*Proceedings of the 1993 Winter Simulation Conference*
*G.W. Evans, M. Mollaghasemi, E.C. Russell, W.E. Biles (eds.)*

# THE THRESHOLD BOOTSTRAP: A NEW APPROACH TO SIMULATION OUTPUT ANALYSIS

Y. B. Kim
Department of Mathematics
New Mexico Institute of Mining
and Technology
Socorro, NM 87801

T. R. Willemain
J. Haddock
G. C. Runger
Department of Decision Sciences
and Engineering Systems
Rensselaer Polytechnic Institute
Troy, NY 12180

## ABSTRACT

The threshold bootstrap (TB) is a promising new method of inference for a single autocorrelated data series, such as the output of a discrete event simulation. The method works by resampling runs of data created when the series crosses a threshold level, such as the series mean. We performed a Monte Carlo evaluation of the TB using three types of data: white noise, first-order autoregressive, and delays in an M/M/1 queue. The results show that the TB produces accurate and tight estimates of the standard deviation of the sample mean and valid confidence intervals.

## 1 INTRODUCTION

The inherent autocorrelation in simulation output creates a challenge for statistical inference. One method of handling this problem is independent replications. However, in some situations of interest it is not practical to generate multiple replications. For instance, a simulation of the entire Kawasaki steel processing plant in Japan takes about eight hours for a single replication on a SPARC workstation [Muro et al. 1991]. Furthermore, it is inefficient to use independent replications and discard the observations in the transient period for all replications.

The alternatives to independent replications are single replication methods. *Batch means* is the most common single replication method. However, its major drawback is that it requires substantial effort to determine the proper batch size [Sargent et al. 1992]. Another approach is ARIMA (time series) modeling of simulation output data. ARIMA modeling is not well suited for widespread use because ARIMA models are inherently complex and the parametric assumptions for ARIMA modeling may not be valid for a particular simulation. We propose a new nonparametric method for inference from a single simulation run.

## 2 BOOTSTRAP METHODS FOR TIME SERIES

Resampling techniques, such as the *jackknife* and *bootstrap methods* [Efron 1982, Léger et al. 1992], have achieved wide acceptance as nonparametric variance estimators. The conventional bootstrap analyzes pseudo-data constructed by resampling the original data. However, the conventional bootstrap assumes independence of the original data, which is not the case in simulation output.

A survey by Kim [1992] detailed three approaches to extending resampling techniques to autocorrelated data sets. In the first approach, an ARIMA model is fit to the data, then pseudo-series are created by resampling residuals and adding them to the fitted model [Thoms and Schucany 1990]. In the second approach, the *moving block bootstrap*, the data series is divided into adjacent blocks of fixed length, and pseudo-data are created by concatenating blocks chosen by resampling without replacement [Liu and Singh 1992, Künsch 1989]. A significant theoretical problem is that the data series are assumed to be *m-dependent*, i.e., only the last $m$ data values influence the current datum. This assumption is not true for heavily loaded queueing systems, which have slowly decaying autocorrelation functions. Furthermore, the choice of block size is not a simple matter. In the third approach, the *stationary bootstrap* [Politis and Romano 1992], the data series are resampled by concatenating blocks whose starting point is chosen at random and whose length is geometrically distributed with some chosen mean $p$. The choice of mean block length creates the same problem as the choice of fixed block length in the moving block bootstrap.

Kim, Haddock and Willemain [1993] described the *binary bootstrap* for binary data. The binary bootstrap resamples alternately from the runs of zeros and ones that comprise any binary series. This is a simpler approach to dividing the data into chunks for

resampling. Rather than taking fixed size blocks or geometrically distributed blocks, the data define runs that are used for resampling.

Empirical comparisons of the binary bootstrap with the batch means method were favorable for the binary bootstrap. For Bernoulli trials (no autocorrelation) and first-order Markov processes, the binary bootstrap produced excellent estimates of the standard deviation of the sample mean. For a heavily loaded M/M/1 queue, the binary bootstrap produced valid confidence intervals for the probability of long delay, at run lengths smaller than required by the batch means method. With longer runs from M/M/1 and D/M/10 queues, the confidence intervals of both methods were comparable in coverage, mean half-width and stability of half-width. One advantage of the binary bootstrap is that it does not require an extensive search for proper batch size.

