Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

STATIONARITY TESTS AND MSER-5: EXPLORING THE INTUITION BEHIND MEAN-SQUARED-ERROR-REDUCTION IN DETECTING AND CORRECTING INITIALIZATION BIAS

William W. Franklin
K. Preston White, Jr.

Department of Systems and Information Engineering
PO Box 400747
University of Virginia
Charlottesville, VA 22904-4747, U.S.A.

ABSTRACT

We explore the reasoning behind MSER-5, an efficient and effective truncation heuristic for reducing initialization bias in steady-state simulation. We also compare MSER-5 with the KPSS stationarity test as one means of investigating the possibility that MSER's effectiveness is the result of its utility as a stationarity measure. Conversely, this comparison also lets us explore whether or not a stationarity test from the time-series literature can be used as an effective initialization bias-control heuristic. Finally, we investigate the use of an alternative form of MSER-5 that uses a variance estimator that adjusts for serial correlation.

1 INTRODUCTION

The problem of *initialization bias* has been well covered in the simulation literature. The most commonly employed means for reducing the bias introduced by unrepresentative initial observations in a simulation experiment is to simply delete some number of observations from the beginning of the trial. By "unrepresentative," we mean observations that have small stationary probabilities and values that are far from the steady-state mean. The long-run statistics for the experiment are then computed using only the remaining observations.

This technique is generally called *truncation*. The challenge for the experimenter who wishes to use this technique is to determine how many observations to delete. If the experimenter fails to delete enough points, the long-run statistics computed from the experiment will remain sufficiently biased to lead the experimenter to incorrect conclusions about the system being studied. If, conversely, the experimenter deletes too many observations, computational efficiency is sacrificed and the confidence intervals around computed statistics are unnecessarily wide. In an age of inexpensive computing, the risk of deleting too many points and sacrificing computational efficiency is often an acceptable cost. There still exist, however, simulations in which the rate of convergence to the steady state is sufficiently slow that computational efficiency remains a concern.

One method of selecting a truncation point that has been periodically demonstrated and reviewed is the MSER-5 heuristic (Spratt 1998), a modification of the MSER heuristic first proposed by White (1997). (See, also, White *et al.* 2000 for a more readily available discussion of MSER-5.) Though it has been shown to work well, requiring little computation bandwidth or experimenter intervention (Robinson 2005), MSER-5 has not been widely embraced. In part, this is because of the lack of sufficiently broad understanding of the intuition behind the truncation heuristic.

In this paper we explore two key notions about the intuition behind MSER-5. The first notion is that it optimizes on the objective function we most often care about in simulation studies, the confidence interval about the mean of a statistic. The second notion is that MSER-5 provides a reasonable method for determining when a level-stationary sequence of observations ceases to be stationary. A full exposition on the theory behind this second notion is beyond the scope of this paper. Instead, we look at the similarity in performance between MSER-5 and KPSS, a well-studied stationarity test first proposed by Kwiatkowski, *et al.* (1991).

2 OPTIMIZATION VS. DETECTION

Most of the truncation heuristics presented in literature approach the problem by attempting to detect the presence or magnitude of bias in a series of observations directly. MSER-5, by contrast, starts from the premise that observations near the end of a simulation run are most representative of the steady-state behavior of the system under study. The heuristic then works *backwards*

from the end of the run, gathering more observations and, thus, refining its estimate of the mean of the sequence of observations. As long as the observations, working backward in time, continue to be representative of the steady-state behavior of the system, the width of the estimated confidence interval around the estimate of the mean will continue to decrease. (The improvement of MSER-5 over MSER was simply to batch 5 observations together to help ensure the monotonic behavior of the decrease in confidence interval width over the range where steady-state behavior is in force. Computational savings also are achieved as a byproduct.) Once observations begin to be encountered that are not drawn from the system in its steady state, the departure from steady-state begins to exert an upward influence on the width of the confidence interval. Thus, MSER-5 selects that portion of the sequence of observations over which we are *most confident* in our estimate of the mean.

We should note, for clarity, that the standard normal estimate of the confidence interval, shown in equation (2), is known to be biased when applied to sequentially correlated observations because of bias in the estimate of variance. However, it is not the actual estimate of the confidence interval in which we are interested. Instead, we use the estimate of the confidence interval as a measure of the *similitude* in the truncated sequence. As (2) is a consistent estimator of the confidence interval, its use to pick out the subsequence mean about which we are *most* confident is appropriate. We address this issue later in this paper by introducing a variant of MSER that uses a variance estimator designed to account for the bias introduced by sequential correlation.

