Automating warm-up length estimation

AUTOMATING WARM-UP LENGTH ESTIMATION

Kathryn Hoad
Stewart Robinson
Ruth Davies

Warwick Business School
The University of Warwick
Coventry, UK

Abstract

There are two key issues in assuring the accuracy of estimates of performance obtained from a simulation model. The first is the removal of any initialisation bias, the second is ensuring that enough output data is produced to obtain an accurate estimate of performance. This paper is concerned with the first issue, and more specifically warm-up estimation. Our aim is to produce an automated procedure, for inclusion into commercial simulation software, for estimating the length of warm-up and hence removing initialisation bias from simulation output data. This paper describes the extensive literature search that was carried out in order to find and assess the various existing warm-up methods, the process of short-listing and testing of candidate methods. In particular it details the extensive testing of the warm-up MSER-5 method.

Keywords

Simulation, Warm-up period, initialisation bias, truncation point, automation, MSER-5

1 INTRODUCTION

Initialisation bias occurs when a model is started in an 'unrealistic' state. The output data collected during the warming-up period of a simulation can be misleading and bias the estimated response measure. The removal of initialisation bias is, therefore, important for obtaining accurate estimates of model performance.

There are five main methods for dealing with initialisation bias (Robinson, 2004):

1. Run-in model for a warm-up period until it reaches a realistic condition (steady-state for non-terminating simulations) and delete data collected from the warm-up period.
2. Set initial conditions in the model so that the simulation starts in a realistic condition.
3. Set partial initial conditions then warm-up the model and delete warm-up data.
4. Run the model for a very long time making the bias effect negligible.
5. Estimate the steady-state parameters from a short transient simulation run (Sheth-Voss et al., 2005).

This paper is concerned with the first method; deletion of the data with initial bias by specifying a warm-up period (or truncation point). The overall aim of the work is to create an automated procedure for determining an appropriate warm-up period that could be used by non (statistically)-expert users and could be included in commercial simulation software. Section 2 describes the extensive literature review that was carried out to find the various warm-up methods in existence. The next section describes the short-listing procedure used to assess these warm-up methods and the results. Section 4 describes the preliminary testing of the short-listed methods. A brief description of the successful short-listed candidate method, MSER-5, can be found in Section 5. Section 6 describes the extensive testing of this method and section 7 provides the conclusions of the work.

2 LITERATURE REVIEW

An extensive literature review of warm-up methods was carried out in order to collect as many published methods and reviews of such methods as possible. Through the literature search we found 44 warm-up methods. Each method was categorised into one of 5 main types of procedure as described by Robinson (2004):

1. Graphical methods – Truncation methods that involve visual inspection of the time-series output and human judgement.

2. Heuristic approaches – Truncation methods that provide (simple) rules for determining when to truncate the data series, with few underlying assumptions.
3. Statistical methods – Truncation methods that are based upon statistical principles.
4. Initialisation bias tests – Tests for whether there is any initialisation bias in the data. They are therefore not strictly methods for obtaining the truncation point but they can be adapted to do so in an iterative manner.
5. Hybrid methods – A combination of initialisation bias tests with truncation methods in order to determine the warm-up period.

A list of these methods and relevant references is provided in Table 1. More detailed information and a summary of each method can be found at the project website: <www.wbs.ac.uk/go/autosimoa/warmup/>

< Table 1 about here>

3 SHORT LISTING WARM-UP METHODS FOR AUTOMATION

The ultimate aim was to find a method that could automatically detect the warm-up period with minimum user intervention and so would be suitable for automation. Due to the large number of methods found it was not feasible to test them all ourselves. It was therefore necessary to whittle down the number of methods to a short list of likely candidates that could then proceed to testing.

3.1 Short Listing Methodology

We decided to grade all the methods, based on what was reported in the literature about each approach, using 4 main criteria:

- *Accuracy and robustness* of the method - i.e. how well the method truncates allowing accurate estimation of the true mean.
- *'Ease' of automation* potential – e.g. methods requiring significant user interaction / judgment are not easily automatable; a large number of parameters to estimate could also hinder the applicability of a method for automation.

- *Generality* - i.e. does a method work well with a large range of initial bias and data output types.
- *Computer time taken* - Ideally we want the analysis method running time to be negligible compared with the running time of the simulation.

In order to arrive at a shortlist, we first rejected methods that failed the 'ease of automation' criterion as this was seen as most important for our purposes. From the methods left we rejected those that had poor reported accuracy and robustness. We then rejected further methods on the grounds of non-generality and excessive computer time taken. We also rejected 'first draft' methods that had been subsequently usurped by improved versions (e.g. MCR by MSER-5).

The remaining methods could then be tested by ourselves and rejected or not rejected accordingly. The aim was to end up with one or more methods that function well according to all our criteria.

3.2 Results of Short Listing

All of the methods have shortcomings and suffer from a lack of consistent, comparable testing across the literature. Key problems are overestimation and underestimation of the truncation point, reliance on restrictive assumptions and the requirement to estimation of a large number of parameters. A graphical summary of the results of this short listing procedure for graphical, statistical and heuristic approaches can be seen in figure 1. The 6 warm-up methods successfully short-listed to go forward to further testing are listed in table 2.

< Figure 1 about here>

The graphical methods were rejected mainly on grounds of ease of automation (since they require user intervention) and accuracy. For instance, Welch's method, one of the more popular warm-up methods, (Law, 1983) requires a user to judge the smoothness and flatness of a moving average plot; this would be difficult to automate. Furthermore, many graphical methods use cumulative statistics which react slowly to changes in system status. Cumulative averages tend to converge more slowly to a

steady-state than do ensemble averages (Wilson and Pritsker 1978a) which frequently leads to overestimation of the truncation point.

The majority of statistical methods were rejected on grounds of ease of automation, generality or accuracy. For instance, the Kelton and Law regression method is criticised in the literature for being complex to code (Kimbler and Knight 1987). This is partially due to the large number of parameters that require estimation.

The majority of heuristic methods were rejected on the grounds of accuracy, generality and ease of automation. For example, the crossing-of-the-mean rule (Fishman 1973, Wilson and Pritsker 1978a, 1978b) was heavily criticised in the literature for being extremely sensitive to the selection of its main parameter, which was system-dependent, and misspecification of which caused significant over or under-estimation of the warm-up length (Pawlikowski 1990). This method was therefore rejected on ease of automation and accuracy grounds.

The initialisation bias tests were considered separately. It was found that when used iteratively in an automated fashion they gave inconsistent and confusing results, i.e. they would switch from detecting bias to not detecting bias multiple times. This was not at all desirable or helpful in an automated system aimed at non-expert users. Therefore, we rejected the 9 initialisation bias tests and the 2 hybrid methods that incorporated them. These were therefore not included in the shortlisting in figure 1.

< Table 2 about here>

4. PRELIMINARY TESTING OF SHORT LISTED METHODS

The short-listed methods were tested using artificial data. The benefits of using artificial data are that they are completely controllable with known testable characteristics such as the mean and L (point at which the initial bias ends). The methods were tested on simple data sets first, i.e. little to no auto-correlation, normal errors, a bias length of between 10% to 50% of the total data length *n* and a mean-shift or linear shaped bias (table 4). The bias was made severe enough to be easily

identifiable by eye. If a method failed to provide adequate results for these very simple data sets then testing was stopped and the method rejected.

4.3 Results of preliminary testing

The ASD and ADD methods required a very large number of replications which was deemed unsatisfactory for our purposes. Both the goodness-of-fit method and Kimbler's double exponential smoothing method consistently and severely underestimated the truncation point (see table 3) and were therefore rejected. The Euclidean distance method failed to return any result on the majority of occasions and was therefore rejected also. MSER-5 gave the most accurate estimates of the true truncation point L without requiring excessive amounts of data (see table 3).

In general the sequential methods assume a monotonic decreasing or increasing bias function and therefore do not cope with the mean shift bias. Methods that analyse all the data given (in one go), using all the information that all the data provides, seem more able to cope with a larger variety of bias types and seem more suited to automation. From the preliminary results obtained, the MSER-5 truncation method performed the best and the most consistently. The MSER-5 method was therefore considered a promising, indeed the only, candidate for automation and further rigorous testing of this method was carried out.

< Table 3 about here>

5 THE MSER-5 WARM-UP METHOD

In the paper that first introduces the MSER (or MCR) method, White (1997) explains the method as follows: "Instead of selecting a truncation point to minimise the MSE, we propose to select a truncation point that minimises the width of the CI about the truncated sample mean …Thus we will seek to mitigate bias by removing initial observations that are far from the sample mean, but only to the extent this distance is sufficient to compensate for the resulting reduction in sample size in the calculation of the confidence interval half width."

Formally, given a finite stochastic sequence of output i of replication j $\{Y_i(j): i=1,2,\ldots,n\}$, the optimal truncation point for this data series is defined as (Linton and Harmonosky, 2002; White Jnr, 1997):

$$d(j)^* = \underset{n>d(j)\geq 0}{\arg\min}\left[\frac{z_{\alpha/2}s(d(j))}{\sqrt{n(j)-d(j)}}\right],$$

where $z_{\alpha/2}$ is the value of the N(0,1) distribution associated with a 100(1-α)% confidence interval and *s(d(j))* is the sample standard deviation of the reserved sequence (i.e. of all data following *d(j), where d(j) is all possible truncation points for replication j*), and *n(j)* is the total number of observations in replication *j*. Since the confidence level α is fixed, $z_{\alpha/2}$ is a constant and can therefore be set arbitrarily to 1, as the purpose of using the above equation is only to compare all data points to find the minimum.

