

AUTOMATED ANALYSIS OF SIMULATION OUTPUT DATA

Stewart Robinson

Warwick Business School
University of Warwick
Coventry, CV4 7AL, U.K.

ABSTRACT

Appropriate analysis of simulation output is important to the success of a simulation study. Many users, however, do not have the skills required to perform such analyses. One way of overcoming this problem is to provide automated tools for analyzing simulation output. An Excel based automated “Analyser” is described that performs an analysis of a single scenario. The Analyser links to a commercial simulation software package, SIMUL8, and provides recommendations on warm-up, number of replications and run-length. Various standard procedures are used in the Analyser with some adaptations to make them suitable for automation. This research demonstrates the potential of the approach. A requirement for further development is more thorough testing of the analysis procedures, particularly for their robustness and generality in use. Further adaptation of the procedures for automation may also be required.

1 INTRODUCTION

The availability of commercial simulation software has placed simulation model development into the hands of non-experts by removing the need for a detailed knowledge of programming code. Today discrete-event simulation is in widespread use being applied in areas such as manufacturing design and control, service system management (e.g. call centres), business process design and management, and health applications. Organisations benefit from improved performance, cost reduction, reduced risk of investment and greater understanding of their operations. Further to this, the widespread use of accessible tools such as Excel to front and back end simulation models has placed simulation model use into the hands of the end user. As a result, these end users need not have specific simulation skills to use their models.

While a welcome development, the prevalence of simulation software and its adoption by non-experts has almost certainly lead to a significant problem with the use of the simulation models that are being developed. The

appropriate analysis of simulation output requires specific skills in statistics that many non-experts do not possess. Decisions need to be made about initial transient problems, the length of a simulation run, the number of independent replications that need to be performed and the selection of scenarios (Law and Kelton, 2000; Robinson, 2004). Appropriate methods also need to be adopted for reporting, comparing and ranking results. The majority of simulation packages only provide guidance over the selection of scenarios through simulation “optimisers” (Law and McComas, 2002). Other decisions are left to the user with little or no help from the software. As a result, it is likely that many simulation models are being used poorly. Indeed, Hollocks (2001) in a survey of simulation users provides evidence to support the view that simulations are not well used. The consequences are that incorrect conclusions might be drawn, at best causing organisations to forfeit the benefits that could be obtained and at worst leading to significant losses with decisions being made based on faulty information.

Alongside developments in simulation software and simulation practice, theoretical developments in the field of simulation output analysis have continued. Many of these developments are reported at the annual Winter Simulation Conference, which has a stream dedicated to the subject (e.g. Ingalls et al, 2004). The focus of the work reported, however, is largely on theoretical developments rather than practical application. For instance, a survey of research into the initial transient problem and methods for selecting a warm-up period found some 26 methods (Robinson, 2002). None of the methods, with the possible exception of time-series inspection (Robinson, 2004) and Welch's method (Welch, 1983), appear to be in common use.

Three problems seem to inhibit the use of output analysis methods:

- Most methods have been subject to only limited testing giving little certainty as to their generality and effectiveness

- Many of the methods require a detailed knowledge of statistics and so are difficult to use, especially for non-expert simulation users
- Simulation software do not generally provide implementations of the methods

AutoMod is one of the few packages that provides some experimental support through AutoStat (www.automod.com/products/autostat/autostat.asp).

One solution to these problems would be to implement an automated output analysis procedure in the simulation software. This would overcome the problem of the need for statistical skills. Such a procedure might involve full automation, giving the user the “answer”, or partial automation, providing guidance on interpretation to the user.

In this paper a prototype automated output analysis tool, which is based in Excel, is described. The tool links to the SIMUL8 software (www.simul8.com) and aims to give the user advice on warm-up and confidence interval generation through either multiple replications or batch means. An overview of the tool is given in the next section. This is followed by a more detailed description and illustration of the approach, focusing on the three main elements in the tool: warm-up determination, selection of the number of replications and run-length determination. The paper concludes with a discussion on what has been learnt from the development of the tool and further developments that are required.

2 OVERVIEW OF THE AUTOMATED OUTPUT ANALYSIS TOOL

The automated analysis tool (“Analyser”) analyses the output from a single scenario. It provides a recommended warm-up period and number of replications or run-length with the aim of obtaining a desired confidence interval width. There are no facilities for selecting and comparing multiple scenarios.