The relationship between MSER-5 and the confidence interval about the estimate of the mean can be shown mathematically. First, note that one way of expressing the MSER test statistic is simply

$$MSER(n,d) = \frac{S^2(n,d)}{n-d}, \qquad (1)$$

where S^2 is the large sample estimate of the variance for the sequence between observations d and n. The MSER truncation heuristic simply picks the value of d that minimizes (1). Note that this selection of d also minimizes the value of the confidence interval estimate, as shown in (2).

$$CI = \frac{z_{\alpha/2} S(n,d)}{\sqrt{n-d}} \qquad (2)$$

Thus, MSER optimizes our estimate of the mean, using the (2) as our objective function.

This relationship between MSER-5 (and confidence interval) and the observations and estimated mean can be vividly seen in the contrived example shown in Figure 1. The "source data" in this case is a simple ramp up to observations centered about 5. The mean is calculated by averaging the observations between the observation number and the end of the series. Likewise, the confidence interval estimate (2) is computed using the points from the observation number to the end. Note how, as we move from the end toward the beginning of the series, both the MSER value and the confidence interval decrease until we have included observations 5 through 20. Once we include observation 4, however, MSER and the confidence interval begin to increase again. Observation 4 is also the point at which "initialization bias" begins to appear in the form of a mean value that is clearly dropping from it's steady state value.

It should be noted that the truncation rules around MSER help to prevent truncation in the instance where no steady-state behavior is achieved. MSER also is effective, by its nature, in assessing cases in which no initialization bias exists, at least none to the point of adversely affecting our estimate of the mean.

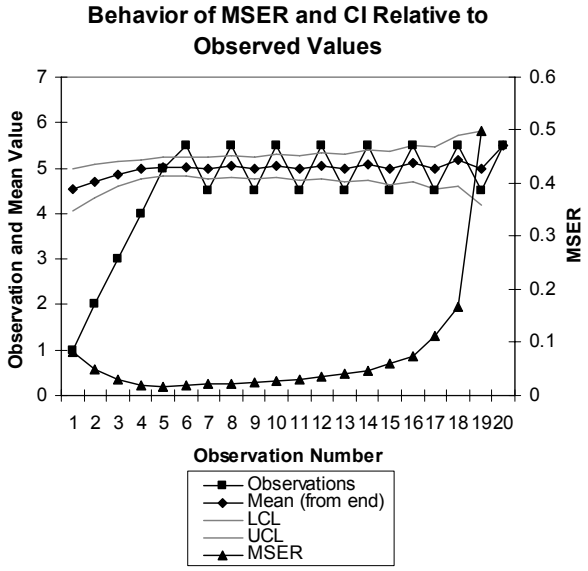

Figure 1: Illustration of MSER behavior

3 ASSESSING STATIONARITY

Experience and illustrations, such as those shown in Figure 1, led us to consider whether or not stationarity tests from the econometric and time-series literature would form the basis of an effective truncation heuristic or, reflexively, if MSER-5 was itself a reasonable level-stationarity assessment tool, especially in the special case where you want to know where a sequence *departs* from stationarity.

This notion of measuring stationarity is not new in simulation literature. Schruben (1982) and Schruben *et al.* (1983) used essentially this approach in creating tests for the presence of initialization bias. These techniques use hypothesis testing to decide if a sequence is stationary or not. Some use stationarity as the null hypothesis while others use stationarity as the alternate hypothesis. These

tests are "converted" into a truncation heuristic by testing different subsequences and determining the first subsequence that does not exhibit bias, one form of nonstationarity. We take the same approach in using the KPSS test as a truncation heuristic.

By contrast, when starting with a truncation heuristic such MSER-5 which does not take the form of a hypothesis test, we must translate the heuristic into a viable yes/no determination of whether or not the sequence is stationary. Fortunately MSER-5 determines its truncation point by locating the global minima of the MSER-5 heuristic. If the MSER-5 heuristic decreases monotonically as the truncation point is moved toward t=0, the initial observation, then the sequence is said to be stationary. This is because, for covariance-stationary sequences, the MSER-5 heuristic decreases as more observations are added. (A paper that presents more detail on the theoretical underpinnings of this behavior is currently in progress.) Furthermore, once a sequence departs from stationarity, the MSER-5 heuristic will begin to increase. This, in fact, is the mathematical behavior that makes MSER-5 a successful truncation heuristic.

