Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research 176 (2007) 332–346
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Stochastics and Statistics

A statistical process control approach to selecting a warm-up period for a discrete-event simulation

Stewart Robinson *

Operational Research and Information Systems Group, Warwick Business School, University of Warwick, Coventry CV4 7AL, United Kingdom

Received 13 January 2005; accepted 6 July 2005
Available online 6 December 2005

Abstract

The selection of a warm-up period for a discrete-event simulation continues to be problematic. A variety of selection methods have been devised, and are briefly reviewed. It is apparent that no one method can be recommended above any other. A new approach, based upon the principles of statistical process control, is described (SPC method). Because simulation output data are often highly autocorrelated and potentially non-normal, the batch means method is employed in constructing the control chart. The SPC method is tested on seven data sets and encouraging results are obtained concerning its accuracy. The approach is also discussed with respect to its ease of implementation, simplicity, generality of use and requirements for parameter estimation.

Keywords: Simulation; Initial transient; Warm-up period; Statistical process control

1. Introduction

The problem of the initial transient has long been discussed in the simulation literature. It is important that it is dealt with appropriately in order to avoid any bias in the estimate of a simulation's steady-state parameters. One of the main approaches for dealing with the initial transient is to run the simulation for a warm-up period and then to delete these data. In using this approach the question of how to choose the warm-up period arises. A variety of methods have been proposed for selecting a warm-up period, but it apparent that no one single method can be recommended above any of the others. Indeed, it appears that simulation practitioners rely upon simple techniques, typically inspecting time-series of the output, to determine the length and the potential effect of the initial transient. Such methods may not be

* Tel.: +44 2476 522132; fax: +44 2476 524539.
E-mail address: stewart.robinson@warwick.ac.uk

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2005.07.014

sufficiently rigorous to give accurate results. All-in-all this suggests that further research into this area is required.

The purpose of this paper is to describe a new approach to determining the warm-up period for the output of a discrete-event simulation model, based upon the principles of statistical process control (SPC). Tests with this method demonstrate its efficacy. One particular advantage is that it is based on an approach that is familiar to many simulation practitioners. Before describing the SPC method, there is a brief review of other methods that have been developed for determining the warm-up period. The SPC method is then described and it is demonstrated by applying it to the output from a model that has been developed for a real-life simulation study. The method is tested by applying it to data from time-series with known characteristics. Conclusions are drawn concerning its effectiveness and some issues for further research are discussed.

2. Methods for selecting a warm-up period

There is a long history of research into the initial transient and methods for both detecting it and removing it to ensure a more accurate estimate of the steady-state parameters of a model. In general, two approaches are used: data deletion and intelligent initialisation (Nelson, 1992). In the first approach, the model is run-in for a warm-up period until it reaches a steady-state and then these data are deleted. The key challenge in this case is to determine the length of the warm-up period. If it is underestimated, there will be some bias in the simulation results. If it is overestimated, output data are wasted and the number of experiments that can be performed in a period of time is reduced. In the second approach, the model is placed in a realistic initial condition at the start of the run. Here the key challenge is to determine what constitutes a realistic starting state. In principle these approaches are the same with both attempting to place the model in a realistic condition from the point at which the output data start being collected. This paper concentrates on the former approach and the challenge of determining an appropriate warm-up period.

A wide variety of methods for estimating the warm-up period have been proposed over the past 40 years. These can be categorised under five headings that are briefly outlined below: graphical methods, heuristic approaches, statistical methods, initialisation bias tests and hybrid methods.

2.1. Graphical methods

These approaches rely upon the visual inspection of time-series of the simulation output. They range from straightforward inspection of the simulation output, to the partitioning of the data into batches and the calculation of moving averages. Among the most popular of these methods are simple time-series inspection (Robinson, 2004) and Welch's method (Welch, 1983) which has been popularised by Law and Kelton (2000) and its inclusion in the AutoMod simulation package. Other methods include ensemble average plots (Banks et al., 2001), the cumulative mean rule (Gordon, 1969; Banks et al., 2001), the deleting the cumulative mean rule (Banks et al., 2001), CUSUM plots (Nelson, 1992) and variance plots (Gordon, 1969).

These methods have the advantage of simplicity and they rely upon few, if any, assumptions about the output data. They rely upon human interpretation and judgement. This has the advantage that it involves the analyst in the decision, but at the same time it may lead to quite different decisions depending on the expertise of the analyst. Some of the methods rely upon the use of cumulative statistics, for instance Welch's method, for which it is acknowledged that there is a tendency to overestimate the length of the initial transient and so waste valuable data (Gafarian et al., 1978; Wilson and Pritsker, 1978; Pawlikowski, 1990; Roth, 1994).

2.2. Heuristic approaches

This set of methods provide rules for determining the length of the initial transient. Some of the graphical methods can be adapted to include heuristic approaches, for instance, Schriber describes a heuristic rule that adds more formality to ensemble average plots (Pawlikowski, 1990). Among the other heuristic methods described in the literature

are: the Conway rule or forward data interval rule (Conway, 1963), the modified Conway rule or backward data interval rule (Gafarian et al., 1978), the autocorrelation estimator rule (Fishman, 1971), the crossing-of-the-mean rule (Fishman, 1973), the marginal confidence rule (White, 1997), and the marginal standard error rules (MSER and MSER-5) (White et al., 2000).

The use of specific rules has the advantage of removing the subjectivity of the graphical methods. This may, however, mean that important patterns in the data are not identified, particularly if a graphical representation of the data is not used in conjunction with the rule. There are few assumptions about the nature of the output data in most of the heuristic approaches and their simplicity makes them reasonably straightforward to implement.

2.3. Statistical methods

These methods rely upon the principles of statistics to estimate the warm-up period. Two key methods fall within this category. Kelton and Law (1983) describe a regression based method that makes a reverse pass through the data, adding data points to the cumulative mean of the data as long as it continues to have a zero slope. Meanwhile the randomisation test for initialisation bias, devised by Yücesan (1993), is based on the random shuffling of batched data and testing for the significance in the difference of two means. A significant difference being indicative of initialisation bias.

These two methods are much more complex than the graphical or heuristic approaches. They require extensive computations and rely upon some restrictive assumptions. These methods do not appear to have been pursued further in the literature.