An overview of the procedure is shown in figure 1. Initial replications (as specified by the user) are performed with the simulation model. For the purposes of this work the simulation model is developed using SIMUL8, but the same procedure could be applied with any simulation tool. As long as the simulation software can be controlled from Excel VBA then the Analyser should be relatively easy to adapt to any simulation software.

The output data from the initial simulation replications are read into the Analyser which then provides recommendations concerning the warm-up period (section 4). Once the user has selected the desired warm-up period, he/she is asked whether the output data are terminating or non-terminating. For a terminating model the Analyser proceeds to determine the number of replications required to obtain a confidence interval of a specified width (section 5). For a non-terminating simulation, the user can select

multiple replications or the batch means method for constructing a confidence interval (section 6). In the latter case, the Analyser determines the batch size and run-length required.

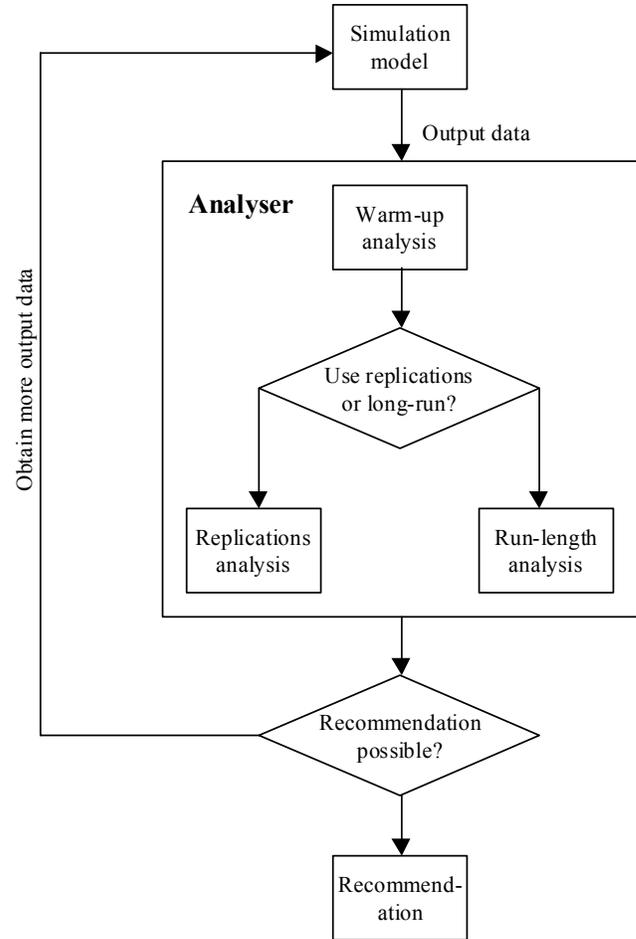


Figure 1: Overview of the Automated Analysis Procedure

If insufficient data are available to determine the warm-up, number of replications or run-length, the tool automatically asks SIMUL8 to perform more replications or a longer run. In either case the number of replications or the run-length is successively doubled until the required parameters can be determined, or the user interrupts the process.

3 ILLUSTRATIVE EXAMPLE

For the purposes of illustration, the Analyser is used with a simple M/M/1 queuing model. The key output data of interest is the number of customers in the system, which is set to a limit of 100. These data are collected in a time-series every time unit. The arrival rate (λ) is set at 1 with the service rate (μ) at 0.67, giving a traffic intensity (ρ) of

more than 1. This gives a steady-state number of customers in the system close to 100.

On loading and running the Analyser in Excel the user is prompted for the name of the SIMUL8 model that is to be analysed. The model is then loaded into SIMUL8 and the user prompted for the number of replications and run-length to use. This information is required to give an initial starting point for the analysis. In this case, 3 replications of 1,000 time units are requested. The Analyser then performs an analysis of the warm-up period required.