The expression for the optimal truncation point can therefore be written explicitly in terms of the output data points:

$$d(j)^* = \underset{n>d(j)\geq 0}{\arg\min}\left[\frac{1}{(n(j)-d(j))^2}\sum_{i=d+1}^{n}\left(Y_i(j)-\bar{Y}_{n,d}(j)\right)^2\right] \quad (1)$$

MSER-m (unlike MCR or MSER) applies equation (1) to a series of $b=\lfloor n/m \rfloor$ batch averages instead of to the raw output data series. MSER-5 is therefore the MSER-m method using batches of 5 data points. Figure 2 shows a working example of the MSER-5 method.

< Figure 2 about here>

From herein the truncation point returned by MSER-5 will be referred to as the *Lsol* value. Any *Lsol* value > n/2 is rejected, as in these cases it is possible that the method has not been provided with enough data to produce a valid result. This would occur if the transient period extends into the 2nd half of the data, the data has not reached steady-state or more data is required because of the high auto-correlation of the data.

In practice, if this occurs, more data could be produced and MSER-5 run again with the extended data set until a valid *Lsol* value was returned.

MSER-5 can sometimes erroneously report a truncation point towards the end of the data series (from here on referred to as an 'end point' *Lsol* value). This is because the method can be overly sensitive to observations at the end of the data series that are close in value (Delaney 1995, Spratt 1998). This is an artefact of the point at which the simulation is terminated (Spratt 1998). This can be avoided most of the time by not allowing the algorithm to consider the standard errors calculated from the last few data points, (we have chosen a default value of 5 points), although this does not completely eradicate the problem. If, however, we reject any truncation point that falls into the second half of the data and simply rerun the algorithm with more data, this almost completely eliminates this problem.

6 FURTHER TESTING OF MSER-5

MSER-5 was tested further on a larger range of artificial data sets and the results analysed by graphical and statistical methods.

6.1 Creation of the artificial data sets.

The artificial data sets were created in two parts: the initial bias functions, a_t, and the steady-state functions X_t (where t = time). These two parts were then combined by superposition (Spratt 1998).

We had previously created a representative and sufficient set of model output data by analysing over 50 'real' models / output and identifying a set of important characteristics (for full details see Hoad, Robinson & Davies 2006 at http://www2.warwick.ac.uk/fac/soc/wbs/projects/autosimoa). By studying the initial transients and steady-state data in this collection and reviewing the literature we decided upon four criteria that would completely specify the bias function a_t: length, severity, shape and orientation; and three criteria to define our steady-state functions: the variance, error terms and the type of auto-correlation of the data.

The bias function, a_t, criteria were:

- The length of the initial bias, L, is described in terms of the percentage of the total data length, *n*. The total data length, *n*, was set at 1000.
- The severity of the initial bias is described by its maximum value. In order to control the severity we let Max $| a_t |_{t \leq L}$ (the maximum value that the bias function, a_t, can take) = M × Q. M is the relative maximum bias value set by us. Q is the difference between the steady-state mean and the 1st (if bias function is positive) or 99th (if bias function is negative) percentile of the steady-state data. If M is set to be greater than 1 then we would expect the bias to be significantly separate from the steady-state data and therefore easier to detect. Likewise, if M is set to a value less than 1 we would expect the bias to be absorbed into the steady-state data and therefore be far harder to detect.
- The shapes of the bias functions were taken from the literature (Cash et al. 1992, Spratt 1998, White et al. 2000) and knowledge of 'real model' warm-up periods. The 7 main shapes used are shown in Table 4 along with their respective mathematical functions.
- There are two possible bias directions: a positive bias is where the biased data starts above the steady-state mean and a negative bias is where the biased data starts below the steady-state mean.

The steady-state function criteria were:

- As these were steady-state data that we were creating the variance was kept constant.
- The error terms, ε_t, are either normally or non-normally (exponentially) distributed. The L'Ecuyer Random Number Generator (L'Ecuyer 1999, Law 2007) is used to generate all the random numbers required.
- The steady-state functions either have no auto-correlation in which case the steady-state data are simply made up by the error term, or have varying complexity of auto-correlation. The actual autoregressive or moving average functions and parameter values were chosen in order to give a varying degree and complexity of correlation with a range of oscillatory/decay behaviour (Box et al. 1994). The equations and parameter values used to create the 12 different steady-state functions are shown in table 5. As desired, it is possible

to mathematically calculate the true mean values for each of the steady-state functions. The equations used are shown in table 6.

The bias functions were incorporated into the steady-state functions by superposition (Spratt 1998). This adds the bias function onto the end of the steady-state function, X_t, to produce the finished data Y_t. For example, for the AR(1) function with parameter ϕ:

$$X_t = \phi X_{t-1} + \varepsilon_t$$
$$Y_t = X_t + a_t$$
etc...

There is therefore no 'run-in' period and no lag between the end of the bias function and the start of the steady-state period. Hence we know precisely the true truncation point and have complete control over the shape and severity of the bias.

< Table 4 about here>
< Table 5 about here>
< Table 6 about here>

6.2 Experimental design

The data sets were either created using single runs or by averaging over 5 replications in order to test the benefit of the smoothing effect of multiple replications. Therefore, in summary, we used 7 parameters to create our artificial data: bias length, severity, shape and orientation, error type, auto-correlation type and single run or replications. It was thought that some or all these parameters would impact on the warm-up method's effectiveness. The levels at which we set the 7 parameters are detailed in table 7. A full factorial design was used leading to 2016 separate sets of artificial data. Further to this we produced another 1032 data sets with no bias or 100% bias in order to test the ability of MSER-5 to correctly identify these circumstances.

6.3 Performance Criteria

What we want to know is whether the warm-up method, MSER-5, is effective. But, what do we mean by effective? And, can we judge this accurately? Existing literature

predominantly uses performance measures that fall into two categories: accuracy of mean measures and accuracy of L measures. Using the literature as a guide (Kelton and Law 1983, Robinson 2005, Spratt 1998) we selected / created the following performance criteria to assess the effectiveness of the chosen warm-up method, MSER-5. The method was run with each type of artificial data set 100 times to allow for the statistical analysis of the results.

1. *Coverage of the true mean*: Ideally, the true mean should fall within the confidence interval around the average of the truncated means. This criterion was also calculated for the data series without truncation for comparison purposes.
2. *Closeness of estimated truncation point (Lsol) to actual L*. This indicates consistent underestimation or overestimation of the true end of the initial bias.

   < Table 7 about here>

   Because of the different shapes and severity of the initial bias functions used in testing, truncating all the functions at some point x prior to the correct value of L would eradicate different amounts of bias from the data sets. It is therefore unclear from just the *Lsol* values how effective MSER-5 has been at removing the initial bias in each case. We therefore decided to calculate the amount of bias that would be removed by truncating each data set at its *Lsol* value.
3. *Percentage bias removed by truncation*: As explained above, the different shapes and severity of the initial bias functions used in testing, causes different amounts of bias to be removed from the data when truncating at the same point. A calculation of the percentage of bias that would be removed by truncating each data set at its *Lsol* value would therefore give a clearer idea of how effective MSER-5 is at identifying initial bias. It is desirable that the method removes a high percentage, ideally 100% without removing much (if any) steady-state data (i.e. *Lsol* > L).

   We calculated the percentage bias removed by determining the area under the bias function that would be deleted by truncating at *Lsol* and comparing this

with the whole area under the bias function. For example, data set '*X*' with data $x_1, x_2, x_3, \ldots, x_n$, is made from adding the bias function data $b_1, b_2, \ldots, b_n$ to the steady-state data $c_1, c_2, \ldots, c_n$ by superposition. (The bias data is different for each data set used in testing.) The true bias truncation point is known to be at $x_{L\ (0< L< n/2)}$ and the MSER-5 method determined a truncation point of $x_{Lsol\ (Lsol \le n/2)}$. Therefore the percentage bias removed by truncating data set '*X*' at *Lsol* is:

$$\begin{cases} = 100\left[\dfrac{\sum_{i=1}^{Lsol}(x_i - c_i)}{\sum_{i=1}^{L}(x_i - c_i)}\right] = 100\left[\dfrac{\sum_{i=1}^{Lsol} b_i}{\sum_{i=1}^{L} b_i}\right] & , for\ Lsol \le L \\ = 100\% & , for\ Lsol > L, \end{cases}$$

All cases where *Lsol* > L will be said to fall into the '100+' category, for the purposes of discussion and analysis.

4. *Analysis of the pattern and frequency of rejections of the estimated truncation point, Lsol, due to insufficient data*: *Lsol* was not accepted if it fell into the second half of the data, (i.e. *Lsol* > *n*/2), as this was assumed to indicate that there were insufficient data for the method to provide a robust estimate of L (see section 5 for full explanation).

It was also presumed that the various parameters used to create the artificial test data would affect the functioning of the warm-up method. We therefore analyse the effect of each parameter separately, as well as the interaction effects between the 7 parameters, upon the performance of the warm-up method as reflected by the performance criteria. Both graphical analysis and chi-squared testing of results were employed. Because the algorithm was run only once rather than in an iterative fashion, all results quoted, unless specifically identified otherwise, refer only to the valid runs i.e. *Lsol* values returned in the first half of the data. Section 6.5 details the results regarding the rejected *Lsol* values with a discussion on when, and possible reasons why, this occurs.

