D:\Data\WP\PAPERS\Bootva30.wpd

Printed: May 8, 2001 (5:23PM)

Written: May 4, 2001

VALIDATION OF TRACE-DRIVEN SIMULATION MODELS:

BOOTSTRAP TESTS

(Fourth version)

JACK P.C. KLEIJNEN[1], RUSSELL C.H. CHENG[2], and BERT BETTONVIL[3]

[1] Department of Information Systems (BIK)/Center for Economic Research (CentER), Tilburg University (KUB), 5000 LE Tilburg, Netherlands; e-mail: kleijnen@kub.nl; phone +3113-4662029; fax: +3113-4663377; http://center.kub.nl/staff/kleijnen

[2] Department of Mathematical Sciences, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom; e-mail: rchc@maths.soton.ac.uk; phone +441703-594550; fax: +441703-595147

[3] Department of Information Systems (BIK)/Center for Economic Research (CentER), Tilburg University (KUB), 5000 LE Tilburg, Netherlands; e-mail: B.W.M.Bettonvil@kub.nl; phone +3113-4662359; fax: +3113-4663377

VALIDATION OF TRACE-DRIVEN SIMULATION MODELS: BOOTSTRAP TESTS

JACK P.C. KLEIJNEN[1], RUSSELL C.H. CHENG[2], and BERT BETTONVIL[1]

[1] Department of Information Systems (BIK)/Center for Economic Research (CentER), Tilburg University (KUB), 5000 LE Tilburg, Netherlands
[2] Department of Mathematical Sciences, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom

Trace-driven (or correlated inspection) simulation means that the simulated and the real systems have some common inputs (say, historical arrival times) so the two systems' outputs are cross-correlated. To validate such a simulation, this paper focuses on the difference between the average simulated and real responses. To evaluate this validation statistic, the paper develops a novel bootstrap technique - based on replicated runs. This validation statistic and the bootstrap technique are evaluated in extensive Monte Carlo experiments with specific single-server queues. These experiments show acceptable type-I and type-II error probabilities.

(Keywords: TIME SERIES, DEPENDENCE, PAIRED OBSERVATIONS, ERROR RATES, POWER)

1. Introduction

We define *validation* as assessing whether a specific simulation model is an acceptable representation of the corresponding real system - given the goal of the simulation model (also see the classic textbook on simulation, Law and Kelton 2000). Many types of validation are used and proposed in practice and theory, but we focus on validation that uses *mathematical statistics*. After all, simulation means experimentation (albeit with a model instead of the real system), and experimentation calls for statistical analysis. Obviously, such an analysis is only part of the whole validation process (other parts are graphical summaries, the Schruben-Turing test on 'face validity', etc.; see Kleijnen 2000). However, when applying mathematical statistics, *correct* statistics should be used.

In this paper we discuss the type of statistical validation that compares data on the real and the simulated systems. Such a comparison makes more sense if both systems are observed under *similar scenarios*; for example, a busy day at the real supermarket should be compared with a busy day at the simulated store. Obviously, real data may pertain to input and output; for example, inputs are customers' arrival times and cashiers' service times at the supermarket, whereas outputs are customers' waiting times.

The most powerful statistical validation is possible if both input and output of the real system are measured. In so-called *trace driven* or *correlated inspection* simulation, analysts feed real input data into the simulation program, in historical order (also see Law and Kelton 2000). After running the simulation program, these analysts compare some summary statistic (namely, the average $\bar{X}$) for the time series of simulated output with the same statistic (namely, $\bar{Y}$) for the historical time series of real output.

In practice, these statistics may be seriously nonnormal. Therefore we use bootstrapping in this paper. Our *main conclusion* for a *specific* simulation that we use to illustrate our method, will be: If a trace-driven simulation model is replicated more than twice, then bootstrapping a simple statistic (namely, $T = \bar{X} - \bar{Y}$) gives statistically acceptable type-I error probability, given the prespecified nominal α value. The power is acceptable, provided the runlength is large enough.