Generally speaking, econometric and time-series literature equates nonstationarity with the presence of a "unit root," where root refers to the characteristic roots in an autoregressive process description of a time series (Hamilton 1994). The presence of a unit root effectively prevents the series from remaining in or returning to a consistent range of values. Of the available unit-root tests, we chose the KPSS test because of the weight of references to that test in recent literature on the topic. Variants and improvements have been described but KPSS appears to be a reasonable canonical reference point.

The mathematical format of the KPSS test (level stationarity form) is given by

$$\hat{\eta}_{\mu} = \frac{1}{n^2} \sum_{t=1}^{n} \frac{P_t^2}{s^2(l)} \tag{3}$$

where P_t is a partial sum process of the residuals from the mean and $s^2(l)$ is an estimate of the variance. More specifically, the KPSS test uses the Phillips-Perron estimator of variance, which was designed to adjust for effects of sequential correlation (Phillips and Perron 1988). The Phillips-Perron variance estimator is given by

$$s^2(l) = \frac{1}{n} \sum_{t=1}^{n} e_t^2 + \frac{2}{n} \sum_{s=1}^{l} w(s,l) \sum_{t=s+1}^{n} e_t e_{t-s} \tag{4}$$

where $w(s,l)$ is an optional weighting function that corresponds to the choice of a spectral window and e_t is the residual from the mean at observation t. The KPSS authors used the Bartlett window (5) as their weighting function and we follow their example, here. We also draw on the guidance of the KPSS authors in choosing lag, l, to be set equal to 8.

$$w(s,l) = 1 - \frac{s}{l+1} \tag{5}$$

Kwiatkowsky *et al.* (1991) demonstrate that the test statistic (4) asymptotically follows the Chi-squared distribution with associated critical values. The null hypothesis, in the case of the KPSS test, is that the series is stationary.

4 ADJUSTING MSER-5 WITH THE PHILLIPS-PERRON VARIANCE ESTIMATOR

With the Phillips-Perron correlation adjusted-variance estimator being computed as part of our investigation, we take advantage of the opportunity to create an "adjusted" form MSER-5. As noted earlier, the numerator of the MSER-5 (1) test statistic is simply the large sample variance estimator, which nominally assumes that samples are uncorrelated. Our adjusted version of MSER-5 replaces the large-sample variance estimator with the Phillips-Perron variance estimator. We then minimize this value to find the truncation point. We refer to this new test as the PPVR test, which stands for "Phillips-Perron Variance Reduction."

5 COMPARISON TESTS

5.1 Data Sets

For continuity of exposition, as well as to facilitate comparisons with previously studied heuristics, we use the same data sets that were used in White *et al.* (2000). For convenience, the details are repeated here.

The data sets are all based on the second-order autoregressive process

$$X_i = \Phi_1 X_{i-1} + \Phi_2 X_{i-2} + a_i \tag{6}$$

with initial conditions $X_1 = X_2 = 0$. a_i is standard normal noise. We use six different combinations of characteristic roots, as listed in Table 1.

Table 1: Equation (4) Parameters and Characteristic Roots

Model Number	Φ_1	Φ_2	Characteristic Roots
1	0.9	0.0	(0, 0.9)
2	0.9	0.0	(0, -0.9)
3	0.25	0.5	(-0.59307, 0.84307)
4	-0.25	0.5	(-0.84307, 0.59307)
5	0.75	-0.5	0.375±0.59948*i*
6	-0.75	-0.5	-0.375±0.59948*i*

The bias functions used were exponential bias (7), mean shift (8), and underdamped oscillation (9).

$$B_i = \begin{cases} Ce^{-0.005(i-1)} & \text{for } i = 1,...,1000 \\ 0 & \text{for } i > 1000 \end{cases} \quad (7)$$

$$B_i = \begin{cases} C/5 & \text{for } i = 1,...,1000 \\ 0 & \text{for } i > 1000 \end{cases} \quad (8)$$

$$B_i = \begin{cases} Ce^{-0.005(i-1)} \sin\left(\frac{\pi i}{200} + \frac{\pi}{2}\right) & \text{for } i = 1,...,1000 \\ 0 & \text{for } i > 1000 \end{cases} \quad (9)$$

These bias functions were selected because of their basis in Cash *et al.* (1992). Three different values of the bias coefficient, $C = 5,10,15$, were used for each of the three bias functions for a total of nine different bias functions.