2.4. Initialisation bias tests

Initialisation bias tests are not strictly methods for determining the warm-up period, but for determining whether initialisation bias is present in a series of data. As such they can be used in conjunction with the methods described above to determine whether the selected warm-up period is at least of sufficient length, although they will not determine that the warm-up period is greater than required should a conservative estimate be made.

Three such tests have been devised by Schruben: the maximum test (Schruben, 1982), Schruben's modified test (Schruben, 1982) and the optimal test (Schruben et al., 1983). Vassilacopoulos (1989) describes the Rank test and Goldsman et al. (1994) describe a family of tests based on the use of batch means. Of these methods, the optimal test appears to be most popular.

The main advantage of these tests is that they enable an analyst to determine whether the initialisation bias has been removed successfully, although with varying degrees of accuracy depending on the nature of the simulation output. The theory underlying the methods is quite complex, but the computations are relatively modest. A key problem is that most of the methods require the estimation of the variance of the data. Also, some of these tests are computationally unstable, sometimes resulting in divisions by zero (Nelson, 1992; White et al., 2000).

2.5. Hybrid methods

This final set of methods employs initialisation bias tests in conjunction with graphical or heuristic approaches in a formal manner. Two key methods are Pawlikowski's sequential method (Pawlikowski, 1990) and the scale invariant truncation point method (Jackway and DeSilva, 1992). The former is complex requiring the use of spectral variance analysis. The latter requires an estimate of the variance and it may need a large amount of data to perform well since it is based on asymptotic theory, although tests performed by the authors suggest that it is reasonably accurate.

2.6. Conclusion on methods for selecting a warm-up period

Tests that have been performed on the methods outlined above demonstrate that they estimate the warm-up period with varying degrees of success, in part dependent on the nature of the output data that are being analysed. There certainly does not appear to be one method that can be recom-

mended above any others. For many, only limited testing has taken place, and few have been reportedly used in simulation case studies. Indeed, the literature and anecdotal evidence would suggest that the only methods in common use, apart from guessing, are the time-series inspection method and Welch's method, both of which are subject to some difficulties. Albeit that a variety of approaches have been proposed over the past 40 years, there is a need both to test these approaches more thoroughly and to devise new methods that could be adopted widely by simulation users. This paper contributes to the latter by describing a new method for estimating the warm-up period that, because it is based on a familiar technique, could be adopted by simulation users.

3. Description of the SPC method

In SPC two forms of variation are identified (Montgomery and Runger, 1994). 'Chance' or 'common' causes are sources of variation that are unavoidable. In manufacturing systems these are caused by small variations in, for instance, the materials being processed, the environment (e.g. heat and light) and the performance of the staff. In these circumstances the process is assumed to be varying according to some fixed distribution about a constant mean. Meanwhile, 'assignable' or 'special' causes of variation can be corrected or eliminated, for instance, a tool may be worn or staff may be poorly trained. In these circumstances the process ceases to vary according to some fixed distribution about a constant mean. A process that is subject only to chance causes of variation is said to be 'in-control'. When assignable causes of variation occur, the process is said to be 'out of control'. The purpose of SPC is to determine when a process has gone out of control, so the process may be corrected.

There is a close relationship between the concepts of SPC and those of transience and steady-state in simulation output analysis. Simulation output that is in steady-state varies about a constant mean according to some fixed distribution, while in transience it does not. As such, a simulation that is in steady-state could be considered to be 'in-control', while during a transient phase it could be considered to be 'out of control'. It seems possible, therefore, that the methods employed in SPC could be used to detect when simulation output is in steady-state, and when it is not. The method presented here employs SPC in this way, with a particular view to identifying the initial transient for a simulation with steady-state parameters. Unlike the implementation of SPC in real-life processes, there is no sense of intervening in a model run in order to return the model to a steady-state, it is assumed to be a self-controlling process. The method is now described by a series of stages that are based on SPC theory.

3.1. Stage 1: Perform replications and collect output data

First, time-series data on a key output statistic need to be collected. Typically these would be throughput for manufacturing systems and customer waiting time in service systems. The output data should be collected over a series of replications in order to provide a number of samples for each data point. The observation interval should be relatively short; for typical manufacturing and service system models a 1 hour interval is probably appropriate.

Standard SPC approaches assume that the data are normally distributed. This, of course, may not be the case for simulation output data. By performing a series of independent replications, however, it can be assumed (via the central limit theorem) that the sample mean at each time interval tends towards normality, unless the data are highly skewed. The larger the sample size the stronger this assumption. Therefore, the more replications that can be performed the better. It is recommended that at least five replications are performed.

The length of each replication also needs to be determined. It is recommended that the run-length is at least twice as long as the estimated length of the initial transient and that at least 20 data points are collected in each replication, although the more data that can be collected the better. The rationale for this is given in stage 3 of the method. An initial estimate of the length of the initial

transient could be obtained by simply inspecting a time-series of the output data to look for where the model appears to be in steady-state.

On completion of this stage a time-series of the key output data for each replication should have been collected, represented by Y_{ji}, where i is the observation number and j is the replication number. From this the sample means at each observation interval can be calculated as follows:

$$\overline{Y}_i = \frac{\sum_{j=1}^{n} Y_{ji}}{n}, \tag{1}$$

where n is the number of replications performed. As a result the vector $\bar{y}$, representing a time-series of sample means, is formed:

$$\bar{y} = (\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_m), \tag{2}$$

where m is the total number of observations made in each replication.

3.2. Stage 2: Test that the output data meet the assumptions of SPC

Standard SPC approaches rely on two key assumptions: that the data are normally distributed and that the data are not correlated. In stage 2 these two assumptions are tested on the second half of the data in the time-series, since these are the data from which the SPC parameters are calculated (stage 3).

Simulation output data are typically autocorrelated. The batch means method is one approach for dealing with this autocorrelation (Law and Kelton, 2000). Indeed, Runger and Willemain (1996) use batch means to deal with autocorrelated data in SPC. By combining observations into batches, the batch means become approximately uncorrelated as the batch size (k) increases. The time-series from Eq. (2) is combined into a series of batches (h) to generate the batch means as follows:

$$\overline{\overline{Y}}_h = \frac{\sum_{i=(h-1)k+1}^{hk} \overline{Y}_i}{k} \quad \text{for } h = 1, 2, \ldots, \left\lfloor \frac{m}{k} \right\rfloor. \tag{3}$$

This results in the following vector of batch means:

$$\bar{\bar{y}} = (\overline{\overline{Y}}_1, \overline{\overline{Y}}_2, \ldots, \overline{\overline{Y}}_b), \tag{4}$$

where b is the number of batches of length k that can be calculated from the time-series, such that $m \geqslant bk$.