4 WARM-UP SELECTION

4.1 Choice of Warm-up Selection Procedures

Initial investigations were carried out to choose the warm-up selection procedures that would be appropriate for automation. The aim was to have 3 procedures so the user could select the warm-up period by comparing the results from the 3 methods. Criteria for choosing a procedure included both theoretical and practical considerations, including:

- *Accurate*: provides an accurate estimate of the length of the initial transient period.
- *Reliable*: consistently estimates the length of the initial transient.
- *General*: can be used across a range of output data types.
- *Easy to implement (in Excel)*: does not require sophisticated statistical procedures.
- *Requires minimum involvement from the user*: on the basis that he/she does not have the necessary expertise.
- *Varied*: for instance, it was not seen as desirable to rely on purely graphical procedures, but a range of methods should be used.

A review of warm-up selection methods reveals some 26 procedures (Robinson, 2002). It is also apparent that many of these procedures have not been fully tested, making it difficult to obtain objective data concerning the selection criteria listed above.

After some consideration, the following 3 procedures were adopted:

- *MSER-5* (White et al, 2000): empirical testing by White et al demonstrates the accuracy of this method. It does not require complex statistical procedures and no user interaction is required.
- *Batch Means Bias Detection* (Goldsman et al, 1994): although not the most accurate bias detection method, a key advantage is that it does not require an estimate of the variance of the output data.
- *Welch's Method* (Welch, 1983): this method seems to be the most popular warm-up selection procedure ap-

pearing in texts such as Law and Kelton (2000) and Robinson (2004). There are, however, some doubts about its accuracy and it may be conservative in its estimates due to its use of cumulative statistics (Gafarian et al, 1978; Wilson and Pritsker, 1978; Pawlikowski, 1990; Roth, 1994).

These three approaches represent a heuristic method, an initialisation bias test and a graphical method respectively. As such, the requirement for a range of methods is also met.

MSER-5 is not a sequential procedure, but makes a warm-up recommendation based on a fixed data set. The Analyser, however, adopts a sequential approach, asking for more data if they are required. In order to fit MSER-5 into this approach, the warm-up recommendation was rejected if the truncation point was at more than half the data available. In this case, more output data would be requested (section 2).

The batch means bias detection test does not specifically identify the warm-up period, but instead tests whether there is bias in the data for a proposed truncation point. The procedure was adapted for the Analyser by starting with a warm-up period of 0 and incrementing this value by 1 until no bias was detected.

4.2 Adaptation of Welch's Method for Automation

While MSER-5 and batch means detection are ripe for automation (i.e. the procedures require no user intervention), Welch's method is not. In this method the user must determine the window size required to give a "smooth" moving average line. The user must also determine the point at which the moving average line becomes smooth and flat ("convergence") to identify the warm-up period. Therefore, in order to automate Welch's method, smoothness and convergence criteria were generated.

Smoothness Criterion Suppose there is an output sequence $\{X_i\}_{1 \leq i \leq n}$ containing n observations. Define the i^{th} jump J_i , as

$$J_i = |X_{i+1} - X_i| \tag{1}$$

This is the absolute difference between the $(i + 1)^{\text{th}}$ and the i^{th} observations. Let the average jump \bar{J} be

$$\bar{J} = \frac{1}{n-1} \sum_{i=1}^{n-1} J_i \tag{2}$$

For a smooth data set we would expect the plot to progress steadily and there would be no sharp upward or downward jumps. For a rough data set we would expect a large number of sharp upward and downward jumps. Hence once a data set has been smoothed through the use of a moving

average we would expect the average jump to have reduced. This principle forms the basis of the smoothness criterion. \bar{J} is computed for the raw data and for the smoothed (moving average) data. The window size of the moving average is increased until the average jump has been reduced to 10% of the jump in the raw data.

Convergence Criterion (Average Difference Rule)

Suppose the moving average plot is deemed to have become smooth and flat at the j^{th} observation X_j . It is reasonable to assume that all following observations will be similar in value. The following statistic should, therefore, have a low value:

$$C_j = \frac{1}{n-1} \sum_{k=j+1}^n (X_k - X_j) \quad (3)$$

where n is the number of points in the moving average. This is known as the convergence criterion.

We need to determine what is a suitably low value for C_j . This is determined by obtaining a value of C_j such that $C_j/M < L$, where $0 < L < 1$. M is the difference between maximum and minimum value of X_i for $i \geq j$. The lower the value of L chosen, the stricter the convergence criterion. Testing showed that a value of $L=0.0025$ ensured that the average difference rule gave convergence points similar to those that would be chosen by visual inspection of the graph.