6.4 Results of testing the MSER-5 heuristic

6.4.1 Coverage of the true mean

The reason for truncating and thus eradicating any initial bias is to provide a non-biased estimate of the true mean of the data. Therefore one way to judge how well MSER-5 is working is to calculate the mean (with 95% confidence intervals) of all data series (where MSER-5 returned a *Lsol* value in the first half of the data) after they have been truncated at their respective *Lsol* values. The test criterion is then whether or not the true known mean of the data falls within these confidence intervals. Due to the nature of the artificial data it is also important to check whether the true mean falls into the equivalent confidence intervals for the non-truncated data. There are therefore 4 possible combinations of results as shown in table 8.

The results in table 8 do not include the averaged ARMA(5,5) data with exponential error as these data sets, when truncated at correct L, did not include the true mean within their confidence intervals and were therefore deemed not to be a fair test of MSER-5. Because of the high auto-correlation of this data type it would be necessary to have more data than only n=1000, in order to achieve a representative estimate of the true mean.

< Table 8 about here>

These coverage results appear to be very good with approximately 80% of the test data covering the true mean after truncation. Of course, as the impact of residual bias is dependent on the run-length of the data beyond the truncation point, all the data sets could be made to fall into this category by running more data after truncation.

6.4.2 L versus *Lsol*

For each true truncation point L, MSER-5 gave a wide range of *Lsol* values (see figure 3 for examples of 2 data sets). It was noted that as the severity of decline in the bias increases the number and severity of underestimations of the warm-up period increases, e.g. the most underestimation occurs in data with exponentially declining

bias. However, judging MSER-5 on *Lsol* values alone is misleading. For example, table 9 shows the mean *Lsol* values with 95% confidence intervals for the two data sets featured in figure 3. Both confidence intervals do not include the true value of L (100) for those sets, and therefore this criterion alone suggests that the MSER-5 method is performing poorly. But for the mean-shift bias data, MSER-5 accurately estimated L on 72% of the runs and only overestimated L to a maximum of 45 (average of 14) data points on the remaining 28% of runs. It would therefore appear that MSER-5 was actually quite accurate in estimating the true warm-up period for this data. The quadratic bias example shows a general underestimation of true L due to the declining nature of the initial bias, but it is unclear how detrimental this underestimation is. The percentage of bias removed by truncation is seen as a more useful measure of the effectiveness of a truncation method (especially regarding underestimation of the truncation point).

< Figure 3 about here>
< Table 9 about here>

6.4.3 Percentage bias removed by truncation

It can be seen in figure 4 that in over 64% of valid runs, (92.6% of the total 201,600 runs were deemed valid), the MSER-5 method removed at least 95% of the bias from the data, and in over 77% of valid runs it removed at least 90% of the bias.

< Figure 4 about here>

Further analysis was then performed to understand the effect of the various parameters used to create the artificial data sets on the functioning of MSER-5. We analysed the main effect of each parameter and any interaction effects between the 7 parameters. Looking at the percentage bias removed results (excluding those for L = 0% and 100%) in more detail, the following observations can be made.

Error type

The type of error (normal or exponential) did not significantly affect the percentage of bias removed. A chi-squared test of the two percentage bias removed distributions gave a p-value of 0.85 indicating no significant difference at the 95% level.

Auto-correlation function type

The stronger the auto-correlation in the data the more difficulty MSER-5 had in accurately removing initial bias. Figure 5 shows the cumulative percentage bias removed for each autocorrelation data type. The closer the cumulative line is to the right of the graph, the better the performance. A chi-squared test of these distributions gave a p-value of 0.000 indicating a significant difference at the 95% level. Three quarters of the cases where less than 40% of the bias was removed were from data sets with very high auto-correlation. This effect was greatly reduced by using averaged data rather than single run data. Figure 5 clearly shows the difference between the ARMA(5,5) results and the rest. The underlying equations help to explain these differences. ARMA(5,5) in particular has an extremely long lag in the auto-correlation with the influence of the first value becoming non significant (at the 5% significance level) at around the 30^{th} value. AR(1) performs nearly as poorly. Because of its large parameter value (0.9) the influence of the first value only becomes non significant at around the 15^{th} value. It should also be noted of the ARMA(5,5) data that due to its high auto-correlation a data length of 1000 data points, even for non-biased data, is generally not enough to produce an accurate estimate of the true mean. Hence asking MSER-5 to find a truncation point for the ARMA(5,5) data sets with 1000 data points is an extremely difficult if not 'unfair' request, but an interesting exercise to see how well the method copes.

< Figure 5 about here>

Averaged replications or single run

MSER-5 removed a greater percentage of bias from data produced by averaging over 5 replications, than for data from a single run (see figure 6). In 76% of valid runs using the averaged data, the MSER-5 method removed at least 95% of the bias from the data, and in 88% of valid runs it removed at least 90% of bias. This is in contrast

with 52% and 67% respectively when single run data were used. A chi-squared test of these 2 distributions gave a p-value of 0.000 indicating a significant difference at the 95% level. This is logical, as averaging over replications reduces variation in the data producing a clearer difference between any initial bias and the steady-state data.

< Figure 6 about here>

Bias shape

The more sharply the initial bias declines, the more likely MSER-5 is to underestimate the warm-up period and to remove increasingly less bias. Removal of mean-shift bias was very successfully achieved, with over 98% of cases having over 99% of the bias removed. This was as expected due to the non-declining nature of this bias. When bias declines sharply, a large amount of the bias is at a relatively low severity and effectively 'hidden' by the variation in the data. Thus, as shown in figure 7, the amount of bias that is removed reduces in direct proportion with the sharpness of decline of the bias function; in descending order: mean-shift, linear, oscillating linear, quadratic, oscillating quadratic, exponential, oscillating exponential. Again, a chi-squared test of these 7 distributions of percentage bias removed produced a p-value of 0.000 indicating a significant difference at the 95% level.

< Figure 7 about here>

Direction of bias

The orientation of the bias (positive or negative) does not significantly affect the ability of MSER-5 to remove bias. This is not unexpected due to the nature of the MSER-5 heuristic.

Bias severity

As the severity increases, MSER-5 removes an increasingly higher percentage of the bias (see figure 8). A chi-squared test of these 3 distributions of percentage bias removed gave a p-value of 0.000 indicating a significant difference at the 95% level.

This is not surprising, since we would expect that it is easier to detect the bias when it is more severe.

< Figure 8 about here>

Bias length

The longer bias was removed slightly more efficiently by MSER-5 than the shorter bias. The shorter bias had a higher percentage of overestimations but this was partly due to overestimations in the longer bias being more likely to fall into the 2nd half of the data and therefore be categorised as *Lsol* rejections. In order to take this artefact of the method in to account figure 9 shows the percentage bias results for the data with a bias length of 10% versus data with 40% bias, including the percentage of cases that were rejections.

< Figure 9 about here>

6.5 Truncation estimates (*Lsol*) that fall into the second half of the data.

It has been suggested that the MSER method can be sensitive to outliers in the steady-state data (Sandikci and Sabuncuoglu 2006). We too have observed this phenomenon, but mainly where these 'outliers' occur just after the true truncation point. However, this is partially alleviated by using averaged replication data rather than single runs and by the fact that MSER-5 batches the data into batches of 5 data points hence further smoothing the data. We have also observed that it can struggle to function properly when faced with highly auto-correlated data. This issue is not isolated to just the MSER-5 method and can be partially alleviated by providing the method with more data. We have also observed 'end-point' *Lsol* values (see section 5).

In order to avoid an 'end-point' *Lsol* value, possible 'outlier' effects and the effect of high auto-correlation, MSER-5 rejects an *Lsol* value if it falls in the second half of the data (i.e. *Lsol* > n/2). This occurred in only 7.4% of the total 201600 runs and over 88% of these were from the highly auto-correlated ARMA(5,5) data sets. Figure 10

shows the distribution of *Lsol* rejections over the test data sets and illustrates the large contribution from ARMA(5,5) and to a lesser extent AR(1). It is not surprising that the two highest auto-correlated sets should cause this phenomenon.

There were also higher numbers of rejections from the data sets with L = 400 than L = 100 as would be expected, since if *Lsol* is an overestimate it is more likely to fall into the second half of the data. Using averaged data rather than single run data slightly increases the probability of getting an 'end point' *Lsol* value but also increases the probability of procuring a more accurate estimate of L.

< Figure 10 about here>

It was hypothesised that giving more data to the MSER-5 method in an iterative fashion would eventually produce a valid *Lsol* value where previously the *Lsol* value had been rejected. To check this each data set with a rejected *Lsol* value was given more data (successively 100 more data points) and the method re-run. It was found that for all of the data sets a valid *Lsol* value (i.e. $Lsol \leq n/2$) was returned by the time the data length was doubled (i.e. n = 2000). See figure 11 for examples of the relatively small amounts of extra data needed to achieve this in the ARMA(5,5) data. The ARMA(5,5) data required the most extra data, whereas for the other data types a further 100 or 200 data points would suffice. In the case of 'end point' rejections it was found that often just adding one more batch of data would suffice.

< Figure 11 about here>

6.6 Testing MSER-5 with data that has no initial bias.

The bias length parameter was set at 0 effectively producing data that had no initial bias. The only parameters now valid were the auto-correlation, error type and single run or averaged data type. These parameters were varied over the same levels as before.

If there is no initial bias in a data set it is hoped that a 'good' warm-up method would return a truncation point of zero or, failing that, truncate a very small amount of the data. It was found that MSER-5 returned a zero truncation value approximately 71% of the time (see table 10). The *Lsol* values of greater than 50 data points were mainly due to the highly auto-correlated AR(1) and ARMA(5,5) data sets.