We incorporated the bias terms in two different ways, addition and injection. The form of the addition of bias was simply

$$Y_i = X_i + B_i . \quad (10)$$

Injection was accomplished by forcing the bias directly into the difference equation as follows.

$$Y_i = \Phi_1 Y_{i-1} + \Phi_2 Y_{i-2} + a_i + B_i \quad (11)$$

With 6 different models, 3 different bias equations, 3 different bias coefficients, and 2 different ways of incorporating the bias, we have a total of 108 different experiments. Each experiment was run on 35 independently generated sequences, each with 10,000 observations.

5.2 Performance Measures

Table 2 is an example of the data that we captured for each of the 108 experiments.

Means:

- Mean – the average of the sample means after truncation has been applied
- |Bias| - absolute value of the difference between the "unbiased" mean of means and the mean of means computed after using each of the candidate heuristics
- Bias Reduction – the ratio of the reduced bias magnitude to the unreduced bias magnitude
- Median – median of the sample means
- Std Dev – standard deviation of the sample means
- 5^{th} Pctl – 5^{th} percentile of the sample means
- 95^{th} Pctl – 95^{th} percentile of the sample means
- *T*-test *p*-value – The *p*-value from a two-sample *t*-test comparing the unbiased sample means to those determined with each heuristic

Table 2: Example Experiment Summary Table

Summary Table for phi-1 = -0.75, phi-2 = -0.5, with Added Mean Bias (C = 10)				
	Unbiased	**MSER-5**	**PPVR**	**KPSS**
Means				
Mean	-0.00032	0.00034	0.00033	-0.00429
|Bias|		0.00066	0.00065	0.00397
Bias Reduction		99.7%	99.7%	98.0%
Median	-0.00018	0.00014	0.00009	-0.00415
Std Dev	0.00378	0.00446	0.00448	0.00602
5th Pctl	-0.00622	-0.00637	-0.00643	-0.01200
95th Pctl	0.00539	0.00732	0.00733	0.00405
T-test p-value		0.01951	0.02504	0.00007
Truncation Points				
Average		1009.74	1008.286	1145.63
Median		1006	1003	978
Std Dev		7.830	10.501	982.96
5th Pctl		1006	1002	964
95th Pctl		1025	1029	1025

Truncation Points:

- Average – average of the truncation point selections
- Median – median truncation point selection
- Std Dev – standard deviation of the truncation point selections
- 5^{th} Pctl – 5^{th} percentile of the truncation point selections
- 95^{th} Pctl – 95^{th} percentile of the truncation point selections

In addition to the above measures, we sought to assess the similarity of performance between KPSS and MSER-5. Of the various ways to accomplish this, the criterion we used as a first-order measure was if either the median or mean KPSS truncation point was within 10% of the MSER-5 truncation point. We assessed other characteris-

tics more qualitatively and describe those in the succeeding section.

5.3 Computation Environment

The bulk of the computation was performed using 'C' code on a cluster of Unix servers. Detailed comparison and summarization was performed using Excel and Excel VBA on a WindowsXP laptop computer.

5.4 Comparison Findings

No heuristic was consistently good in detecting bias in the underdamped oscillation cases. In a number of cases, acceptable bias reduction was achieved but we credit that to the fact that the mean of the oscillation bias is, by its nature, less than the mean introduced by a monotonic function of the same amplitude. For the remainder of this section, we refer only to the mean and exponential bias cases.

One of the key findings of the experiment runs was that, for mean and exponential bias cases, the MSER-5 and PPVR results were highly similar, usually choosing truncation points that were within 10 points of one another and, thus, yielded similar bias reduction results. The similarities were so compelling that we identified a new "winner" category of MSER/PPVR to distinguish those cases where the bias reduction or *t*-test results were identical or nearly so. This result leads us to assess that, for at least these sets of data, MSER-5 performs as well as a measure that explicitly seeks to adjust for sequentially correlated observations.

MSER-5 (and PPVR) resulted in better bias reduction and *t*-test performance in almost every case, including the underdamped oscillation cases, than did KPSS. Of the 108 experiments, KPSS yielded better bias reduction in only 4 cases and yielded better *t*-test *p*-values in only 7 cases. In the case of the *t*-test "wins" by KPSS, the p-values were only slightly above the .05 critical value that we selected as the minimum value we would accept in declaring a winner among the three heuristics.