Encouraged by the success of the binary bootstrap with binary time series, we investigated a generalization that does not require binary data or non-binary data clipped to binary form. We call the generalization the *threshold bootstrap* (TB).

## 3 THE THRESHOLD BOOTSTRAP

The threshold bootstrap works as follows:

Step 0: Obtain a time series with N values.

Step 1: Select a threshold value, such as the sample mean.

Step 2: Divide the series into runs that are either above or below the threshold (high and low runs).

Step 3: Create a bootstrap replication by concatenating runs chosen alternately from the populations of high and low runs. Resample the runs at random with replacement. Truncate the concatenated runs when their total length exceeds N.

Step 4: Compute the desired statistic, such as the sample mean.

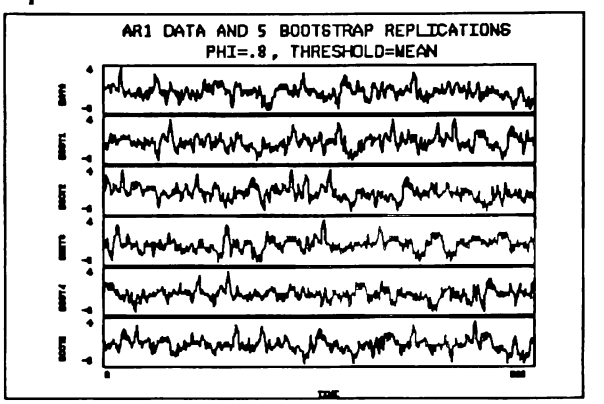
Step 5: Repeat Steps 3 and 4 a total of B times.

Step 6: Analyze the statistics from Steps 4 and 5 as if they were from independent replications.

## 4 PLAUSIBILITY STUDY

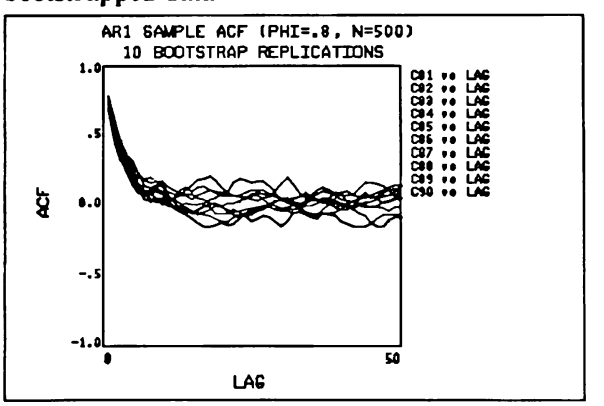
We began by conducting a plausibility study to see whether concatenation of blocks randomly defined by crossing a threshold preserved the autocorrelation structure of a data series. In this plausibility study, which used first-order autoregressive (AR1) data, we first compared the time plots of the original data series and several pseudo-series which were generated by the threshold bootstrap. Figure 1 shows that there were no gross differences between the original data series and the pseudo-series.

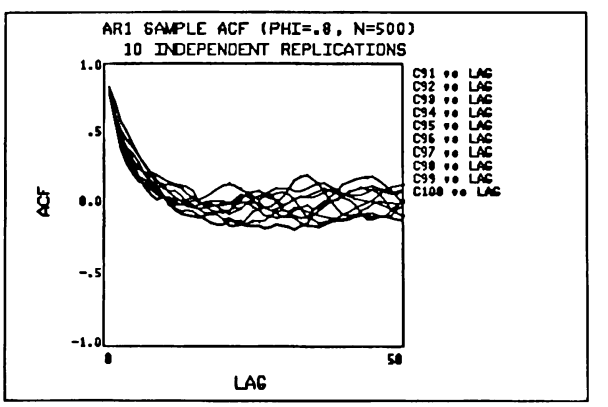
**Figure 1: Timeplots or AR1 series and bootstrap replications**




We next compared the autocorrelation function of original data series and pseudo-data series. Figure 2 shows the autocorrelation function of ten

**Figure 2: Autocorrelation functions of independent replications of AR1 data and bootstrapped data**



independent AR1 series and the bootstrapped replications of one of the independent series. Both sets of autocorrelation functions were similar in shape and variability.