The remainder of the paper is organized as follows. §2 introduces some notation. §3 recapitulates EFRON's bootstrapping of time series based on 'blocks', which we interpret as simulation runs. §4 derives a bootstrap procedure for trace-driven simulations. §5 illustrates the bootstrapped validation statistic: that section designs a Monte Carlo experiment with single-server queueing models that generate 'real' and simulated sojourn times. §6 interprets the results of this extensive Monte Carlo experiment. §7 presents conclusions and topics for future research. The appendix provides a theoretical underpinning: it proves that as the number of runs n tends to infinity, the estimated density function (EDF) of the bootstrap validation statistic T^* tends uniformly to the EDF of the original statistic T.

2. Notation and Bootstrapping in General

Consider the following realistic simulation problem. In trace-driven simulation, the simulated and the real systems have some common inputs $\boldsymbol{A}$; for example, the simulated and the real queueing systems use the same historical sequence of arrival times (we use capital letters for random variables, and bold letters for matrices including vectors). The real system is so complex that not all input variables are traced: the simulation model has at least one more input variable (e.g., service time) that is sampled using a pseudorandom number stream $\boldsymbol{R}$. There are s simulation runs that use the same trace and non-overlapping streams $R^{(r)}$, with integers r = 1, ..., s and $s \geq 1$.

The real system generates a *time series* of (autocorrelated) outputs $W_{i;\,t}$, whereas the simulated system generates outputs $V^{(r)}_{i;\,t}$ for run i = 1, ..., n and t = 1, 2, ..., k; for example, sojourn time of job t on day i. For simplicity we assume that k is a constant: $k_i = k$ (but this assumption does not affect the basic idea of our bootstrap method).

The real output time series is characterized through a single *performance measure* X_i; an example is the average sojourn time on day i. A crucial assumption is that these X_i are identically and independently distributed (IID). This assumption may hold if each run starts in the empty state. Obviously, we focus on *terminating* simulations.

To validate the simulation model statistically, we compare the real and the simulated performance measures X and (say) Y. One solution was given by Kleijnen, Bettonvil, and Van Groenendaal (1998), assuming that (X, Y) is bivariate *normal*. However, in case of short runs (for example, k = 10) the performance measures may be seriously nonnormal. This nonnormality is not well handled by conventional techniques, if n is small so that the central limit theorem does not apply. Therefore we use bootstrapping in this paper.

Bootstrapping enables estimating the distribution of *any* statistic, but different statistics have different sensitivities to scales, and so on; see the seminal book on bootstrapping (outside simulation) Efron and Tibshirani (1993, pp. 54-56, 162-177), next abbreviated to EFRON. For the validation of trace-driven simulations Kleijnen, Cheng, and Bettonvil (2000) bootstrap six statistics. We, however, focus on the simplest statistic that nevertheless gives good results, namely the *average deviation* between the real and the simulated performance measures, $T = \sum_{i=1}^{n} (X_i - Y_i)/n = \bar{X} - \bar{Y}$.

In general, *bootstrapping* takes a random sample of size n - with replacement - from the original n IID observation. Bootstrapping has not yet been applied frequently in simulation. Simulation yields (autocorrelated) time series, and EFRON (p. 396) warns: '... problems of dependence do not appear to be well understood and are an important area for further research'.

Further, EFRON (pp. 115, 383) cautions: 'bootstrapping is not a uniquely defined concept [...] alternative bootstrap methods may coexist'. Indeed, we shall show that bootstrap methods require the art of modeling; computer power does not suffice.

Moreover, we wish to test the *hypothesis* that the simulation model is 'valid', that is, the null-hypothesis is that the real and the simulated systems have equal means. EFRON (pp. 220-236) does discuss hypothesis testing, but that discussion does not apply here since trace-driven simulation does not give independent X and Y.

Our *main discovery* is: One simulation replicate is certainly a valid model for another simulation replicate, so the hypothesis of a valid simulation model is guaranteed to hold! So if we have more than a single simulation replication ($s \geq 2$), we can obtain the bootstrap distribution of any validation statistic under the null-hypothesis of a valid trace-driven simulation model. This idea can be generally applied to any trace-driven simulation.

3. EFRON's Bootstrapping of Time Series

EFRON (p. 91) assumes a sample of n IID observations $\boldsymbol{Z}_i$ with $i = 1, ..., n$. The sample data is summarized through a statistic $T = s(\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n)$. *Bootstrapping* means that the original observations $\boldsymbol{Z}_i$ are randomly resampled with replacement, n times. So, if the superscript * indicates bootstrapping, then the bootstrap observations are $\boldsymbol{Z}_i^*$.

