data can provide considerable additional protection against premature stopping, apparently by ensuring that the variance of the sample mean is more accurately estimated. And it is this which improves the coverage to closer to the specified level. This happens even

when a simple and conservative method, based on a heuristic and Schruben's test, is used. Since premature stopping and hence the production of overly optimistic confidence intervals is a chronic problem of sequential simulation, this is a valuable contribution. It appears that initial loading of the system can have a similar, although less predictable effect.

The results also emphasise the need to test new proposals for transient deletion methods for more than how they do on mean values – at least variances and preferably the entire distribution of the process of interest should be considered. A method which does this is reported in [2].

Acknowledgement

The two referees contributed carefully considered, substantial suggestions for improving this paper.

References

[1] D.J. Daley, The serial correlation coefficients of waiting times in a stationary single server queue, Journal of the Australian Mathematical Society 8 (1968) 683–699.
[2] M. Eickhoff, K. Pawlikowski, D. McNickle, Detecting the duration of initial transient in steady state simulation of arbitrary performance measures, in: Proceedings ACM ValueTools07, Nantes, France, 23–25 October 2007.
[3] G. Ewing, K. Pawlikowski, D. McNickle, Akaroa-2: exploiting network computing by distributed stochastic simulation, in: 13th European Simulation Multiconference, Warsaw, Poland, SCSC, June 1999, pp. 175–81.
[4] G. Ewing, D. McNickle, K. Pawlikowski, Spectral analysis for confidence interval estimation under multiple replications in parallel, in: Proceedings of the 14th European Simulation Symposium, Dresden, October 2002, pp. 52–61.
[5] A.V. Gafarian, C.J. Ancker, T. Morisaku, Evaluation of commonly used rules for detecting "steady state" in computer simulation, Naval Research Logistics Quarterly 78 (1978) 511–529.
[6] B. Ghorbani, Initial transient phase of steady state simulation: methods of its length detection and their implementation in Akaroa2, M.Com. Thesis, Computer Science and Software Engineering, University of Canterbury, 2004.
[7] D. Goldsman, L. Schruben, J. Swain, Tests for transient means in simulated time series, Naval Research Logistics 41 (1994) 171–187.
[8] D. Gross, J.F. Shortle, J.M. Thompson, C.M. Harris, Fundamentals of Queueing Theory, Wiley, New Jersey, 2008.
[9] P. Heidelberger, P.D. Welch, A spectral method for confidence interval generation and run length control in simulations, Communications of the ACM 24 (4) (1981) 233–245. April.
[10] P. Heidelberger, P.D. Welch, Simulation run length control in the presence of an initial transient, Operations Research 31 (1983) 1109–1144.
[11] K. Hoad, S. Robinson, R. Davies, Automating warm-up length simulation, in: S.J. Mason et al. (Eds.), Proceedings of the 2008 Winter Simulation Conference, 2008, pp. 532–540.
[12] W.D. Kelton, A.M. Law, The transient behavior of the M/M/s queue, with implications for steady-state simulation, Operations Research 33 (1985) 378–396.
[13] A.M. Law, Simulation Modelling and Analysis, fourth ed., McGraw-Hill, Boston, 2007.
[14] J.-S.R. Lee, K. Pawlikowski, D. McNickle, Sequential steady-state simulation: rules of thumb for improving the accuracy of the final results, in: Proceedings of the ESS99, Erlangen, Germany, 1999, pp. 618–622 (October 26–28).
[15] D. McNickle, Autocorrelations in the input and output processes in simple Jackson networks, New Zealand Operational Research 12 (1984) 109–118.
[16] D. McNickle, K. Pawlikowski, G. Ewing, Refining spectral analysis for confidence interval estimation in sequential simulation, in: Proceedings of the ESS, Budapest, October 2004, pp. 99–103.
[17] D. McNickle, K. Pawlikowski, C. Stacey, Detection and significance of the initial transient period in quantitative steady-state simulation, in: Proceedings od the Eighth Australian Teletraffic Research Seminar, RMIT Melbourne, 6–8 December 1993, pp. 193–202.
[18] K. Pawlikowski, Steady state simulation of queueing processes: a survey of problems and solutions, ACM Computing Surveys 22 (1990) 123–170.
[19] K. Pawlikowski, D. McNickle, G. Ewing, Coverage of confidence intervals from sequential steady-state simulation, Simulation Practice and Theory 6 (1998) 255–267.
[20] L. Schruben, Detecting initialisation bias in simulation output, Operations Research 30 (3) (1982) 569–590.
[21] L. Schruben, H. Singh, L. Tierney, Optimal tests for initialisation bias in simulation output, Operations Research 31 (6) (1983) 1167–1178.