Only 135 (5.6%) of the total 2400 *Lsol* values produced were rejected because they fell into the second half of the data. Of these 126 (93%) were from the highly auto-correlated ARMA(5,5) data.

< Table 10 about here>

6.7 Testing MSER-5 with data sets that have 100% initial bias.

The bias length parameter was set at 100% producing data that had not yet reached steady-state. The other 6 parameters were varied over the same levels as before.

Ideally we would like to see a 100% rejection rate (i.e. *Lsol* > n/2) as none of these data sets had finished warming-up. However, the value of the *Lsol* returned was highly dependent on how severe the bias was, how sharply the bias declined and how highly auto-correlated the data were. For data with 100% mean shift bias it was impossible to tell that these data were biased and therefore MSER-5 returned mainly zero *Lsol* values. The following results do not therefore includes the mean-shift bias values. The total percentage of *Lsol* rejections was 61%.

Figure 12 illustrates that the less severe the bias the more likely MSER-5 was to return a valid *Lsol* value rather than rejecting it. It also shows a difference between using averaged and single run data. Using averaged data produced more *Lsol* rejections than single run data due to its reduced variation. Also, the more highly auto-correlated the data, the more likely MSER-5 was to return a valid *Lsol* value rather than to reject it. The only exception to this being the ARMA(5,5) data that due to its very high auto-correlation and too short a run length always produced a very high number of rejections. The shape of the bias had the most impact. The

oscillating exponential and exponential bias decline so rapidly that MSER-5 returned valid *Lsol* values nearly 65% and 90% of the time respectively (see figure 13).

< Figure 12 about here>
< Figure 13 about here>

7. CONCLUSION

All of the warm-up methods found have shortcomings and suffer from a lack of consistent, comparable testing across the literature. Key problems are overestimation and underestimation of the truncation point, relying on restrictive assumptions and requiring estimation of a large number of parameters. Requiring that a method be easily automatable is a further restriction. If you were to relax this requirement, methods that were rejected by us at the short-listing stage become viable. We rejected all the graphical methods due to the need for user intervention and judgement, but it is

possible that some of these methods could feasibly be adapted for automation. Welch's method for example would require a reliable method to ascertain when the data becomes smooth and flat. The authors have seen examples where using MSER-5 on data smoothed by Welch's method can reliably give estimated truncation points that could be seen by eye as likely points where the data flattens out.

MSER-5 is not model or data type specific and is therefore a very general method. It does not require estimation of any parameters and can function adequately without user intervention. It has been shown to perform robustly and effectively for the majority of data sets tested. It is quick to run and fairly simple to understand. It is therefore an ideal candidate for automation and inclusion into an automated analysis system.

The testing approach used in this paper was found to be robust, comprehensive and simple to carry out. It is the authors' intention that it could be usefully utilized by readers for testing any warm-up method.

The next stage of work is to create a heuristic framework around MSER-5 to facilitate its incorporation into an automated analyser for implementation into simulation software. This framework needs to include a 'failsafe' mechanism and an iterative procedure for procuring more data when required.

REFERENCES

Alexopoulos, C., and A. F. Seila. 1998. *Output data analysis, Handbook of simulation*, 225-272. New York: Wiley.

Banks, J., J. S. Carson, B. L. Nelson, and D. M. Nicol. 2001. *Discrete-event system simulation*. 4th ed. New Jersey: Prentice-Hall.

Bause, F., and M. Eickhoff. 2003. Truncation point estimation using multiple replications in parallel. In *Proceedings of the 2003 Winter Simulation Conference*, 414-421.

Beck, A. D. 2004. Consistency of warm up periods for a simulation model that is cyclic in nature. In *Proceedings of the Simulation Study Group, OR Society*, 105-108.

Box, G. E., G. M. Jenkins, and G. C. Reinsel. 1994. *Time series analysis:forecasting and control*, 3rd ed. New Jersey: Prentice-Hall.

Bratley, P., B. Fox, and L. Schrage. 1987. *A guide to simulation*, 2nd ed. New York: Springer-Verlag.

Cash, C. R., D. G. Dippold, J. M. Long, and W. P. Pollard. 1992. Evaluation of tests for initial-condition bias. In *Proceedings of the 1992 Winter Simulation Conference*, 577-585.

Conway, R. W. 1963. Some tactical problems in digital simulation. *Management Science* 10(1): 47-61.

Delaney, P.J. 1995. *Control of initialisation bias in queuing simulations using queuing approximations*. M.S. thesis, Department of Systems Engineering, University of Virginia.

Fishman, G. S. 1971. Estimating sample size in computing simulation experiments *Management Science* 18: 21-38.

Fishman, G. S. 1973. *Concepts and methods in discrete event digital simulation*. New York: Wiley.

Fishman, G. S. 2001. *Discrete-event simulation, modeling, programming, and analysis*. New York: Springer-Verlag.

Gafarian, A. V., C. J. Ancker Jnr, and T. Morisaku. 1978. Evaluation of commonly used rules for detecting 'steady-state' in computer simulation. *Naval Research Logistics Quarterly* 25: 511-529.

Gallagher, M. A., K. W. Bauer Jnr, and P. S. Maybeck. 1996. Initial data truncation for univariate output of discrete-event simulations using the Kalman Filter. *Management Science* 42(4): 559-575.

Glynn, P.W., and D. L. Iglehart. 1987. A New Initial Bias Deletion rule. In *Proceedings of the 1987 Winter Simulation Conference*, 318-319.

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

Gordon, G. 1969. *System simulation*. New Jersey: Prentice-Hall.

Jackway, P. T., and B. M deSilva. 1992. A methodology for initialisation bias reduction in computer simulation output. *Asia-Pacific Journal of Operational Research* 9: 87-100.

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

Kimbler, D. L., and B. D. Knight. 1987. A survey of current methods for the elimination of initialisation bias in digital simulation. *Annual Simulation Symposium* 20: 133-142.

Lada, E. K., and J. R. Wilson. 2006. A wavelet-based spectral procedure for steady-state simulation analysis *European Journal of Operational Research* 174: 1769-1801.

Lada, E. K., J. R. Wilson, and N. M. Steiger. 2003. A wavelet-based spectral method for steady-state simulation analysis. In *Proceedings of the 2003 Winter Simulation Conference*, 422-430.

Lada, E. K., J. R. Wilson, N. M. Steiger, and J. A. Joines. 2004. Performance evaluation of a wavelet-based spectral method for steady-state simulation analysis. In *Proceedings of the 2004 Winter Simulation Conference*, 694-702.

Law, A. M., and W. D. Kelton. 2000. *Simulation modelling and analysis. New York:* McGraw-Hill.

Law, A. M. 1983. Statistical analysis of simulation output data. *Operations Research* 31: 983-1029.

Lee, Y-H., and H-S. Oh. 1994. Detecting truncation point in steady-state simulation using chaos theory. In *Proceedings of the 1994 Winter Simulation Conference*, 353-360.

Lee, Y-H., K-H. Kyung, and C-S. Jung. 1997. On-line determination of steady-state in simulation outputs. *Computers industrial engineering* 33(3): 805-808.

Linton, J. R., and C. M. Harmonosky. 2002. A comparison of selective initialization bias elimination methods. In *Proceedings of the Winter Simulation Conference*, 1951-1957.

Ma, X., and A. K. Kochhar. 1993. A comparison study of two tests for detecting initialization bias in simulation output. *Simulation* 61(2): 94-101.

Mahajan, P. S., and R.G. Ingalls. 2004. Evaluation of methods used to detect warm-up period in steady-state simulation. In *Proceedings of the 2004 Winter Simulation Conference*, 663-671.

Nelson, B. L. 1992. *Statistical analysis of simulation results, Handbook of industrial engineering*. 2nd ed. New York: John Wiley.

Ockerman, D. H., and D. Goldsman. 1999. Student t-tests and compound tests to detect transients in simulated time series. *European Journal of Operational Research* 116: 681-691.

Pawlikowski, K. 1990. Steady-state simulation of queueing processes: A survey of problems and solutions. *Computing Surveys* 122(2): 123-170.

Robinson, S. 2004. *Simulation. The practice of model development and use.* England: John Wiley & Sons Ltd.

Robinson, S. 2005. A statistical process control approach to selecting a warm-up period for a discrete-event simulation. *European Journal of Operational Research* 176: 332-346.

Rossetti, M. D., and P. J. Delaney. 1995. Control of initialization bias in queueing simulations using queueing approximations. In *Proceedings of the 1995 Winter Simulation Conference*, 322-329.

Rossetti, M. D., Z. Li, and P. Qu. 2005. Exploring exponentially weighted moving average control charts to determine the warm-up period. In *Proceedings of the Winter Simulation Conference*, 771-780.

Roth, E., and N. Josephy. 1993. A relaxation time heuristic for exponential-Erlang queueing systems. *Computers & Operations research* 20(3): 293-301.

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.

Sandikci, B., and I. Sabuncuoglu. 2006. Analysis of the behaviour of the transient period in non-terminating simulations *European Journal of Operational Research* 173: 252-267.

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

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

Sheth-Voss, P. A., T. R. Willemain, and J. Haddock. 2005. Estimating the steady-state mean from short transient simulations. *European Journal of Operational Research* 162(2): 403-417.