This bootstrap sample gives one observation on the bootstrap statistic $T^* = s(\boldsymbol{Z}_1^*, ..., \boldsymbol{Z}_n^*)$. To estimate the distribution of this statistic, the whole procedure is repeated b times. Sorting these b observations on T^* gives the order statistics $T^*_{(1)}, ..., T^*_{(b)}$, and the estimated α quantile of its distribution, $T^*_{(\lfloor b\alpha \rfloor)}$. This gives a two-sided 1- α confidence interval for the original statistic T, ranging from the lower estimated $\alpha/2$ quantile to the upper 1 - $\alpha/2$ quantile.

For *time series* (which do not give IID sample observations), EFRON (pp. 99-102) presents 'moving blocks'. For simulation applications we interpret these blocks as *runs*. We assume that these runs give IID performance measures. In the example of §5, each run starts in the empty state and is of constant length k; we do not eliminate the transient phase.

4. Bootstrapping of Validation Tests in Trace-driven Simulation

We assume a 'reasonable' number of IID runs, namely $n = 10$. Further, we assume a sensible number of simulation replications, namely $s = 10$ (also see Schmeiser's (1982) rule of thumb for s).

In our bootstrap application we define $\boldsymbol{Z}_i = (Y_i^{(r)}, Y_i^{(r')})$ with $r, r' = 1, .., s$ and $r \neq r'$ ($Y_i^{(r)}$ denotes the performance measure calculated from run i with trace $\boldsymbol{A}_i$, in replicate r). So we compare the s simulation replicates per run i (with trace $\boldsymbol{A}_i$); that is, we *condition* or *block* on the trace. This results in the bootstrap validation statistic T^*. Repeating b times gives a two-sided (1 - α) confidence interval for T^*, under the *null-hypothesis* of a valid trace-driven simulation model.

We also have s observations on the original validation statistic that uses $(X_i, Y_i^{(r)})$, under the *alternative hypothesis*. We reject the simulation model if any of these s values (or equivalently, the maximum) falls outside the (1 - α/s) confidence interval: *Bonferroni*'s inequality.

In the appendix we prove that as n tends to infinity, the EDF of T^* tends uniformly to the EDF of T.

5. Example: Designing Monte Carlo Experiments with Queuing Simulations

We focus on an $\alpha = 0.10$ type-I error rate of the validation test. For the bootstrap sample size we take a classic value: b = 1,000; see EFRON (p. 275).

We investigate M/M/1 models that generate 'real' and simulated sojourn times $W_{i;\,t}$ and $V_{i;\,t}^{(r)}$, resulting in average sojourn times per run, X_i and Y_i. These models have real and simulated traffic loads (say) ρ and $\tilde{\rho}$. To study the *type I error* of the validation tests, we use simulated and real systems with equal traffic rates: $\tilde{\rho} = \rho$. However, the simulation model is imperfect: real and simulated service times use different pseudorandom numbers (arrival times are traced).

To study the *type II* error, we use unequal simulated and real rates: $\tilde{\rho} \neq \rho$ (for specific values see Table 1 in §6) Note that the traffic rates affect not only the means but also the variances of the real and simulated performance measures; bootstrapping takes care of any nuisance parameters.

We give results for k = 10 (short runs, so high nonnormality) versus k = 1,000. We set n = 10 (higher n would give better convergence of the bootstrap distribution), and $\rho = 1$ (real terminating system with very high traffic) and various $\tilde{\rho}$.

We use 1,000 macro-replications; each macro-replication either rejects or accepts the simulation model. We use a generator proposed by L'Ecuyer (1999), called MRG32k3a, with a cycle length of the order 2^{191}. We select seeds randomly.

6. Results of the Monte Carlo Experiments

For our validation statistic T^* with $\alpha = 0.10$ we find Monte Carlo estimates of the *type I* error probability (say) $\hat{\alpha}$ of 0.028 for k = 10 and 0.088 for k = 1,000. So, bootstrapping our statistic does give an acceptable - albeit conservative - $\hat{\alpha}$: We do not reject H_0: $E(\hat{A} |\ H_0) \leq \alpha$ where $\hat{A}$ denotes the Monte Carlo estimator with observed values $\hat{\alpha}$. (Bonferroni implies conservatism.)