A key problem with using the batch means method is determining an appropriate batch size (k). Fishman (1996) and Alexopoulos and Seila (1998) both present a number of methods for selecting a value of k. Fishman (1978) proposes a procedure based on the Von Neumann (1941) test for correlation. The batch size is doubled until the null hypothesis that there is no correlation between the batches is accepted. An advantage of this approach is that it can be applied to small sample sizes; as few as $b = 8$ batches. For this reason it is adopted here with a slight modification. In order to maximise the number of batches and so increase the sample size for calculating the SPC parameters, the minimum batch size for which the null hypothesis is accepted is sought. This is achieved by doubling the batch size until the null hypothesis is accepted and then by successively testing batch sizes that are at the mid-point between the last accepted and last rejected batch sizes. For instance, if $k = 32$ is accepted, then $k = 24$ would be tested. If this passes the procedure, then $k = 20$ would be tested, otherwise, $k = 28$ would be tested. The procedure stops when the smallest integer value of k is found for which the null hypothesis is accepted.

The data should also be tested for normality at each value of k. Various methods could be used to test for normality, for instance, the chi-square test and the Kolmogorov–Smirnov test. The Anderson–Darling test (Anderson and Darling, 1954) is generally preferred because it places more weight on the tails of the distribution and so has greater power. It should be noted that as k increases the batch means tend towards normality (Law and Kelton, 2000).

If during the tests for autocorrelation and normality the number of batches (b) falls below 20, further model runs are required (the rationale for this is given in the description of stage 3 below). This can be achieved either by performing more replications or by performing longer runs. It is recommended that the run-length is doubled until the conditions for autocorrelation and normality are met.

3.3. Stage 3: Construct a control chart for the batch means data

Once a value of k has been found for which the assumption of low autocorrelation and normality

are met, a control chart can be constructed for the data in Eq. (4). The population mean (μ) and standard deviation (σ) of the data in $\bar{\bar{y}}$ (Eq. (4)) are estimated from the data in the second half of the time-series. In other words, if the total number of batches (b) in the series is 100, then the mean and standard deviation are calculated for the data from batch 51 to batch 100. Note that if b were 99 then the range from batch 51 to batch 99 would be used.

In calculating the mean and standard deviation in this fashion, the assumption is being made that the mean and standard deviation are stable in the second half of the time-series. This is, of course, correct if the output data are in steady-state. Should this not be the case, then it is unlikely that the method will show that the data are in-control and a longer run-length will be required (stage 4). The estimation of parameters from the second half of a time-series is a procedure followed in other initial transient analyses, for instance, Schruben et al. (1983).

It is because the data in the second half of the series are used for parameter estimation that it is recommended that at least 20 data points are collected for each replication. The mean and standard deviation are then estimated from a sample size of at least 10. This also implies that the tests for autocorrelation and normality are based on a sample size of at least 10. This is reasonable since both Fishman's procedure and the Anderson–Darling test can be used with sample sizes as small as eight (D'Agostino and Stephens, 1986). Obviously the number of observations needs to be multiplied by k, in order to ensure that there are 20 data points in the batch means time-series. It is recommended that the run-length is at least twice the estimated length of the initial transient, so the mean and standard deviation are calculated on what are believed to be steady-state data.

The standard deviation is estimated as follows:

$$\hat{\sigma} = \sqrt{\frac{1}{\left\lfloor \frac{b}{2} \right\rfloor} \sum_{h=b-\left\lfloor \frac{b}{2} \right\rfloor+1}^{b} s_h^2}, \tag{5}$$

where s_h^2 is the variance for each batch mean in $\bar{\bar{y}}$ (Eq. (4)). This is calculated from the original data by generating batch means of size k for each individual replication and then determining the variance of the batch means that make up each value in $\bar{\bar{y}}$. For example, if five replications have been performed, s_3^2 would be the variance of the five individual batch means of size k that make-up the batch mean $\bar{\bar{\bar{Y}}}_3$.

From these data, three sets of control limits (CL) are calculated:

$$\mathrm{CL} = \hat{\mu} \pm z\hat{\sigma}/\sqrt{n} \quad \text{for } z = 1, 2, 3. \tag{6}$$

A control chart is then constructed showing the mean ($\hat{\mu}$), the three sets of control limits, and the time-series data $\bar{\bar{y}}$.

3.4. Stage 4: Identify the initial transient

The final stage is to view the plot of the control chart in order to determine the warm-up period required. During the initial transient the time-series data $\bar{\bar{y}}$ (Eq. (4)) will be 'out of control'. Once the simulation output reaches steady-state, the data will be 'in-control'. Montgomery and Runger (1994) identify a number of rules for determining whether a control chart is showing data that are out of control:

- A point plots outside a 3-sigma control limit.
- Two out of three consecutive points plot outside a 2-sigma control limit.
- Four out of five consecutive points plot outside a 1-sigma control limit.
- Eight consecutive points plot on one side of the mean.

In the context of determining the length of the initial transient, a fifth rule is added here:

- Initial points all plot to one side of the mean (as per expected bias).

The warm-up period can be selected by identifying the point at which the time-series data are in-control and remain in-control.

If a steady-state cannot be found this could mean one of two things. Firstly, the model has not been run for long enough to reach a steady-state. This can simply be addressed by returning to stage 1 and running the model for longer; it is

recommended that the run-length is doubled. Secondly, the model output does not ever reach a steady-state and that the output should be classified as transient.

A summary of the SPC method is provided in Fig. 1.