Spratt, S. C. 1998. *An evaluation of contemporary heuristics for the startup problem*. M. S. thesis, Faculty of the School of Engineering and Applied Science, University of Virginia.

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

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

White Jnr, K. P., M. J. Cobb, and S. C. Spratt. 2000. A comparison of five steady-state truncation heuristics for simulation. In *Proceedings of the 2000 Winter Simulation Conference*, 755-760.

Wilson, J. R., and A. A. B. Pritsker. 1978a. A survey of research on the simulation startup problem. *Simulation* 31(2): 55-58.

Wilson, J. R., and A. A. B. Pritsker. 1978b. Evaluation of startup policies in simulation experiments. *Simulation* 31(3): 79-89.

Yucesan, E. 1993. Randomization tests for initialization bias in simulation output. *Naval Research Logistics* 40: 643-663.

Zobel, C. W., and K. P. White Jnr 1999. Determining a warm-up period for a telephone network routing simulation. In *Proceedings of the 1999 Winter Simulation Conference*, 662-665.

ACKNOWLEDGEMENTS

This work is part of the Automating Simulation Output Analysis (AutoSimOA) project (www.wbs.ac.uk/go/autosimoa) that is funded by the UK Engineering and Physical Sciences Research Council (EP/D033640/1). The work is being carried out in collaboration with SIMUL8 Corporation, who is also providing sponsorship for the project.

FIGURES

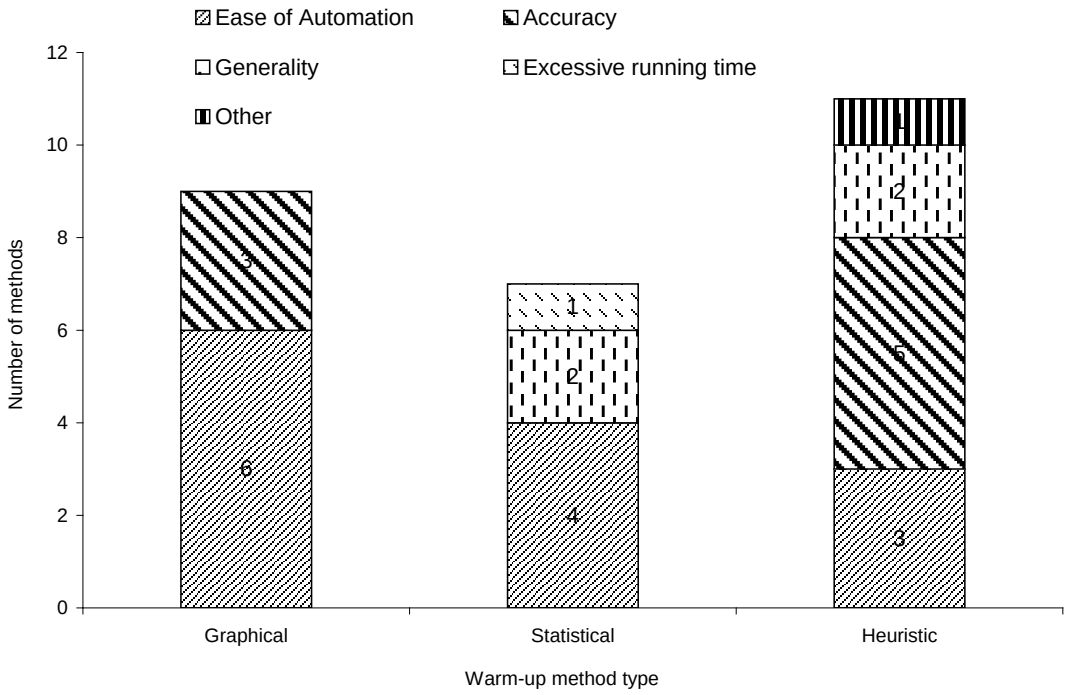


Figure 1: Summary of short listing results – reasons for rejection.

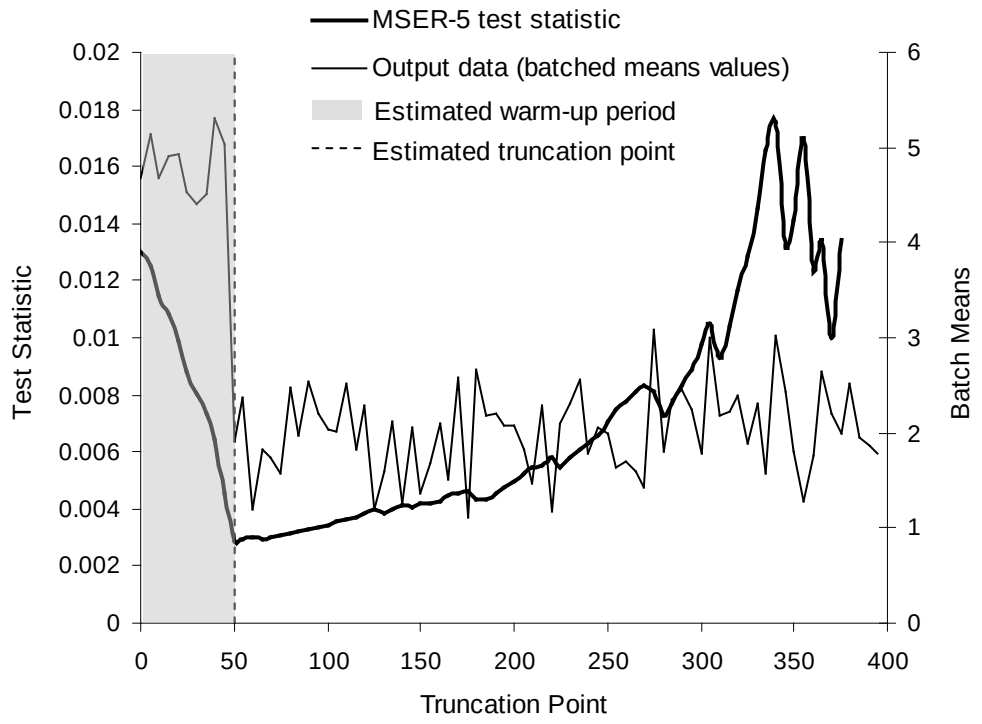


Figure 2: Example of the MSER-5 method at work.

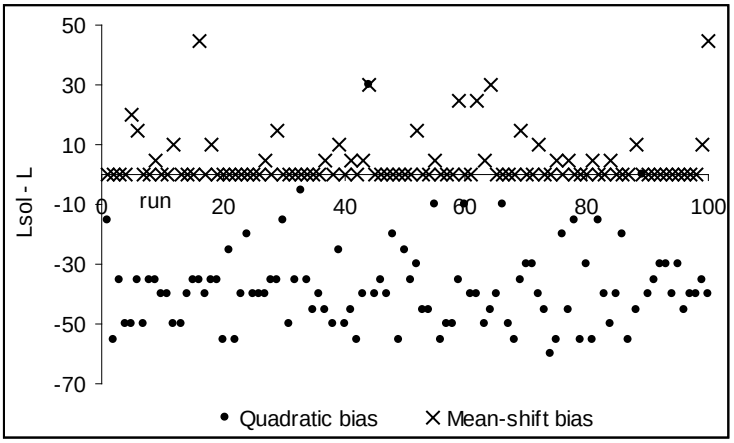


Figure 3: *Lsol* – L values for the positive quadratic and mean-shift bias functions used on single run data, with Normal(1,1) errors and MA(2) auto-correlation, a bias severity value of 2 and true L = 100.

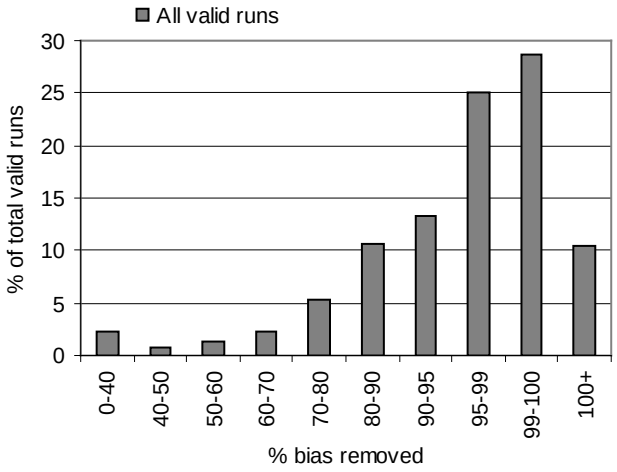


Figure 4: This graph shows the distribution of the percentage of bias removed from each data set where a valid *Lsol* value was returned. '100+' indicates that 100% of the bias was removed by truncation, but there was also over estimation of L so more data was removed than required.