Next we estimate the *type II* error probabilities β. Table 1 shows their complements, the *power* 1 - β, for short and long runs respectively. Obviously, our statistic has more power as the simulated traffic load $\tilde{\rho}$ deviates more from the real load ρ = 1. However, at our high traffic rates, the short run does not give estimated performance measures accurate enough to detect serious non-validity (simulated traffic rate is up to 40% wrong): Throwing a coin has more power! Our long runs, however, have more than 80% probability of detecting traffic rate differences of only 4%.

Table 1: Estimated Power $1 - \hat{\beta}$ of Validation Statistic T for M/M/1 with Varying Simulated Traffic Rate $\tilde{\rho}$
(Real Traffic Rate $\rho = 1$, Number of Simulation Replicates $s = 10$, Number of Runs $n = 10$, Nominal $\alpha = 0.10$, Bootstrap Sample Size $b = 1000$)

(A) Number of Customers per Run $k = 10$

$\tilde{\rho}$	$1 - \hat{\beta}$
0.8	.394
0.9	.149
1.0	.028
1.2	.072
1.4	.353

(B) $k = 1{,}000$

$\tilde{\rho}$	$1 - \hat{\beta}$
0.96	.874
0.98	.404
1.00	.088
1.01	.338
1.04	.808

7. Conclusions and Future Research

In general, the *bootstrap* is a versatile tool that enables estimation of the distribution of any statistic (say) $T(\boldsymbol{Z})$, for any type of distribution for $\boldsymbol{Z}$. However, this tool requires mastering the art of modeling: The analysts still have to interpret their problems.

More specifically, for the *validation* of simulation models we focused on a statistical test for trace-driven simulations with IID responses Y. We investigated a specific validation statistic $T(X, Y)$ with IID real response X; the trace makes X and Y cross-correlated. We developed a bootstrap technique that uses runs, while conditioning on the trace.

We applied the *classic* bootstrap technique to a *specific data set*, namely $\boldsymbol{Z} = (Y^{(r)}, Y^{(r')})$ with $r, r' = 1, ..., s$, and $r' \neq r$. This gives an estimated 1- α confidence interval for the bootstrapped statistic T^*. Next the *real* performance measure X is used to compute $T(X, Y)$, and to test the null-hypothesis of equal means. That hypothesis is rejected if T falls outside the bootstrapped confidence interval.

To illustrate this test, we used M/M/1 simulations. Whether our Monte Carlo results also hold for other simulations, requires further research. In the mean time, the current results *might* be seen as rules of thumb.

Our experiments gave the following *main conclusion*. The validation statistic gives an acceptable - but conservative - type I error rate $\hat{\alpha}$. It has good power if runs are long enough, considering the traffic rate.

Kleijnen et al. (2000) gives many more details. For example, they discuss three cases for the number of simulation replicates, each with different bootstrap techniques, namely s is 1, 2, or more (we discussed only $s = 10$). They show that conditional resampling indeed yields more powerful tests than nonconditional sampling does. They prove that the minimum value for the bootstrap sample b is $2/\alpha - 1$; of course, this minimum b gives smaller power. They perform a 2^3 Monte Carlo experiment for the factors k, n, and ρ. They also examine M/G/1 simulation models where G stands for gamma distributed service times; the real system remains M/M/1. Finally, they simulate other priority rules, namely shortest processing time (SPT) and longest processing time (LPT).

In *future research* we might extend our bootstrapping to other terminating simulations (e.g., random runlength K instead of constant k; queueing networks instead of single-server systems), steady-state simulations (with IID subruns), and non-stationary simulations (conditioning on non-stationary trace variables ***A*** when bootstrapping).

Whereas we use runs, EFRON uses overlapping blocks. Such a sampling procedure has also been explored in non-terminating, stationary simulation; see Sherman (1995).

We might also study a more general null-hypothesis: $|E(X) - E(Y)| < \delta$ with δ some positive constant (instead of zero).

Since bootstrapping uses simulation (to resample the original values $\boldsymbol{Z}_i$), typical simulation problems may be further explored. For example, in quantile estimation we may save computer time by using Wald's sequential probability ratio test (SPRT); see Ghosh and Sen (1991) (instead of taking a fixed bootstrap sample size *b*). Variance reduction techniques - such as antithetic and importance sampling - may also be applied; see Hall (1992).