At this point the smoothness and convergence criteria have only been tested on a limited set of data. Their performance seems reasonable, although there was a tendency to slightly underestimate the length of the initial transient as compared to visual inspection of the moving average graph. Further testing and refinement is required for these criteria.

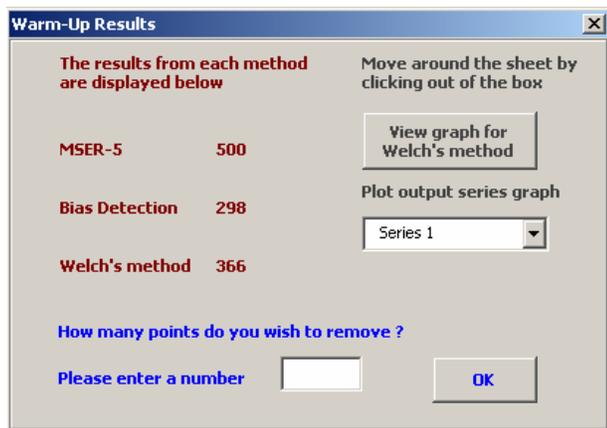


Figure 2: Warm-up Results for M/M/1 Model

4.3 Example of Warm-up Selection

Figure 2 shows the results obtained from the analysis of the output data from the M/M/1 SIMUL8 model. Figure 3 shows the moving average chart for Welch's method. There is some variation in the suggested warm-up period with, in this case, MSER-5 giving the most conservative estimate at 500 observations and the bias detection giving the smallest value at 298. The user is asked to enter the number of observations to delete. The choice depends on how conservative he/she wishes to be. This will depend on the context within which the simulation is being used, especially the desirability of accuracy over time to perform the experiments. Here a value of 300 is used.

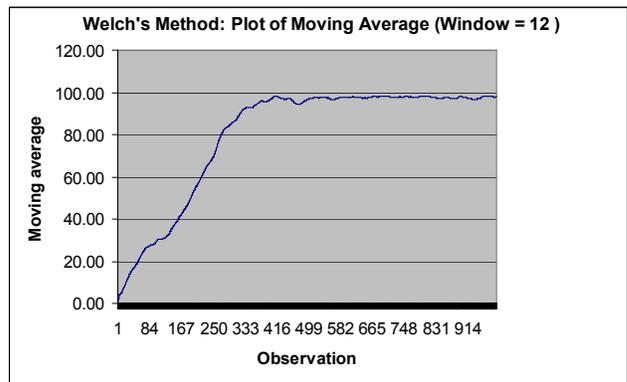


Figure 3: Moving Average Graph for Welch's Method

5 SELECTION OF NUMBER OF REPLICATIONS

Once the warm-up period has been selected the user is prompted for whether the simulation is terminating or non-terminating (figure 4). If the model is terminating, then the Analyser continues by selecting the number of replications. If the model is non-terminating, the user is given the option of using a single long run and the batch means method for confidence interval construction, or to use multiple replications (figure 5). In this section the selection of the number of replications is described. The determination of the batch size and the run-length for the batch means method is discussed in the next section.

At this point a warning will appear if the original run-length specified (section 3) is less than 10 times the warm-up period selected. This is based on Banks et al's (2001) recommendation. The user can choose to ignore this warning and continue with the run-length as specified.