## 5 MONTE CARLO EVALUATION

Encouraged by the plausibility study, we proceeded to Monte Carlo comparison of the TB with independent replication (IR). We focused on whether the TB accurately estimated the standard deviation of the sample mean and achieved nominal 90% coverage for confidence intervals for the sample mean. The study used three kinds of data: white noise, first-order autoregressive time series (i.e., $X_t = \phi X_{t-1} + a_t$) with parameter $\phi$=0.8, and delays before service in an M/M/1 queue with utilization $\rho$=0.8. We studied white noise to confirm that the TB results match conventional results in the case of no autocorrelation. The AR1 data resembled series encountered in finance and statistical process control. The M/M/1 data represented the type of data encountered in manufacturing and challenged the TB to work with very slowly decaying autocorrelations.

The study used series of various lengths. The white noise and AR1 data series had 20,000 points. The queueing data series had both 20,000 and 100,000 points; by the standards of single-series methods in discrete event simulation, even 100,000 points represent a modest series length. In general for single series methods of inference, the greater the degree of autocorrelation, the longer the run lengths needed to achieve stable estimates.

For each type of series, the Monte Carlo study used 100 independent replicates. The IMSL package generated the random deviates. We performed B=500 replications of the TB on each of the 100 independent replicates, thereby generating 100 estimates of the standard deviation of the sample mean and 100 confidence intervals. We compared the TB estimates against known theoretical values and against IR. Our experience is that the number of bootstrap replications B must be 500 or more; some trials with B=100 did not produce valid confidence intervals.

Table 1 summarizes the Monte Carlo results.

**Table 1: Summary of Simulation Results**

| | Type of Data | | | |
|---|---|---|---|---|
| **Statistic** | **White Noise** | **AR1 $\phi$ = .8** | **M/M/1 $\rho$ = .8** | **M/M/1 $\rho$ = .8** |
| # Data Points | 20,000 | 20,000 | 20,000 | 100,000 |
| **Average of Sample Mean** | | | | |
| Theory | .000 | .000 | 3.200 | 3.200 |
| IR ± 1.65 SE | -.000±.001 | -.003±.005 | 3.217±.041 | 3.187±.016 |
| TB ± 1.65 SE | -.000±.001 | -.001±.006 | 3.217±.039 | 3.189±.016 |
| **SD of Sample Mean** | | | | |
| Theory | .007 | .035 | .314* | .141* |
| IR | .007 | .031 | .248 | .099 |
| TB ± 1.65 SE | .007±.000+ | .035±.000+ | .248±.013 | .109±.002 |
| **Confidence Interval Coverage** | | | | |
| Nominal | .900 | .900 | .900 | .900 |
| TB ± 1.65 SE | .870±.055 | .890±.052 | .940±.039 | .900±.050 |

* Asymptotic approximation [Whitt 1989]

Note: Independent Replications (IR) results based on 100 independent replications. Threshold Bootstrap (TB) results based on B=500 bootstrap replications.

The section labeled "Average of Sample Mean" shows that the TB achieves unbiased estimates equivalent imprecision to those achieved by IR. The section labeled "SD of Sample Mean" shows that the TB produces estimates of uncertainty comparable in accuracy to IR. The section labeled "Confidence Interval Coverage" shows that TB confidence intervals are valid, even when using sample sizes that are small by conventional simulation standards (e.g., M/M/1 with N=20,000 points). Overall, the results in Table 1 document that the TB achieves with single samples the kind of results otherwise available with IR.

## 6 SUMMARY AND CONCLUSIONS

The threshold bootstrap is a promising new method of inference for single autocorrelated data series, such as arise in large discrete event simulations. The method works by resampling runs of data created when the series crosses a threshold level, such as the series mean.

Our plausibility study showed that TB pseudo-series have the same appearance and, more importantly, the same autocorrelation structure as the data from which they are generated. Our Monte Carlo evaluation showed that the TB can produce unbiased estimates of the series mean, accurate and precise estimates of the standard deviation of the mean, and valid confidence intervals for the mean.

The TB has some advantages over alternative single series methods of inference. First, it is conceptually simple. Second, it does not require the calculation of an appropriate batch size. Third, the TB algorithm is inherently suited to implementation on parallel computers, in which each processor can produce and analyze a pseudo-series.