Appendix

We give a theoretical justification for our proposed method by considering its asymptotic behavior as *n* tends tot infinity.

The basic form of the result is most easily understood for the case $s = 1$. We discuss this case first, and then show how it extends to $s > 1$. When $s = 1$, $T_4 = \Sigma Z_i/n$ is formed from a univariate random sample, $\{Z_i = X_i - Y_i;\ i = 1, ..., n\}$. Denote a corresponding bootstrap sample by $\{Z_i^* = X_i^* - Y_i^*;\ i = 1, ..., n\}$ and let $T_4^* = \Sigma Z_i^*/n$ be the value of T_4calculated from such a bootstrap sample. Then we have the following theorem.

Theorem 1. If $E(Z^2) < \infty$, then

$$E(T_4^*) = T_4 \rightarrow E(T_4), \text{ with probability } 1$$

and

$$\sup_z |Pr[\sqrt{n}(T_4 - E(T_4)) \leq z] - Pr[\sqrt{n}(T_4^* - T_4) \leq z]| \rightarrow 0 \tag{1}$$

as $n \rightarrow \infty$.

Proof: This result is proved as Theorem 6.7 in Hjorth (1994), where our *Z* is Hjorth's *X*. The key argument is that both T_4 and T_4^* are asymptotically normally distributed, and that, with probability 1, both $E(T_4^*)$ (= T_4) $\rightarrow E(T_4)$ and $Var(T_4^*) \rightarrow Var(T_4)$ as $n \rightarrow \infty$. The limiting distribution of T_4^* is thus exactly the same as that of T_4. See also Singh (1981) and Bickel and Freedman (1981). □

The fact that the two limiting distributions of T_4 and T_4^* are identical, is asymptotical justification for the construction of the confidence intervals described in the main text, to test if $E(Z) = 0$; that is, if $E(X) = E(Y)$.

The case $s = 2$ is almost identical to the case $s = 1$. Thus we take our 'original' sample as being $\{Z_i = Y_i^{(2)} - Y_i^{(1)};\ i = 1, ..., n\}$, and form bootstrap samples from this sample. Again Theorem 1 applies to show that the bootstrap distribution of T_4^* tends to that of T_4. However, here T_4 is being used to compare the outputs from two simulated samples both obtained from the same simulation model. Thus its distribution is that obtained in the knowledge that both samples $\{Y_i^{(j)}\}$ with $j = 1, 2$ are drawn from the same distribution. The null hypothesis that the $\{X_i\}$ have the same distribution as either of the simulated samples $\{Y_i^{(j)}\}$ $j = 1, 2$ can then be tested. We simply take the two test statistics $T_4^{(j)} = \sum_{i=1}^{n} (X_i - Y_i^{(j)})/n$ where the real data sample $\{X_i\}$ is compared with each of the two simulated samples $\{Y_i^{(j)}\}$ $j = 1, 2$. These can then be compared with quantiles of the bootstrap distribution of T_4^*, to see if they can be regarded as arising from the same distribution.

Consider now the case $s > 2$. Then (1) still applies if we can again show that T_4 and T_4^* are asymptotically normally distributed, and that, with probability 1, both $E(T_4^*)$ (= T_4)→ $E(T_4)$ and $Var(T_4^*) \rightarrow Var(T_4)$ as $n \rightarrow \infty$. There are two cases: unconditional and conditional respectively.

Case 1: The unconditional case is where the bootstrap sample has the form

$$\{Z_i^* = Y_{J(i)}^{U(i)} - Y_{J(i)}^{V(i)};\ i = 1,\ ...,\ n\} \tag{2}$$

where the $J(i)$, $i = 1, 2, ..., n$, are i.i.d. random variables uniformly distributed over the subscripts $\{i = 1, 2, ..., n\}$ and the $(U\,(i), V\,(i))$, $i = 1, 2, ..., n$ are i.i.d. pairs of random values selected from the $s(s - 1)$ distinct pairs $C = \{(k, l); k, l = 1, 2, .., s, k \neq l\}$, with all pairs being equally likely to be selected. The Z_i^* are identically distributed and from the definition $T_4^* = \Sigma Z_i^*/n$, it follows that T_4^* is asymptotically normal if it has finite mean and finite non-zero variance. Now