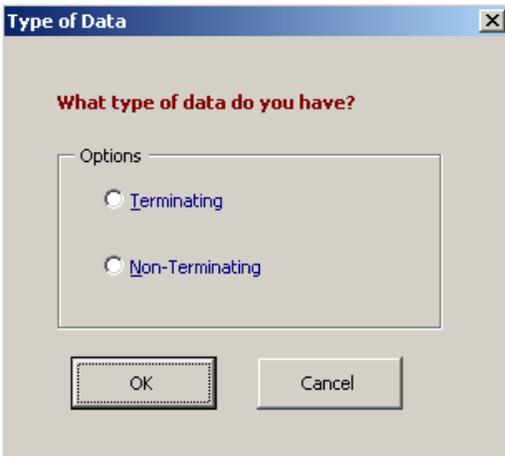


Figure 4: Prompt for Terminating or Non-Terminating Simulation

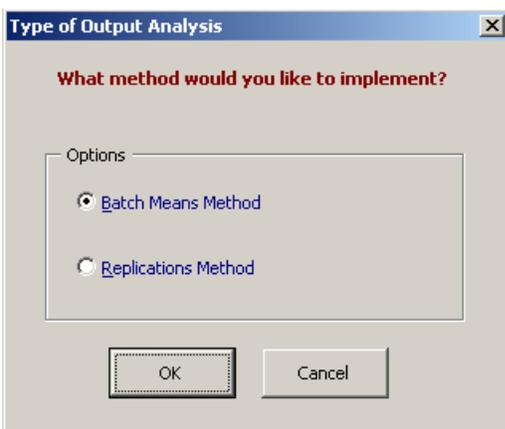


Figure 5: Prompt to Use Replications or Batch Means Method

5.1 Example of Selection of Number of Replications

The basic approach is to run sufficient replications to obtain a confidence interval of a specified precision. The user is given two options for determining the number of replications required (figure 6). The Analyser can perform the analysis on a set of normal runs, that is, runs using standard random number streams. Alternatively “antithetic runs” can be employed, in which replications with the normal random number streams are paired with replications with the antithetic values of those streams (Law and Kelton, 2000). This is one method of variance reduction which aims to reduce the total number of replications required; with varying degrees of success (Law and Kelton, 2000)! The “both” option (figure 6) will allow the user to compare the number of replications required when using just normal runs or paired normal and antithetic runs.

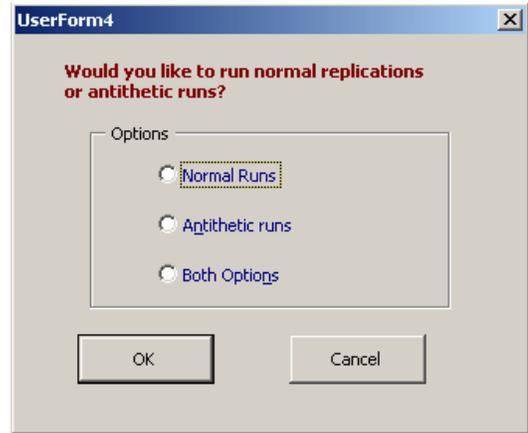


Figure 6: Prompt for use of Normal or Antithetic Runs

Note that because it is not possible to set SIMUL8 to run in antithetic mode remotely from Excel, the use of antithetic runs is not enabled. The Analyser does, however, have the capability to perform analyses with antithetic runs.

Following selection of normal or antithetic runs the Analyser asks for two parameters for performing a replications analysis (figure 7). First, the user needs to determine the significance level (α) to be used for the confidence interval calculation. Second, the precision required from the confidence interval needs to be specified. This is defined as the desired half width of the confidence interval, expressed as a percentage of the mean (“deviation”).

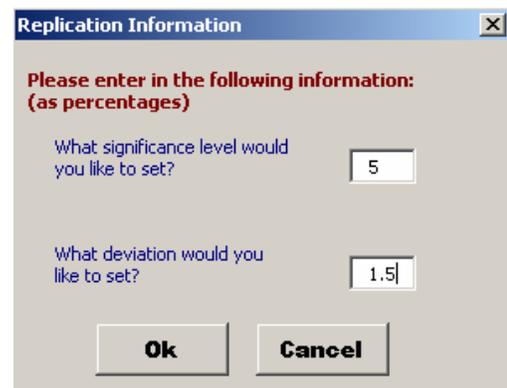


Figure 7: Prompt for Replications Parameters

The Analyser will continue to run replications until a confidence interval with the specified α and precision is obtained. It does this by successively doubling the number of replications performed from the base number (section 3).

Figure 8 shows the results from the replications analysis with the M/M/1 example model. In this case 5 replications are required to obtain a deviation of less than 1.5%. The replications graph (figure 9) shows how the confidence interval narrows as more replications are performed. In this case, the original 3 replications were insufficient to

obtain the desired precision. As a result the number of replications was doubled to 6 leading to the conclusion that 5 replications are required.