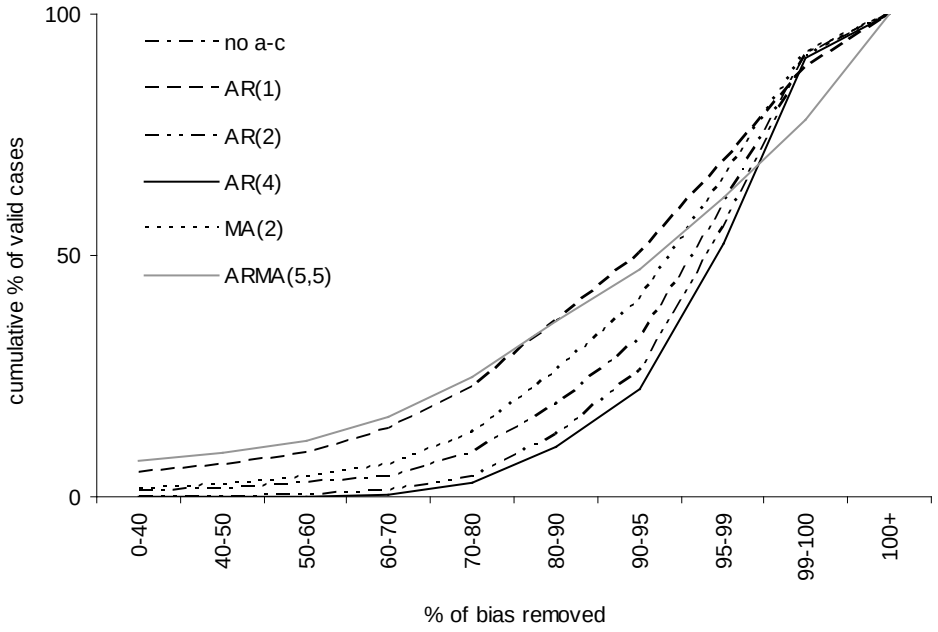


Figure 5: The cumulative percentage of bias removed by truncation, for each different autocorrelation type.

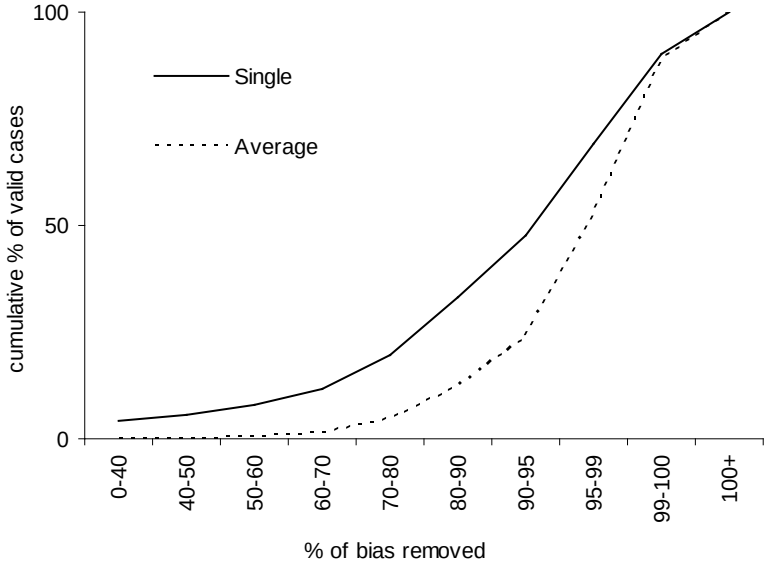


Figure 6: The cumulative percentage of bias removed by truncation, for each 'averaged' and 'single' data set where a valid *Lsol* value was returned.

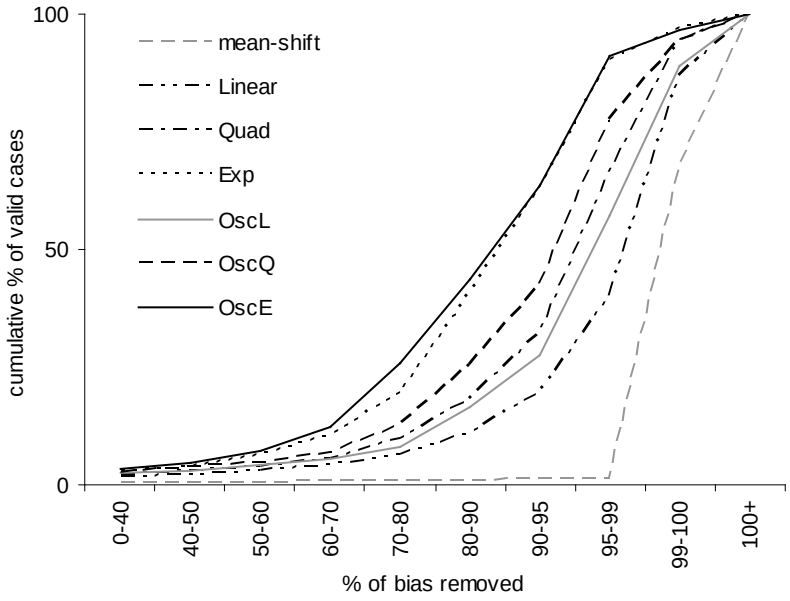


Figure 7: The cumulative percentage of bias removed by truncation, for each different bias shape.

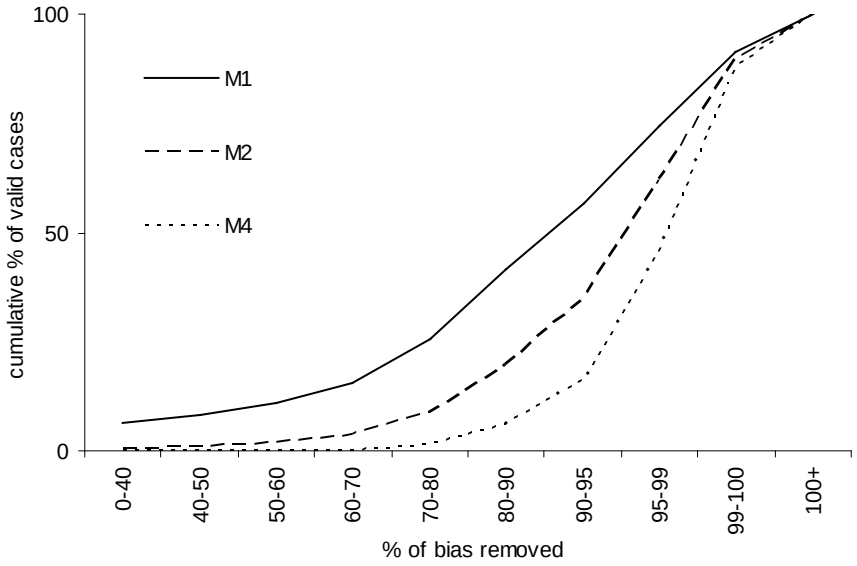


Figure 8: The cumulative percentage of bias removed by truncation, for each data set with varying severity of bias where a valid *Lsol* value was returned.

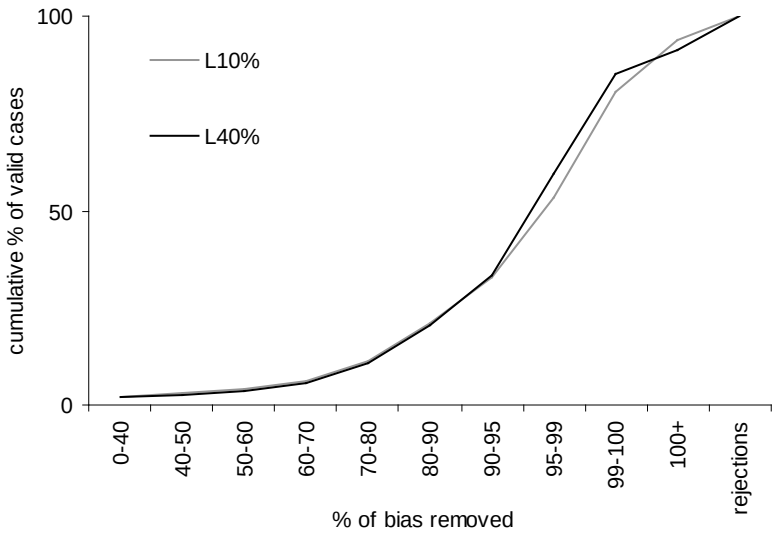


Figure 9: The cumulative percentage of bias removed by truncation, for each data set with 10% bias and 40% bias.

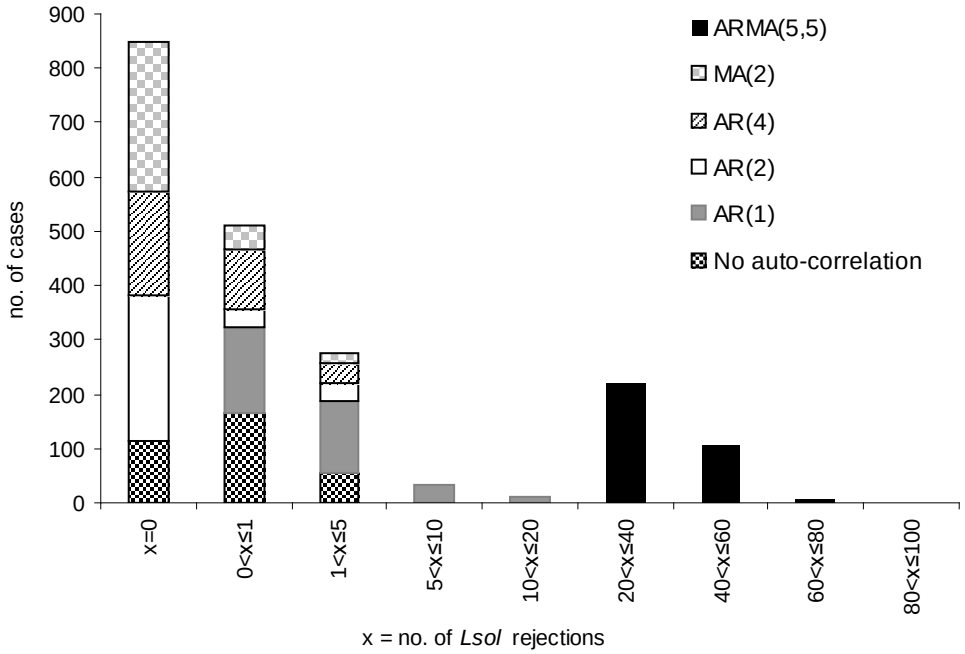


Figure 10: Distribution of *Lsol* rejections over the test data sets with respect to the different auto-correlation types.