4. Illustrative example: The SPC method applied to data from a help desk simulation

In order to illustrate the SPC method for selecting a warm-up period it is now applied to the time-series output from a simulation model of a

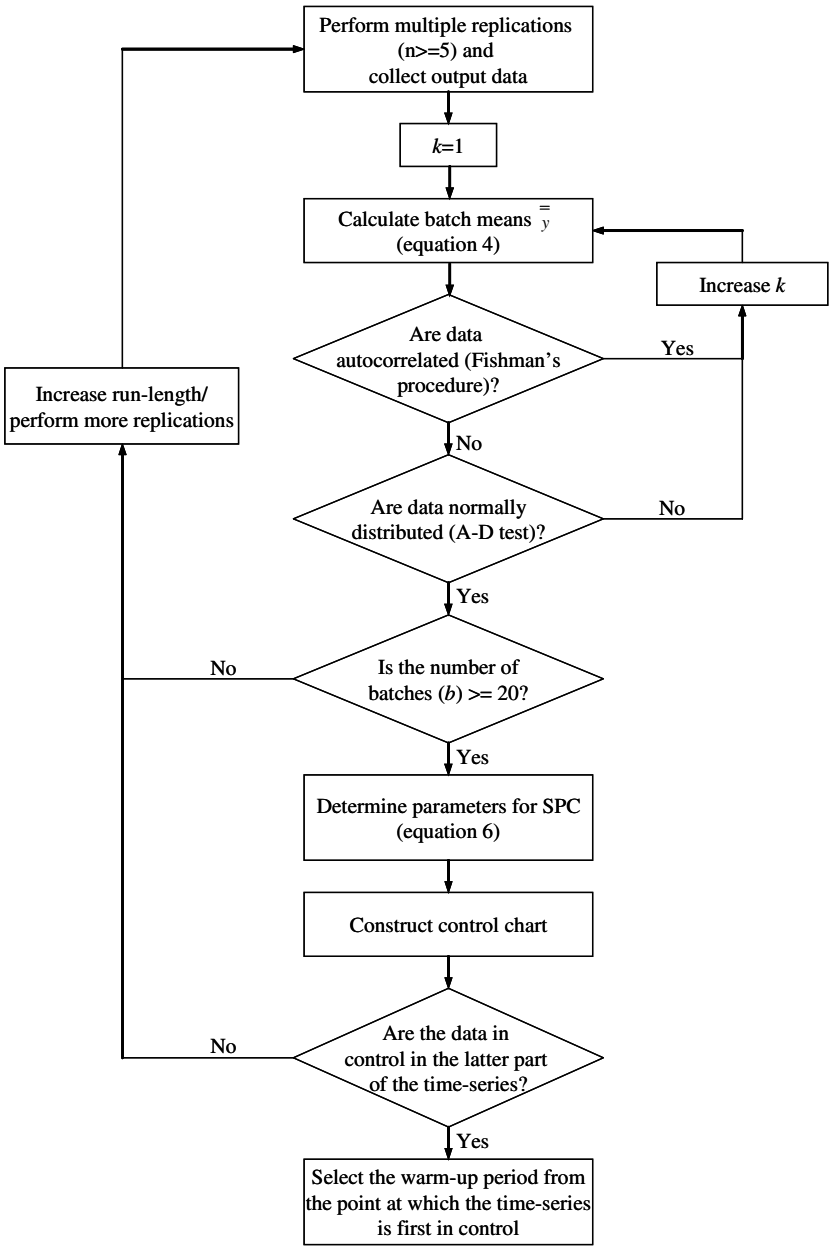

Fig. 1. Summary of the SPC method for choosing a warm-up period.

customer support help desk. The model was developed and used to investigate the operation of a real-life system.

The help desk receives requests for help via telephone, email or face-to-face contact. These may be attended to immediately, or logged and resolved via telephone contact or a visit to the customer. Stochasticity is caused by the random arrival of requests and the variability in service times. Various performance measures are of interest, including the number of jobs resolved, staff utilisation and the time requests spend in the system. Obviously, the length of the initial transient is not known.

Fig. 2 shows a time-series of the mean time the requests that are completed each day have spent in the system. Five replications were performed with the model, each of 250 days duration. The time-series shows the mean result from these replications. Because requests arrive at different rates during the day and the number of staff varies during the day, this model could be described as having steady-state cycle parameters (Law and Kelton, 2000). In order to remove the cycle from the time-series of the output data it is therefore necessary to set the observation period to days rather than hours.

The output data are tested to see if they meet the autocorrelation and normality assumptions of the SPC method, and the batch size (k) is increased until the assumptions are met. The results are shown in Table 1. Whereas the requirement for lack of correlation is met with a batch size of $k = 1$, the Anderson–Darling tests concludes that the data are not normally distributed. The value of k is doubled until the requirements for lack of correlation and normality are both met at $k = 4$. If the batch size is then reduced to $k = 3$, the autocorrelation requirement is not met and so a batch size of $k = 4$ is used for determining the SPC parameters.

The SPC parameters are shown in Table 2 and the control chart in Fig. 3. The mean and 3-sigma control limits are denoted by solid lines and the other control limits by dashed lines.

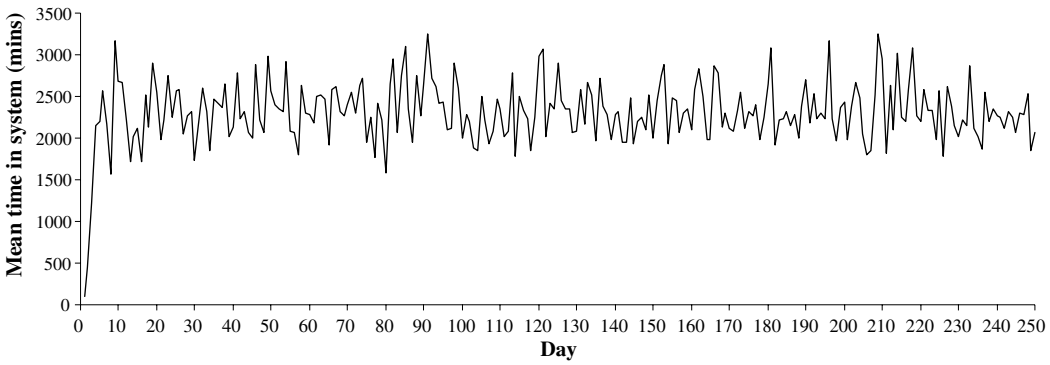

Fig. 2. Help desk model: mean time in system results from five replications.