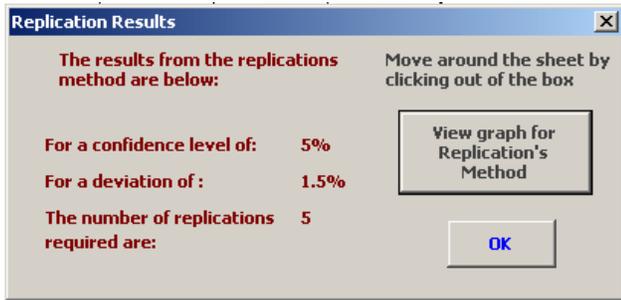


Figure 8: Results of Replications Analysis

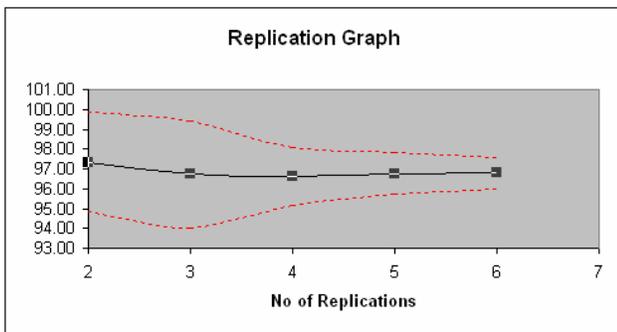


Figure 9: Graph for Replications Analysis

6 BATCH MEANS CONFIDENCE INTERVAL FOR SELECTING THE RUN-LENGTH

6.1 Choice of Batch Means Procedures

The key issue in the batch means method is the choice of the batch size to ensure independence in the data. Investigations into methods of batch size selection revealed 8 “non-overlapping batch means” approaches and some further methods based on overlapping batch means (Meketon and Schmeiser, 1984), spaced batch means (Fox et al, 1991) and weighted batch means (Bischak et al, 1993).

As for the selection of the warm-up period, the aim was to have 3 procedures in the Analyser. The user could then select the batch size and run-length from the results of these procedures. The criteria for selecting an algorithm for inclusion in the Analyser were the ease of understanding, the extent to which an algorithm had been tested, the robustness of the algorithm and the computational efficiency. Based on these criteria the following 3 algorithms were selected:

- Fishman’s algorithm (Fishman, 1978)
- Law and Carson’s algorithm (Law and Carson, 1979)
- ABATCH algorithm (Fishman and Yarberry, 1997).

Further to this, with each algorithm the ability to implement spaced batch means, as a method of reducing correlation between batches, was included.

Since neither Fishman’s nor the ABATCH algorithms are sequential procedures they were adapted to enable more output data to be collected when necessary. If the confidence intervals are too wide, the batch means are correlated or, in the case of Fishman’s algorithm, there are less than 10 batches, the run-length of the simulation will be increased. The process continues until the above conditions are met.

6.2 Example of Batch Means/Run-Length Selection

Figure 10 shows the form to enter the parameters for Fishman’s algorithm. There are similar input forms for Law and Carson’s and the ABATCH algorithms. Here the user is presented with a set of default values that he/she may adjust if desired. The parameters are the size of the confidence interval (1- α) to be constructed and the desired (half) width of the confidence interval. This is the precision required expressed as a percentage of the mean (as for the replications method, section 5.1).

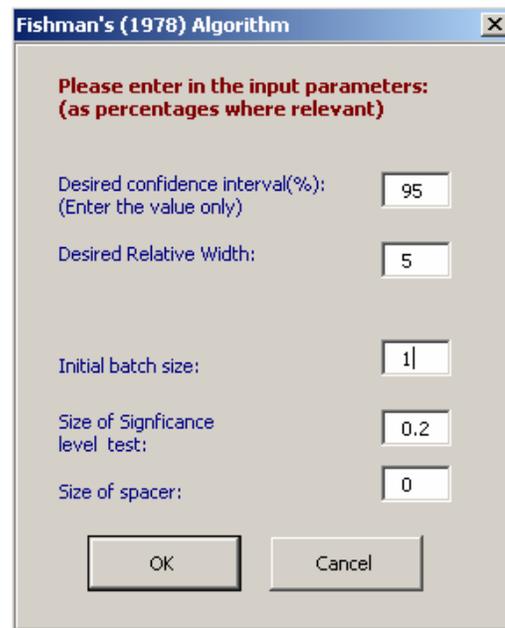


Figure 10: Prompt for Parameters for Fishman’s Algorithm

The initial batch size is set to 1 and is successively doubled until independence and the required confidence interval width are achieved. The size of the significance level test refers to the value used in the von Neumann test for independence that is part of Fishman’s algorithm. A value of 0.2 is used for the two sided test in line with Fishman’s recommendation.