Further research on the TB should focus on, first, development of a sound theoretical explanation for our empirical results; second, investigation of which types of time series and statistics are suitable for bootstrapping; third, implementation on parallel computers to bring closer the day of real-time simulation output analysis.

## REFERENCES

Efron, B. 1982. *The Jackknife, the Bootstrap and Other Resampling Plans*, CBMS:NSF, Philadelphia.

Kim, Y. B. 1992. Output Analysis of Single Replication Methods in Simulation Experiments, Ph.D. Dissertation, Department of Decision Sciences, Rensselaer Polytechnic Institute, Troy, NY.

Kim, Y. B., J. Haddock, and T. R. Willemain. 1993. "The Binary Bootstrap: Inference with Autocorrelated Binary Data", *Comm. Stat: Simulation and Computation*, 22, 205-216.

Künsch, H. R. 1989. "The Jackknife and the Bootstrap for General Stationary Observations", *Ann. Statist.*, 17, 1217-1241.

Liu, R., and K. Singh. 1992. Moving Blocks Jackknife and Bootstrap Capture Weak Dependence. In *Exploring the Limits of Bootstrap*. R. LePage and L. Billard (eds.). Wiley, New York.

Léger, C., D. N. Politis, and J. P. Romano. 1992. "Bootstrap Technology and Applications", *Technometrics*, 34, 378-398.

Muro, Z., N. Yanagihara, and H. Fujimoto. 1991. "Super Simulation Shell for Plants", In *Proceedings 1991 Winter Simulation Conference*. B.L. Nelson, W.D. Kelton and G.M. Clark (eds.). 289-293.

Politis, D. N. and J. P. Romano. 1992. The Stationary Bootstrap. In *Exploring the Limits of Bootstrap*. R. LePage and L. Billard (eds.). Wiley, New York.

Sargent, R.G., K. Kang, and D. Goldsman. 1992. "An Investigation of Finite-Sample Behavior of Confidence Interval Estimator", *Opns. Res.* 40, 898-913.

Thoms, L. A. and W. R. Schucany. 1990. "Bootstrap Prediction Intervals for Autoregression", *J. Amer. Statist. Assoc.*, 85, 486-492.

Whitt, W. 1989. "Planning Queueing Simulation", *Management Science*, 35, 1341-1366.

## AUTHOR BIOGRAPHIES

**YUN BAE KIM** is an Assistant Professor in the Department of Mathematics at New Mexico Institute of Mining and Technology. He received the B.S. degree in Industrial Engineering at Sung Kyun Kwan University, Seoul, Korea, the M.S. degree in Industrial and Systems Engineering from the University of Florida, and the Ph.D. in Engineering Science from Rensselaer Polytechnic Institute in 1992. His interests focus on simulation modeling and output analysis.

**THOMAS R. WILLEMAIN** is an Associate Professor in the Department of Decision Sciences and Engineering Systems at Rensselaer Polytechnic

Institute. He received the B.S.E degree in Electrical Engineering from Princeton University of 1969 and the S.M. and Ph.D. in Electrical Engineering from Massachusetts Institute of Technology in 1970 and 1972, respectively. His research interests include time series analysis, forecasting, and the process of model formulation.

**JORGE HADDOCK** is an Associate Professor of Industrial Engineering and Operations Research in the Department of Decision Sciences and Engineering Systems at Rensselaer Polytechnic Institute. He holds a BSCE from the University of Puerto Rico, a MSMgtE from Rensselaer, and a PhD in Industrial Engineering from Purdue University. Professor Haddock's primary teaching interests include Operations Research and Production Planning and Inventory Control courses at the undergraduate and graduate levels. His primary research interests involve modeling of manufacturing/production and inventory control systems, as well as the design and implementation of simulation modeling and analysis tools.

**GEORGE C. RUNGER** is an Assistant Professor in the Department of Decision Sciences and Engineering Systems at Rensselaer Polytechnic Institute. He received the B.S. in Industrial Engineering from Cornell University in 1974 and the Ph.D. in Statistics from the University of Minnesota in 1981. His research interests include the measurement and analysis of on-line manufacturing data (particularly for high-volume, sensor-based data acquisition systems), statistical quality control for autocorrelated data, adaptive real-time control, and novel experimental designs for small sample sizes using partial or incomplete prior knowledge.