Table 1
Autocorrelation and normality tests for help desk model

Batch size (k)	Result of Von Neumann test for correlation		Result of Anderson–Darling test for normality	
	Von Neumann statistic	Test result ($\alpha = 0.025$)	p-Value	Test result ($\alpha = 0.05$)
1	−0.169	Not correlated	0.004	Not normal
2	−1.734	Not correlated	0.010	Not normal
4	−0.878	Not correlated	0.758	Normal
3	−2.386	Correlated	0.008	Not normal

Table 2
SPC parameters for help desk model ($k = 4$)

Mean ($\hat{\mu}$)	Standard deviation ($\hat{\sigma}$)	Control limits		
		1-Sigma	2-Sigma	3-Sigma
2328.91	355.15	2328.91 ± 158.83	2328.91 ± 317.65	2328.91 ± 476.48

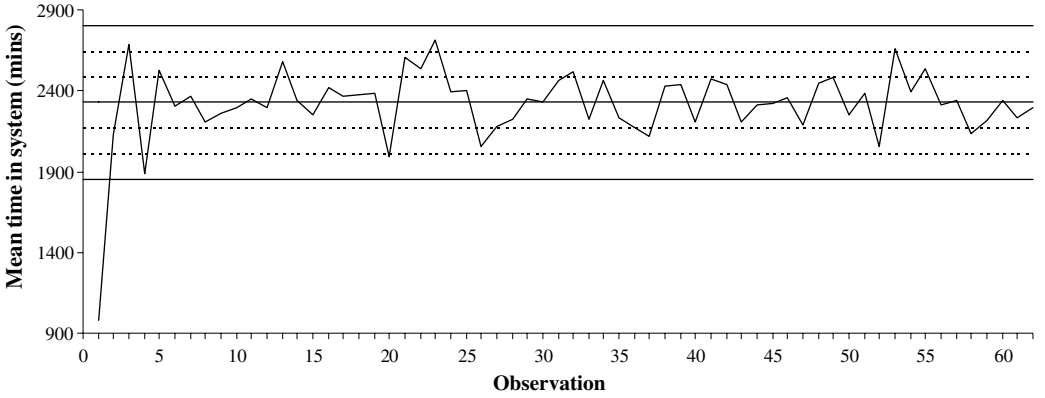

Fig. 3. Batch means control chart for the help desk model.

Observation 1 is well below the lower 3-sigma control limit and observation 2 is just above the lower 1-sigma control limit. Because the model starts with no requests in the system it is expected that there will be a negative bias in the initial data. Therefore, according to the fifth rule for the data being out of control, the first two data points are not in-control. Beyond observation 2 the time-series does appear to remain in-control. Therefore, a warm-up period of 2 batches, that is, 8 days ($2k$), is selected.

5. Testing the SPC method

5.1. Test models

The method is tested by applying it to time-series data generated from models with known characteristics. The four models used to generate the data have been used previously in the literature for investigating the initial transient problem. Model 1 is a first-order autoregressive (AR(1)) process with an error term that is normally distributed (Cash et al., 1992). Let X_t be the tth term in the series given by:

$$X_t = \Phi X_{t-1} + \varepsilon_t, \tag{7}$$

where

$$\varepsilon_t = N(0, (1 - \Phi)^2). \tag{8}$$

The constant Φ was set to 0.9 and X_0 to -0.96, giving a steady-state mean of 0. The model was used to generate sets of time-series (replications) with 250 observations and 500 observations, that is, $m = 250$ and $m = 500$ respectively.

Model 2 is the same as model 1, except that the error term (ε_t) is exponentially distributed with a mean of 1 (Goldsman et al., 1994). The steady-state mean is, therefore, 10. Again, time-series with 250 and 500 observations were generated using this model.

The third model is an M/M/1 queuing model (Cash et al., 1992). The arrival rate (λ) was set to 0.8 and the service rate (μ) to 1, giving a traffic intensity (ρ) of 0.8. The initial condition of the queue was empty. The steady-state mean delay for customers is 4 (Winston, 1994). The model

was run with $m = 2000$ arrivals and then with $m = 4000$ arrivals.

The final model is based on an M/M/1 queuing model and was used by Goldsman et al. (1994). It is referred to as GSS M/M/1. Let Z_t be the delay of the tth customer in an M/M/1 queue:

$$X_t = Z_t + 1 - a_p, \tag{9}$$

where

$$a_p = [1 - (p/4000)]^2 \quad \text{for } p = 1, \ldots, 4000, \tag{10}$$

$$a_p = 0 \quad \text{for } p > 4000. \tag{11}$$

For the M/M/1 model the arrival rate (λ) was set to 0.5 and the service rate (μ) to 1, giving a traffic intensity (ρ) of 0.5. Again, the initial condition was empty. Little's formula can be used to determine the steady-state mean of Z_t (which is 1) and so the steady-state mean of the model output is 2. The model was run for $m = 8000$ observations.

5.2. Criteria for assessing the SPC method

The SPC method was applied to the data from each model. Various criteria are used to assess the efficacy of the method. These have been adapted from Kelton and Law (1983) criteria for testing their method for identifying the warm-up period.

Of particular interest is the estimate of the steady-state mean obtained having deleted the data from the warm-up period identified by the SPC method. This measure is of interest since the main motivation for deleting initial transient data is to obtain a more accurate estimate of the steady-state mean. Similarly the bias, calculated as the difference between the estimated mean and the actual mean, is of interest.

Two criteria are used to assess the 90% confidence intervals calculated for the steady-state mean. First the coverage, that is the percentage of occasions on which the confidence interval covers the actual mean value. This is of interest, because it shows that the steady-state mean has been correctly identified. For 90% confidence intervals, the mean coverage would be expected to be 90%. The expected half length of the confidence interval is also determined. The interest here is in a half length that is narrow, since this implies a greater level of certainty about the value of the steady-state mean.

For the purposes of comparison, the results above are calculated with and without the use of the SPC method, that is, with and without deletion. The effect of deleting versus not deleting the initial data is identified by recording the percentage of occasions on which the absolute bias from deleting the data is greater than if the data had not been deleted; this obviously implies the method is performing poorly.