Table 1: Results of Batch Means Analysis

Algorithm	Overall batch mean	Standard deviation	95% Confidence Interval				Batch size	Batches	Data points used
			Lower	Upper	Size of half width	Relative width			
Fishmans:	97.54	1.44	97.10	97.98	0.442	0.453%	16	43	688
Law and Carson:	97.48	1.21	96.91	98.05	0.566	0.581%	30	20	600
ABATCH:	97.51	0.88	97.06	97.96	0.450	0.462%	48	17	816

The final parameter allows the user to adopt a spaced batch means method. The default value of 0 implies the straightforward non-overlapping batch means approach.

Table 1 shows the results obtained from the M/M/1 example model. In all cases the relative width of the confidence interval is easily achieved. For the ABATCH algorithm additional output data (816 + 300 warm-up) were required beyond the 1,000 originally requested (section 3). The batch size required varies between 16 and 48 and the run-length between 600 and 816. Selection of which recommendation to use would depend on the user's preference and the context of the modelling work. There is, of course, a trade-off between a larger batch size, giving greater likelihood of independence, and the number of batches, giving greater precision in the confidence interval for a given run-length.

7 DISCUSSION

The discussion above demonstrates that it is possible to link an automated output analysis tool (in Excel) to a simulation model (in a commercial software package). The Analyser as described should be able to link to any SIMUL8 model and perform an analysis on a time-series of the output data. The analysis provides a recommendation on the warm-up period required and the number of replications/run-length needed to achieve a confidence interval of a specified precision. Indeed, the Analyser could be used with any simulation package as long as the software can be controlled from Excel and the output data can be read into Excel. As yet such linking has not yet been tried.

At present this work only aims to act as a proof of concept. There are many limitations with the Analyser and further work needs to be carried out. A key issue is the adoption and adaptation of the procedures used by the Analyser. The procedures currently used were chosen on the basis that they met certain criteria. Since many of the procedures have been subject to only limited testing, this selection was based largely on literature based judgments.

Rigorous testing, particularly for the generality and robustness of the procedures is required.

Some alterations have been made to the algorithms used in the Analyser. This has occurred for two reasons. First, some of the procedures as described in the literature work on a fixed data set and are not sequential procedures where the quantity of output data can be increased when required. In particular the MSER-5, Fishman's algorithm and the ABATCH algorithm were adapted to become sequential procedures. These adaptations need further testing.

The second reason for altering the algorithms is because some procedures require significant user intervention and so cannot be directly automated. Welch's method requires the user to inspect a graph for smoothness and flatness, while adjusting the window length of a moving average. The procedure was adapted to include criteria for smoothness and flatness (convergence). These criteria gave reasonable results, but require much more detailed testing.

The stopping procedure for the replications method also needs further investigation. The problem with stopping when a confidence interval of a specified precision is reached is that it assumes that the confidence interval narrows monotonically with successive replications. This is not the always the case. This could be addressed by performing a specified number of replications more than the Analyser recommends in order to check that the confidence interval precision remains within the bounds specified.

8 CONCLUSION

An automated analysis tool is described which provides a recommendation concerning the warm-up period and number of replications/run-length required for a simulation model. The key advantage of this approach is that it guides a non-expert simulation user in making these decisions with only limited training. An automated approach such as this aims to improve the use of simulation models at the experimentation stage. If developed further, this may help to address the need to ensure simulation models are used properly and appropriately.

ACKNOWLEDGMENT

I acknowledge the help of Masters students Kunal Govindia, Paras Shah and Nitin Sood in developing the automated output analysis tool.