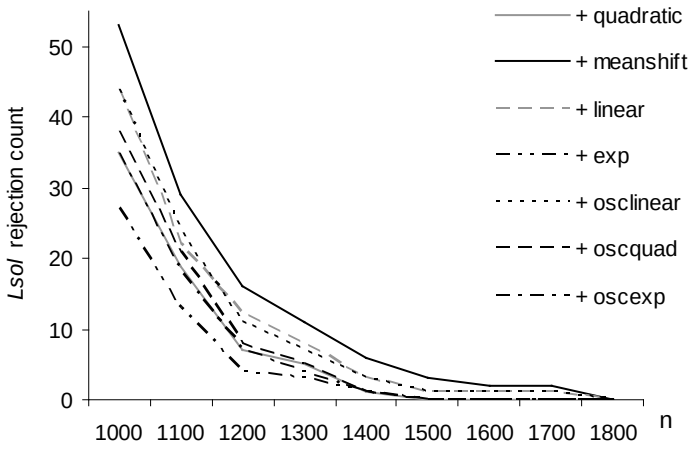


Figure 11: The number of *Lsol* rejections for differing amounts of data given to MSER-5, with respect to the differing bias shapes (using N(1,1) M2 L40 ARMA(5,5) positively biased averaged data).

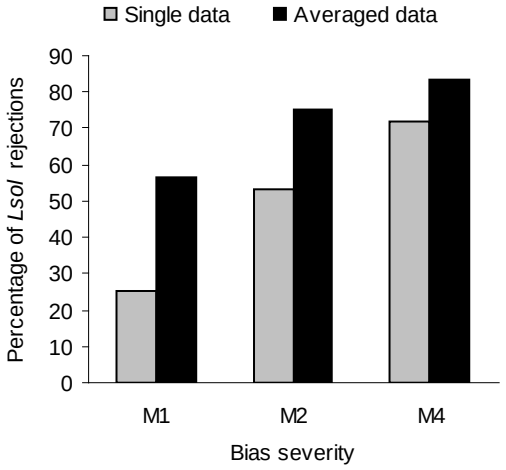


Figure 12: Percentage of *Lsol* rejections for data with 100% bias, divided into single run or averaged data and bias severity value (M) (*excludes mean-shift bias*).

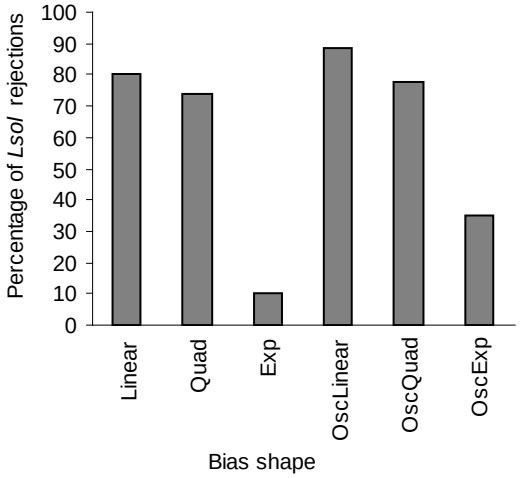


Figure 13: Percentage of *Lsol* rejections for data with 100% bias, described by their bias shape (*excludes mean-shift bias*).

TABLES

Table 1. Methods for determining the warm-up period.

Method Type	Method	References
Graphical	Simple Time Series Inspection	Gordon (1969)
	Ensemble (Batch) Average Plots	Banks et al. (2001)
	Cumulative-Mean Rule	Gordon (1969), Wilson and Pritsker (1978a), Gafarian et al. (1978), Nelson (1992), Roth and Josephy (1993), Roth (1994), Banks et al. (2001), Fishman (2001), Bause and Eickhoff (2003), Sandikci and Sabuncuoglu (2006)
	Deleting-The-Cumulative-Mean Rule	Roth and Josephy (1993), Roth (1994)
	CUSUM Plots	Nelson (1992)
	Welch's Method	Law (1983), Pawlikowski (1990), Alexopoulos and Seila (1998), Law and Kelton (2000), Banks et al. (2001), Linton and Harmonosky (2002), Bause and Eickhoff (2003), Mahajan and Ingalls (2004), Sandikci and Sabuncuoglu (2006)
	Variance Plots (or Gordon Rule)	Gordon (1969), Wilson and Pritsker (1978a), Gafarian et al. (1978), Pawlikowski (1990)
	Exponentially Weighted Moving Average Control Charts	Rossetti et al. (2005)
	Statistical Process Control Method (SPC)	Law and Kelton (2000), Mahajan and Ingalls (2004), Robinson (2005)
Heuristic	Ensemble (Batch) Average Plots with Schribner's Rule	Wilson and Pritsker (1978a), Wilson and Pritsker (1978b), Pawlikowski (1990)
	Conway Rule or Forward Data-Interval Rule	Conway (1963), Fishman (1973), Wilson and Pritsker (1978b), Gafarian et al. (1978), Wilson and Pritsker (1978a), Bratley et al. (1987), Pawlikowski (1990), Yucesan (1993), White (1997), Mahajan and Ingalls (2004)
	Modified Conway Rule or Backward Data-Interval Rule	Wilson and Pritsker (1978a), Gafarian et al. (1978), White (1997), Lee et al. (1997)
	Crossing-Of-The-Mean Rule	Wilson and Pritsker (1978a), Gafarian et al. (1978), Wilson and Pritsker (1978b), Pawlikowski (1990), White (1997), Lee et al. (1997), Mahajan and Ingalls (2004)
	Autocorrelation Estimator Rule	Fishman (1971), Wilson and Pritsker (1978a), Pawlikowski (1990)
	Marginal Confidence Rule or Marginal Standard Error Rules (MSER)	White (1997), White et al. (2000), Linton and Harmonosky (2002)
	Marginal Standard Error Rule m, (e.g. m=5, MSER-5)	White et al. (2000), Mahajan and Ingalls (2004), Sandikci and Sabuncuoglu (2006)
	Telephone Network Rule	Zobel and White (1999)
	Relaxation Heuristics	Kimbler and Knight (1987), Pawlikowski (1990), Roth and Josephy (1993), Roth (1994), Linton and Harmonosky (2002)
	Beck's Approach for Cyclic Output	Beck (2004)
	Tocher's Cycle Rule	Pawlikowski (1990)

	Kimbler's Double Exponential Smoothing Method	Kimbler and Knight (1987)
	Euclidean Distance (ED) Method	Lee et al. (1997)
	Neural Networks (NN) Method	Lee et al. (1997)
Statistical	Goodness-Of-Fit Test	Pawlikowski (1990)
	Algorithm for a Static Dataset (ASD)	Bause and Eickhoff (2003)
	Algorithm for a Dynamic Dataset (ADD)	Bause and Eickhoff (2003)
	Kelton and Law Regression Method	Kelton and Law (1983), Law (1983), Kimbler and Knight (1987), Pawlikowski (1990), Roth and Josephy (1993), Roth (1994), Gallagher et al. (1996), Law and Kelton (2000), Linton and Harmonosky (2002)
	Glynn & Iglehart Bias Deletion Rule	Glynn and Iglehart (1987)
	Wavelet-Based Spectral Method (WASSP)	Lada et al. (2003), Lada et al. (2004), Lada and Wilson (2006)
	Queueing Approximations Method (MSEASVT)	Rossetti and Delaney (1995)
	Chaos Theory Methods (methods M1 and M2)	Lee and Oh (1994)
	Kalman Filter Method	Gallagher et al. (1996), Law and Kelton (2000)
	Randomisation Tests for Initialisation Bias	Yucesan (1993), Mahajan and Ingalls (2004)
Initialisation bias tests	Schruben's Maximum Test (STS)	Schruben (1982), Law (1983), Schruben et al. (1983), Yucesan (1993), Ockerman and Goldsman (1999), Law and Kelton (2000)
	Schruben's Modified Test	Schruben (1982), Nelson (1992), Law (1983), White et al.(2000), Law and Kelton (2000)
	Optimal Test (Brownian bridge process)	Schruben et al. (1983), Kimbler and Knight (1987), Pawlikowski (1990), Ma and Kochhar (1993), Law and Kelton (2000)
	Rank Test	Vassilacopoulos (1989), Ma and Kochhar (1993), Law and Kelton (2000)
	Batch Means Based Tests – Max Test	Cash et al (1992), Lee and Oh (1994), Goldsman et al. (1994), Law and Kelton (2000), White et al. (2000)
	Batch Means Based Tests – Batch Means Test	Cash et al. (1992), Goldsman et al (1994), Ockerman and Goldsman (1999), White et al. (2000), Law and Kelton (2000)
	Batch Means Based Tests – Area Test	Cash et al. (1992), Goldsman et al (1994), Ockerman and Goldsman (1999), Law and Kelton (2000)
	Ockerman & Goldsman Students t-tests Method	Ockerman and Goldsman (1999)
	Ockerman & Goldsman (t-test) Compound Tests	Ockerman and Goldsman (1999)
Hybrid	Pawlikowski's Sequential Method	Pawlikowski (1990)
	Scale Invariant Truncation Point Method (SIT)	Jackway and deSilva (1992)