5.3. Test results

Table 3 shows how the data have been sampled from each of the test models. For the first three models a total of 500 replications have been performed. For the GSS M/M/1 model, 250 replications have been performed. For models 1 to 3, two run-lengths have been used; the second being double the first. This enables the effect of a greater run-length to be investigated. In every case the SPC method has been applied to time-series that are the mean of $n = 5$ replications (as per Eq. (2)), giving a sample of 100 (500/5) applications of the method for models 1 to 3 and a sample of

Table 3
Sample of data collected from the test models

Model	Model no.	Total reps.	Run length	n	Sample	Steady-state mean
AR(1) Normal a	Model 1	500	250	5	100	0
AR(1) Normal b	Model 1	500	500	5	100	0
AR(1) Exp. a	Model 2	500	250	5	100	10
AR(1) Exp. b	Model 2	500	500	5	100	10
M/M/1 $\rho = 0.8$ a	Model 3	500	2000	5	100	4
M/M/1 $\rho = 0.8$ b	Model 3	500	4000	5	100	4
GSS M/M/1	Model 4	250	8000	5	50	2

50 (250/5) for model 4. Table 3 also reports the actual mean for each model.

Table 4 shows the results obtained for the four models when no data are deleted from the time-series. The results are reported for time-series that are the mean of five replications ($n = 5$). For each of the assessment criteria the mean and 95% confidence intervals are given based on the sample of results obtained for each model.

Table 5 shows the results from applying the SPC method and deleting the selected quantity of data from each time-series. Apart from the assessment criteria, the deletion point is reported with a 95% confidence interval. The number of occasions on which a warm-up period was estimated with the method is also shown. For models 1–3 there were a number of occasions on which the SPC method could not identify a warm-up period with the data that were presented. The reasons for this are discussed below. Note that this is the reason why the percentage coverage for these models is not an exact integer as would be expected for a sample size of 100.

The results in Table 4 (the no deletion case) show a poor performance in estimating the steady-state mean for all models with the exception of model 3, the M/M/1 model. The results show that it is only for model 3 that the 95% confidence interval for the coverage includes 90%, as would be expected. For models 1 and 4, the confidence interval for the steady-state mean never covers the actual mean value in the experiments performed. Note that for all the models there is a negative bias, although the confidence intervals for both the M/M/1 model runs cover zero.

When applying the SPC method the results are, in general, much improved. There is some variation in the deletion point (warm-up period) identified, with the confidence interval being around ±15% of the mean truncation point for some of the results (Table 5). When the results are compared to Table 4 there is a clear improvement in

Table 4
Results for test models: no data deleted

Model	Estimated mean	Bias	Coverage of 90% CIs	Expected half length of 90% CIs
AR(1) Normal a	−0.0385 ± 0.0005	−0.0385 ± 0.0005	0.0 + 0.03[a]	0.0056 ± 0.0004
AR(1) Normal b	−0.0191 ± 0.0004	−0.0191 ± 0.0004	0.0 + 0.03	0.0041 ± 0.0003
AR(1) Exp. a	9.5463 ± 0.0529	−0.4537 ± 2.3954	67.0 ± 9.3	0.5748 ± 0.0414
AR(1) Exp. b	9.7907 ± 0.0395	−0.2093 ± 0.0395	79.0 ± 8.1	0.4317 ± 0.0291
M/M/1 $\rho = 0.8$ a	3.9264 ± 0.0901	−0.0736 ± 0.0901	89.0 ± 6.2	0.8619 ± 0.0903
M/M/1 $\rho = 0.8$ b	3.9755 ± 0.0625	−0.0245 ± 0.0625	90.0 ± 6.0	0.6500 ± 0.0589
GSS M/M/1	1.8275 ± 0.0065	−0.1725 ± 0.0065	0.0 + 0.06	0.0516 ± 0.0054

[a] When the number of successes (x) out of a set of trials (n) is small, it is appropriate to calculate an upper bound instead of a confidence interval as follows $\frac{1}{2n}\chi^2_{2(x+1),\alpha}$ (Miller et al., 1990).

Table 5
Results for test models: data deleted using SPC method

Model	Warm-up est.	Deletion point	Estimated mean	Bias	Coverage of 90% CIs	Expected half length of 90% Cis	% Bias > from deletion
AR(1) Normal a	89	77.0 ± 8.1	−0.0003 ± 0.0007	−0.0003 ± 0.0007	91.0 ± 6.0	0.0071 ± 0.0006	0.0 + 0.03[a]
AR(1) Normal b	96	97.1 ± 14.7	0.0002 ± 0.0005	0.0002 ± 0.0005	87.5 ± 6.7	0.0046 ± 0.0003	0.0 + 0.03
AR(1) Exp. a	83	68.6 ± 9.8	10.0123 ± 0.0771	0.0123 ± 0.1086	89.2 ± 6.8	0.6969 ± 0.0492	33.7 ± 10.3
AR(1) Exp. b	98	91.2 ± 14.8	10.0230 ± 0.0533	0.0230 ± 0.0533	87.8 ± 6.6	0.4814 ± 0.0334	41.8 ± 9.9
M/M/1 $\rho = 0.8$ a	88	501.7 ± 89.4	4.0446 ± 0.1460	0.0446 ± 0.1460	86.4 ± 7.3	1.0580 ± 0.1951	51.1 ± 10.6
M/M/1 $\rho = 0.8$ b	90	1006.4 ± 164.6	3.9759 ± 0.0888	−0.0241 ± 0.0888	84.4 ± 7.6	0.7000 ± 0.0908	56.7 ± 10.4
GSS M/M/1	50	3159.1 ± 211.4	1.9968 ± 0.0092	−0.0032 ± 0.0092	90.0 ± 8.5	0.0046 ± 0.0003	0.0 + 0.06

[a] When the number of successes (x) out of a set of trials (n) is small, it is appropriate to calculate an upper bound instead of a confidence interval as follows $\frac{1}{2n}\chi^2_{2(x+1),\alpha}$ (Miller et al., 1990).

the estimate of the mean, with the absolute mean bias being reduced for every model. The coverage is approximately as expected (90%) for all models, and in each case the confidence interval of the coverage includes 90%. The expected half-length of the confidence interval for the mean is greater for all models when the data are deleted. This is a consequence of having a smaller sample size on which to calculate the confidence interval. On the other hand, the estimated mean is more accurate.

By using the SPC method the bias is reduced in all cases for models 1 and 4 (see '% bias > from deletion'). The bias is reduced in the majority of cases for model 2. For model 3, however, the bias is reduced in about 50% of cases and increased in the other 50%. This is not surprising since there is little bias in the M/M/1 data when no data are deleted. Note that on average the absolute bias is reduced for model 3, although only marginally for the run length of 4000.