REFERENCES

- Banks, J., J.S. Carson, B.L. Nelson and D.M. Nicol. 2001. *Discrete-Event System Simulation*, 3rd ed. Prentice Hall, Upper Saddle River, NJ.
- Bischak, D., W.D. Kelton and S. Pollock. 1993. Weighted Batch Means for Confidence Intervals in Steady-State Simulations. *Management Science*, 39, 1002-1019.
- Fishman, G. 1978. Grouping Observations in Digital Simulation. *Management Science*, 24, 510-521.
- Fishman, G. and L. Yarberr. 1997. An Implementation of the Batch Means Method. *INFORMS Journal of Computing*, 9, 296-310.
- Fox, F., D. Goldsman and J. Swain. 1991. Spaced Batch Means. *Operations Research Letters*, 10, 255-263.
- Gafarian, A.V., C.J. Ancker and T. Morisaku. 1978. Evaluation of Commonly Used Rules for Detecting "Steady State" in Computer Simulation. *Naval Research Logistics Quarterly*, 25, 511-529.
- Goldsman, D., L.W. Schruben and J.J. Swain. 1994. Tests for Transient Means in Simulated Time Series. *Naval Research Logistics*, 41, 171-187.
- Hollocks, B.W. 2001. Discrete-Event Simulation: An Inquiry into User Practice. *Simulation Practice and Theory*, 8, 451-471.
- Ingalls, R.G., M.D. Rossetti, J.S. Smith and B.A. Peters. 2004. *Proceeding of the 2004 Winter Simulation Conference*. IEEE, Piscataway, NJ.
- Law, A. and J. Carson. 1979. A Sequential Procedure for Determining the Length of a Steady-State Simulation. *Operations Research*, 27, 1011-1025.
- Law, A.M. and W.D. Kelton. 2000. *Simulation Modeling and Analysis*, 3rd ed. McGraw-Hill, New York.
- Law, A.M. and M.G. McComas. 2002. Simulation-Based Optimization. *Proceedings of the 2002 Winter Simulation Conference* (Yücesan, E., Chen, C-H., Snowden, S.L. and Charnes, J.M., eds.). IEEE, Piscataway, NJ, 41-44.
- Meketon, M. and B. Schmeiser. 1984. Overlapping Batch Means: Something for Nothing?. *Proceedings of the 16th Winter Simulation conference*. IEEE, Piscataway, NJ, 226-330.
- Pawlikowski, K. 1990. Steady-state Simulation of Queueing Processes: A Survey of Problems and Solutions. *Computing Surveys*, 22 (2), 123-170.
- Robinson, S. 2002. A Statistical Process Control Approach for Estimating the Warm-up Period. *Proceeding of the 2002 Winter Simulation Conference* (Yücesan, E., Chen, C-H., Snowden, S.L. and Charnes, J.M., eds.). IEEE, Piscataway, NJ, 439-446.
- 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.
- Welch, P.D. 1983. The Statistical Analysis of Simulation Results. *The Computer Performance Modeling Handbook* (Lavenberg, S., ed.). Academic Press, New York, 268-328.
- White, K.P., M.J. Cobb and S.C. Spratt. 2000. A Comparison of Five Steady-State Truncation Heuristics for Simulation. *Proceedings of the 2000 Winter Simulation Conference* (J.A. Joines, R.R. Barton, K. Kang and P.A. Fishwick, eds.). IEEE, Piscataway, NJ, 755-760.
- Wilson, J.R. and Pritsker, A.B. 1978. A Survey of Research on the Simulation Startup Problem. *Simulation*, 31, 55-59.

AUTHOR BIOGRAPHY

STEWART ROBINSON is Professor of Operational Research at Warwick Business School. He holds a BSc and PhD in Management Science from Lancaster University. Previously employed in simulation consultancy, he supported the use of simulation in companies throughout Europe and the rest of the world. He is author/co-author of three books on simulation. His research focuses on the practice of simulation model development and use. Key areas of interest are conceptual modelling, model validation and output analysis. Stewart is also managing a project to investigate the use of artificial intelligence for representing human decision-making in simulation models. His e-mail address is:
stewart.robinson@warwick.ac.uk
 and his Web address is:
www.btinternet.com/~stewart.robinson1/SR.HTM