KPSS chose an earlier median truncation point in most instances but also displayed a much greater degree of dispersion than either MSER-5 or the PPVR version of MSER, as measured by both the standard deviation and the range between the 5^{th} and 95^{th} percentiles.

Given our definition of "similar" (< 10% difference in truncation points), MSER-5 and KPSS were judged to be similar in roughly half of the cases, leading us to the conclusion that they do not behave in a similar fashion on this set of experiments.

6 CONCLUSIONS AND FOLLOW-ON SUGGESTIONS

Our goals in this paper were to:

- share with the WSC community the intuition behind mean-squared-error reduction as a bias elimination heuristic
- begin our exploration of MSER-5 as a stationarity measure by comparing its performance on canonical simulation literature sequences to a trusted stationarity measure
- explore the possibility that a stationarity measure from the econometric/time-series literature might be a good initialization bias reduction heuristic
- explore a modified version of MSER-5 that used a correlation-adjusted estimate of variance.

We found that using a correlation-adjusted variance estimate, in place of the large sample-variance estimate, does *not* yield an improvement in results. We also found that the KPSS stationarity measure does not yield an improvement in bias reduction over MSER-5. Instead, KPSS showed greater variability of results within each experiment. Because of KPSS' lack of performance on this data set, we were not able to make any statements about MSER-5's utility as a stationarity measure based on comparisons with KPSS.

Our continuing research includes:

- adapting MSER-5 for use as a stationarity measure and testing it on sequences from the time-series and econometric literature
- investigating MSER's behavior on highly correlated sequences to determine its limits of effectiveness, if such limits exist.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Stewart Robinson and Nitin Sood of the University of Warwick, U. K., for sharing their Excel VBA code for automating MSER-5.

REFERENCES

Cash, C. R., D. G. Dippold, J. M. Long, B. L. Nelson, and W. P. Pollard. 1992. Evaluation of tests for initial-condition bias. In *Proceedings of the 1992 Winter Simulation Conference*, ed. J. J. Swain, D. Goldsman, R. C. Crain, and J. R. Wilson, 577-585 Institute of Electrical and Electronics Engineers, Piscataway, NJ.

Hamilton, J. D. 1994. *Time Series Analysis*. Princeton, NJ: Princeton University Press

Kwiatkowski, D., P.C.B Phillips, P. Schmidt, and Y. Shin. 1991. Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root. *Journal of Econometrics*, 54:159-178.

Phillips, P. C. B., and P Perron. 1988. Testing for a unit root in time series regression. *Biometrika*, 75(2):335-346.

Robinson, S. 2005. Automated analysis of simulation output data. In Proceedings of the 2005 Winter Si-

mulation Conference, ed. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 763-770

Spratt, S. C. 1998. Heuristics for the startup problem. M.S. Thesis, Department of Systems Engineering, University of Virginia.

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

White, K. P, M. J. Cobb, and S. C. Spratt. 2000. A comparison of five steady-state truncation heuristics for simulation. In *Proceeding of the 2000 Winter Simulation Conference*, ed. J. A. Joines, R. R. Barton, K. Kang, and P. A. Fischwick, 755–760.

Schruben, L. W. 1982. Detecting initialization bias in simulation output. *Operations Research*, 30(3):151-153.

Schruben, L. W., H. Singh, and L. Tierney. 1983. Optimal tests for initialization bias in simulation output. *Operations Research*, 30(6):1167-1178.

AUTHOR BIOGRAPHIES

WILLIAM W. FRANKLIN is a graduate student in systems engineering at the University of Virginia. He is also a statistical analyst at a large financial services firm. Mr. Franklin received a B.S. and an M.S. in electrical engineering from the University of Texas in 1988 and 1989, respectively. Prior to becoming a statistician, he worked in the defense industry, specializing in signal processing, and spent 12 years as an IT professional. His email address is <wwf6r@viginia.edu>.

K. PRESTON WHITE, JR., is Professor of Systems Engineering at the University of Virginia and a past Board chairman of the Winter Simulation Conference. He received the B.S.E., M.S., and Ph.D. degrees from Duke University. He has held faculty appointments at Polytechnic University and Carnegie-Mellon University and served as Distinguished Visiting Professor at Newport News Shipbuilding and at SEMATECH. He is a member of INFORMS, SCS, and INCOSE and a senior member of IEEE and IIE. He sits on the Advisory Board of VMASC. His email address is <kpwhite@virginia.ed>.