For some datasets the SPC method was not able to identify a warm-up period. This happened because a suitable batch size could not be found to reduce the autocorrelation in the data sufficiently or to ensure that the data are normally distributed. A longer run-length would probably overcome this problem. Indeed, the results show a greater level of success when the run-length is increased. On one occasion a warm-up period could not be identified because the data did not meet the criteria for being in-control. This also implies that a longer run-length is required.

The results in Tables 4 and 5 also indicate the effect of increasing the run-length for models 1 to 3, in this case, doubling it. For the case where no data are deleted, the bias and expected half length are both reduced. The coverage improves for model 2, with little or no effect for models 1 and 3. With the SPC method employed the expected half length falls when the run-length is increased. The mean bias only improves for models 1 and 3. Although the mean bias is greater with an increased run-length for model 2, the confidence interval is much narrower, indicating that it may have improved. For all models the coverage is slightly reduced with an increased run-length. This is assumed to be because the greater quantity of data narrows the confidence interval for the steady-state mean. The deletion point becomes greater with more data, and is approximately doubled for model 3. This seems to be mainly a result of using larger batch sizes when more data are available. With fewer data there are less grounds on which to reject the tests for autocorrelation and normality. As such, with more data there is a tendency to have larger batch sizes. Note that when the run-length is doubled, since the tests are performed on the second half of the time-series, they are performed on a completely different set of data.

In summary, employing the SPC method improves the estimate of the steady-state mean over the case where no data are deleted. With the SPC method the coverage of the mean is as expected. This suggests that the method is providing valid results and so it is identifying a satisfactory warm-up period. An increased run-length gives a slight improvement in the estimate of the steady-state mean, but the SPC method does suggest that more data need to be deleted as the run-length is increased. The method fails to identify a warm-up period for a number of cases because the requirements for autocorrelation and normality are not met. In these cases a greater run-length would probably mean that a warm-up period could be identified.

6. Discussion

When evaluating a method for estimating the warm-up period, consideration should be given to more than its accuracy. As already stated, although many methods for estimating the warm-up period have been proposed, it appears that very few are used in practice. Any evaluation of a method should pay attention, therefore, to its potential for adoption. Gafarian et al. (1978) use various criteria in their comparison of methods for detecting the warm-up period. Those criteria, with some amendments, are used here:

- *Accuracy in estimating the warm-up period.* Is the initialisation bias detected and removed?
- *Simplicity.* Is the approach accessible to the average practitioner? Considerations include

the requirement for complex computations and the transparency of the approach.
- *Ease of implementation.* Can the approach easily be implemented? Is it possible to automate the procedure?
- *Assumptions and generality in their use.* Is the method restricted to specific circumstances or can it be relied upon in a wide range of situations?
- *Estimation of the parameters.* Do many parameters need to be estimated? If so, the efficacy of the approach may be compromised.

The tests demonstrate the *accuracy* of the SPC method. The coverage of the confidence intervals for the steady-state means are all at the level expected. One concern with the accuracy of the method is the use of small sample sizes (as small as 10) for the statistical tests and the estimation of the control chart parameters. This could be mitigated by enforcing a larger minimum number of batches, but this would require longer run-lengths for some models.

The SPC method has a significant advantage in terms of *simplicity*. This is a method in common use for quality control in industry, and therefore many simulation users will be familiar with the underlying concept. On the down side, the need to test for autocorrelation and normality complicate the approach. The method can, however, be *implemented* with relative ease, for instance, in a spreadsheet. The use of the SPC rules mean that the procedure could also be automated, although there should be some care in this respect, since interpretation is required to guard against, for instance, a control limit being broken when this simply represents a Type I error. Indeed, a strength of the method is that it maintains a view of the time-series, enabling the user to interpret the data while being guided by the SPC rules.

The key *assumptions* behind the SPC method are the need for normality and low autocorrelation. Albeit that these are aided by the batching of data and use of multiple replications, these assumptions may become restrictive in some circumstances. Indeed, these assumptions could not be met with the data available for some of the experiments performed in the tests, leading to a failure to identify a warm-up period. It is probable that the assumptions could have been met by running the models for longer, enabling larger batch sizes to be generated. Rather than use batch means for dealing with non-normality and autocorrelation, alternative methods for constructing control charts should be investigated. Beyond Runger and Willemain (1996) approach various methods have been proposed, for instance, averaging subgroups of data (this is similar to the batch means method) (Alwan and Radson, 1992), exponentially weighted moving average charts (MacCarthy and Wasusri, 2001; Lu and Reynolds, 1999), a CUSUM based scheme (Atienza et al., 2002) and a procedure based on the detection of outliers and level shifts (Atienza et al., 1998). The batch means method is preferred here because it employs a control chart approach that is likely to be familiar to most simulation users.

Finally, the method does require the estimation of five *parameters*, namely the number of replications, the run-length, the batch size, and the mean and variance of the data. The normality and autocorrelation criteria guide the user over the first three parameters, requiring that one or more is increased until both criteria are met. The mean and variance are estimated from the second half of the series, an approach that is used in other methods, for instance, the optimal test (Schruben et al., 1983). This, of course, does not guarantee an accurate estimate of the mean and variance, especially if the data have not settled to a steady-state or they demonstrate some unusual pattern in this part of the time-series. Indeed, one of the main reasons why the method failed to cover the actual mean was because of a poor estimate of the steady-state mean from the second half of the data.

7. Conclusions

A new approach for determining the warm-up period of a steady-state simulation is described. This is based upon statistical process control theory. The method requires the estimation of the SPC parameters, the inspection of a control chart and the interpretation of the chart in the light of

standard SPC rules for being in-control. As such it can be described as a hybrid of a graphical and a heuristic method.

Further work should investigate the effect of the number of replications and run-length on the results obtained from the SPC method. There is obviously a relationship between these and the batch size required. What is the nature of this relationship? Other SPC methods that deal with non-normality and autocorrelated data should also be explored. Could these methods be employed to better effect, particularly when the simulation output is problematic in these respects?

Acknowledgments

The author thanks Agisilaos Ioannou for helping to develop the analysis software used for this paper, and for his help in searching the literature. Also thank you to various colleagues for their advice.

The description of the SPC method is based on the papers:

Robinson, S., 2002. Selecting a Warm-up Period: A Statistical Process Control Approach. Operational Research Society, Simulation Workshop, March 2002. Operational Research Society, Birmingham, UK, pp. 69–78.