Table 2: six warm-up methods short-listed to be taken forward to further testing:

Statistical methods	Heuristics
Goodness-of-Fit (GoF) test (*Pawlikowski, 1990*)	MSER-5 (*White, 1997; White et al., 2000*)
Algorithm for a Static Data Set (ASD) (*Bause and Eickhoff, 2003*)	Kimbler's Double Exponential Smoothing (*Kimbler and Knight, 1987*)
Algorithm for a Dynamic Data Set (ADD) (*Bause and Eickhoff, 2003*)	Euclidean Distance Method (ED) (*Lee et al., 1997*)

Table 3: Results of preliminary testing for goodness-of-fit, Kimbler's double exponential smoothing, MSER-5 and Euclidean distance methods

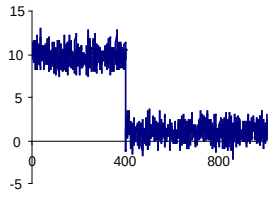
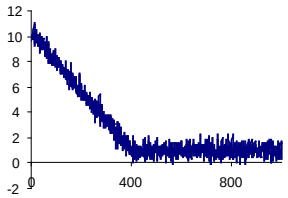
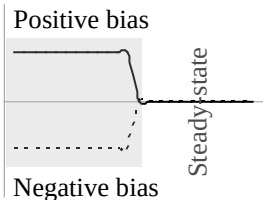
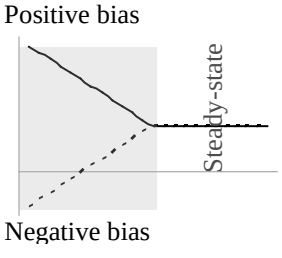
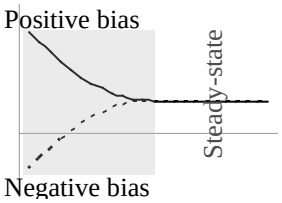
	Mean-shift bias, N(1,1) data.	Linear bias, N(1,1) data.
	 True L = 400	 True L = 400
Goodness-of-Fit	*n* = 1000	
Mean estimated L & (range):	1.5 (2)	1 (0)
Kimbler's D.Exp.S method	*n* = 1000	
Mean estimated L & (range):	17 (34)	33.7 (74)
MSER-5	*n* = 1000 to 1005	
Mean estimated L & (range):	407 (45)	381.5 (20)
ED	*n* = 1000 to 3300	
Mean estimated L & (range):	No results given	No results given
The data length, n, was initially set at 1000 data points and increased incrementally if a method was unable to supply an estimate of L for a specific data set with that number of data.		

Table 4: The 5 main bias shapes and their functions.

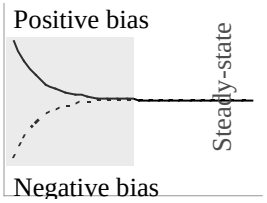
1.	Mean Shift:	$a(t) = \begin{cases} \pm QM, & t = 1,...,L \\ 0, & t = L+1,...,n \end{cases}$	
2.	Linear:	$a(t) = \begin{cases} \pm\left(\frac{QM}{1-L}\right)(t-L), & t = 1,...,L \\ 0, & t = L+1,...,n \end{cases}$	

3. Quadratic: $$a(t)=\begin{cases}\pm\dfrac{QM(L-t)^2}{(L-1)^2}, & t=1,...,L\\ 0, & t=L+1,...,n\end{cases}$$



4. Exponential:

$$a(t)=\begin{cases}\pm QM\left[\exp\left\{\mathrm{Ln}\left(\dfrac{QM}{k}\right)\Big/(L-1)\right\}\right]^{(1-t)}, & t=1,...,L\\ 0, & t=L+1,...,n\end{cases}$$



Where k = 0.005, is the value of $a(t)$ at $t = L$.

5. Oscillating (decreasing): $$a(t)=\begin{cases}\pm QM\Psi\,\mathrm{Sin}\left(\dfrac{\pi\,\mathrm{t}}{f}\right), & t=1,\ldots,L\\ 0, & t=L+1,\ldots,n\end{cases}$$

Where f is the frequency of oscillation for the Sin function. The number of cycles in the oscillating bias $a(t)$, t = 1,…,L, was set at 10, hence $f = L/10$. Ψ is either a linear, quadratic or exponentially decreasing function:

Linear decreasing function: $\Psi=\dfrac{t-L}{1-L}$,

Quadratic decreasing function: $\Psi=\dfrac{\left(L-\left(t-\left(\dfrac{f}{2}-1\right)\right)\right)^2}{(L-1)^2}$,

Exponentially decreasing function: $\Psi=\exp\left\{\left(t-\dfrac{f}{2}\right)\dfrac{\ln(k)}{L}\right\}$

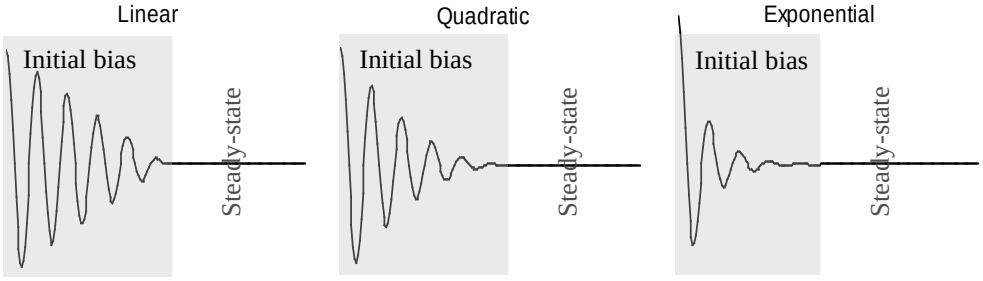


Table 5: Equations and parameter values of the steady-state functions.

Auto-correlation type	**Equation**	**Parameter values**
AR(1)	$X_t^{(1)} = \phi_1 X_{t-1}^{(1)} + \varepsilon_t$	$\phi_1 = 0.9$
AR(2)	$X_t^{(2)} = \phi_2 X_{t-1}^{(2)} + \phi_3 X_{t-2}^{(2)} + \varepsilon_t$	$\phi_2 = -0.25,\ \phi_3 = 0.5$
AR(4)	$X_t^{(3)} = \phi_4 X_{t-1}^{(3)} + \phi_5 X_{t-2}^{(3)} + \phi_6 X_{t-3}^{(3)} + \phi_7 X_{t-4}^{(3)} + \varepsilon_t$	$\phi_4 = -0.45,\ \phi_5 = 0.3,$ $\phi_6 = 0.2,\ \phi_7 = 0.1$
MA(2)	$X_t^{(4)} = \varepsilon_t + \phi_8 \varepsilon_{t-1} + \phi_9 \varepsilon_{t-2}$	$\phi_8 = 0.25,\ \phi_9 = 0.5$
ARMA(5,5)	$X_t^{(5)} = 1 + \varepsilon_t + \sum_{i=1}^{5} \left[\frac{1}{2^i} \left(X_{t-i}^{(5)} + \varepsilon_{t-i} \right) \right]$	
No auto-correlation	$X_t^{(6)} = \varepsilon_t$	
Where ε_t are random variates drawn from the Exp(1) or N(1,1) distributions.		

Table 6: True mean calculations for each steady-state model.

Auto-correlation type	Equation of the mean, μ
AR(q), where q = 1, 2, 4	$\mu = \frac{\mu_\varepsilon}{1 - \sum_{i=1}^{q} \phi_i}$
MA(q), where q = 2	$\mu = \mu_\varepsilon \left[1 + \sum_{i=1}^{q} \phi_i \right]$
ARMA(5,5)	$\mu = 32 + 63\mu_\varepsilon$
where μ_ε is the mean of the error function.	

Table 7: Factorial design of 7 factors with varying numbers of levels, with n = 1000

Error	**M**	**L**	**Auto-Correlation**	**Bias direction**	**Bias Shape**	**Data type**
N(1,1)	1	0 (0% of n)	None	+ (positive)	Mean-shift	Single run
Exp(1)	2	100 (10% of n)	AR(1)	- (negative)	Linear trend	
	4	400 (40% of n)	AR(2)		Quadratic trend	
		1000 (100% of n)	AR(4)		Exponential trend	Data averaged over 5 replications
			MA(2)		Oscillating (linearly decreasing)	
			ARMA(5,5)		Oscillating (quadratically decreasing)	
					Oscillating (exponentially decreasing)	

Does true mean fall within the 95% CI around the estimated mean?		Truncated data	
		No	Yes
Non-truncated data	No	19.8%	72.5%
	Yes	0%	7.7%
Total		19.8%	80.2%

Table 8: The percentage of cases that fall into the 4 possible combinations of coverage results.

Bias shape	mean-shift	quadratic
Mean *Lsol*	103.95 ± 1.75	62.55 ± 2.83
(Minimum, Maximum) *Lsol* value	(100, 145)	(40, 130)

Table 9: Mean *Lsol* values with 95% CIs for the data seen in figure 3.

		Normal errors	Exponential errors
Single run data	$Lsol = 0$	72.8%	75.2%
	$Lsol \leq 50$	96.1%	91.3%
Averaged data	$Lsol = 0$	66.7%	70.1%
	$Lsol \leq 50$	93.8%	91.9%

Table 10: Percentage of cases where MSER-5 returned *Lsol* values of zero and less than 50, for single run and averaged data with normal and exponential errors.

Figure Captions

Table Captions