Robinson, S., 2002. A statistical process control approach for estimating the warm-up period. In: Yucesan, E., Chen, C.-H., Snowden, J.L., Charnes, J.M. (Eds.), Proceeding of the 2002 Winter Simulation Conference. Institute of Electrical and Electronic Engineers, Piscataway, NJ, pp. 439–446.

References

Alexopoulos, C., Seila, A.F., 1998. Output data analysis. In: Banks, J. (Ed.), Handbook of Simulation. Wiley, New York, pp. 225–272.

Alwan, L.C., Radson, D., 1992. Time-series investigation of subsample mean charts. IEE Transactions 24, 66–80.

Anderson, T.W., Darling, D.A., 1954. A test of goodness of fit. Journal of the American Statistical Association 49, 765–769.

Atienza, O.O., Tang, L.C., Ang, B.W., 1998. A SPC procedure for detecting level shifts of autocorrelated processes. Journal of Quality Technology 30 (4), 340–351.

Atienza, O.O., Tang, L.C., Ang, B.W., 2002. A CUSUM scheme for autocorrelated observations. Journal of Quality Technology 34 (2), 187–199.

Banks, J., Carson, J.S., Nelson, B.L., Nicol, D.M., 2001. Discrete-Event System Simulation, third ed. Prentice-Hall, Upper Saddle River, NJ.

Cash, C.R., Nelson, B.L., Dippold, D.G., Long, J.M., Pollard, W.P., 1992. Evaluation of tests for initial-condition bias. In: Swain, J.J., Goldsman, D., Crain, R.C., Wilson, J.R. (Eds.), Proceedings of the 1992 Winter Simulation Conference. IEEE, Piscataway NJ, pp. 577–585.

Conway, R.W., 1963. Some tactical problems in digital simulation. Management Science 10, 47–61.

D'Agostino, R.B., Stephens, M.A., 1986. Goodness-of-fit Techniques. Marcel Dekker, New York.

Fishman, G.S., 1971. Estimating sample size in computing simulation experiments. Management Science 18 (1), 21–38.

Fishman, G.S., 1973. Concepts and Methods in Discrete Event Digital Simulation. Wiley, New York.

Fishman, G.S., 1978. Grouping observations in digital simulation. Management Science 24, 510–521.

Fishman, G.S., 1996. Monte Carlo: Concepts, Algorithms, and Applications. Springer-Verlag, New York.

Gafarian, A.V., Ancker, C.J., Morisaku, T., 1978. Evaluation of commonly used rules for detecting "Steady State" in computer simulation. Naval Research Logistics Quarterly 25, 511–529.

Goldsman, D., Schruben, L.W., Swain, J.J., 1994. Tests for transient means in simulated time series. Naval Research Logistics 41, 171–187.

Gordon, G., 1969. System Simulation. Prentice-Hall, Upper Saddle River, NJ.

Jackway, P.T., DeSilva, B.M., 1992. A methodology for initialization bias reduction in computer simulation output. Asia-Pacific Journal of Operational Research 9, 87–100.

Kelton, W.D., Law, A.M., 1983. A new approach for dealing with the startup problem in discrete event simulation. Naval Research Logistics Quarterly 30, 641–658.

Law, A.M., Kelton, W.D., 2000. Simulation Modeling and Analysis, third ed. McGraw-Hill, New York.

Lu, C.-W., Reynolds, M.R., 1999. EWMA control charts for monitoring the mean of autocorrelated processes. Journal of Quality Technology 31 (2), 166–188.

MacCarthy, B.L., Wasusri, T., 2001. Statistical process control for monitoring scheduling performance—addressing the problem of correlated data. Journal of the Operational Research Society 52 (7), 810–820.

Miller, I., Freund, J.E., Johnson, R.A., 1990. Probability and Statistics for Engineers, fourth ed. Prentice-Hall, London.

Montgomery, D.C., Runger, G.C., 1994. Applied Statistics and Probability for Engineers. Wiley, New York.

Nelson, B.L., 1992. Statistical analysis of simulation results. In: Salvendy, G. (Ed.), Handbook of Industrial Engineering, second ed. Wiley, New York (Chapter 102).

Pawlikowski, K., 1990. Steady-state simulation of queueing processes: A survey of problems and solutions. Computing Surveys 22 (2), 123–170.

Robinson, S., 2004. Simulation: The Practice of Model Development and Use. Wiley, Chichester, UK.

Roth, E., 1994. The relaxation time heuristic for the initial transient problem in M/M/k queueing systems. European Journal of Operational Research 72, 376–386.

Runger, R.C., Willemain, T.R., 1996. Batch-means control charts for autocorrelated data. IIE Transactions 28, 483–487.

Schruben, L.W., 1982. Detecting initialization bias in simulation output. Operations Research 30 (3), 569–590.

Schruben, L., Singh, H., Tierney, L., 1983. Optimal tests for initialization bias in simulation output. Operations Research 31 (6), 1167–1178.

Vassilacopoulos, G., 1989. Testing for initialization bias in simulation output. Simulation 52 (4), 151–153.

Von Neumann, J., 1941. Distribution of the ratio of the mean square successive difference and the variance. Annals of Mathematical Statistics 12, 367–395.

Welch, P., 1983. The statistical analysis of simulation results. In: Lavenberg, S. (Ed.), The Computer Performance Modeling Handbook. Academic Press, New York, pp. 268–328.

White, K.P., 1997. An effective truncation heuristic for bias reduction in simulation output. Simulation 69 (6), 323–334.

White, K.P., Cobb, M.J., Spratt, S.C., 2000. A comparison of five steady-state truncation heuristics for simulation. In: Joines, J.A., Barton, R.R., Kang, K., Fishwick, P.A. (Eds.), Proceedings of the 2000 Winter Simulation Conference. IEEE, Piscataway, NJ, pp. 755–760.

Wilson, J.R., Pritsker, A.B., 1978. A survey of research on the simulation startup problem. Simulation 31, 55–59.

Winston, W.L., 1994. Operations Research: Applications and Algorithms, third ed. Duxbury Press, Belmont, CA.

Yücesan, E., 1993. Randomization tests for initialization bias in simulation output. Naval Research Logistics 40, 643–663.