$$E(T_4^*) = E(\sum Z_i^*/n) = \frac{1}{s(s-1)} \sum_{(u,v) \in c} \mu_n\ (u,\ v)$$

where

$$\mu_n\ (u,\ v) = \frac{1}{n} \sum_{i=1}^{n} (Y_i^u - Y_i^v).$$

But for each fixed pair (u, υ), the $(Y_i^u - Y_i^\upsilon)$, $i = 1, 2, ..., n$ are mutually independent. Therefore by the strong law of large numbers $\mu_n\,(u, \upsilon) \rightarrow E(T_4)$ almost surely. This applies to each pair (u, υ), and as s is fixed it follows that $E(T_4^*) \rightarrow E(T_4)$ with probability 1 also.

An analogous argument also shows that $Var(T_4^*) \rightarrow Var(T_4)$ with probability 1.

This completes the proof for the unconditional case, showing that Theorem 1 still applies.

Case 2: Consider now the conditional case proposed in the main text. A bootstrap sample in this case has the form

$$\{Z_i^* = Y_i^{U(i)} - Y_i^{V(i)};\ i = 1,\ ...,\ n\} \tag{3}$$

where again $(U\,(i), V\,(i))$, $i = 1, 2, ..., n$ are defined as below (2). We thus have

$$E(Z_i^*) = \frac{1}{s(s-1)} \sum (Y_i^u - Y_i^v) \tag{4}$$

and

$$E(Z_i^{*2}) = \frac{1}{s(s-1)} \sum_{(u,v) \in c} (Y_i^u - Y_i^v)^2 \tag{5}$$

and it follows that $E(T_4^*)$ and $Var(T_4^*)$ are exactly the same as in the unconditional case. However, the form of the moments defined in (4) and (5) shows that the Z_i^* are not identically distributed. Thus we need an additional assumption to guarantee that T_4^* is asymptotically normal.

Theorem 2. Let T_4^* be calculated from the conditional bootstrap sample (3) where $s > 2$. Let $\tau = E_{\boldsymbol{Y}}\,[Z^* - E(Z^*)]^2 < \infty$, and $\kappa = E_{\boldsymbol{Y}}\,|\,Z^* - E(Z^*)|^3 < \infty$, where the outer expectations are taken with respect to $\boldsymbol{Y} = (Y^{(1)}, Y^{(2)}, ..., Y^{(s)})$, the s observations simulated. Then with probability 1,

$$\sup_{z} |Pr[\sqrt{n}(T_4 - E(T_4)) \le z] - Pr[\sqrt{n}(T_4^* - T_4) \le z]| \to 0$$

Proof: Let $B_n = \sum_{i=1}^{n} Var(Z_i^*)$, $C_n = \sum_{i=1}^{n} E(|Z_i^* - E(Z_i^*)|^3)$. Then by the strong law of large numbers $n^{1/2} B_n^{-3/2} C_n \to \tau^{-3/2}\kappa$ with probability 1 as $n \to \propto$. Thus $B_n^{-3/2} C_n \to 0$ with probability 1 as $n \to \propto$. It follows from Lyapunov's Theorem (given in Petrov (1995) as Theorem 4.9, for example) that T_4^* is asymptotically normally distributed with probability 1.

As we have already shown, with probability 1, $E(T_4^*) \to E(T_4)$ and $Var(T_4^*) \to Var(T_4)$, so the Theorem follows. □

References

Bickel, P.J. and Freedman, D.A. (1981) Some asymptotic theory for the bootstrap. *Annals of Statistics,* 9, 1196-1197

Efron, B. and R.J. Tibshirani (1993), *Introduction to the Bootstrap*. Chapman & Hall, New York

Ghosh, B.K. and P.K. Sen (1991), *Handbook of Sequential Analysis.* Marcel Dekker, New York

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York

Hjorth, J.S.U. (1994) *Computer Intensive Statistical Methods,* Chapman & Hall, London

Kleijnen, J.P.C. (2000), Strategic directions in verification, validation, and accreditation research: a personal view. *Proceedings of the 2000 Winter Simulation Conference,* edited by J.A. Joines, R.R. Barton, K Kang, and P.A. Fishwick, pp. 909-916

---, B. Bettonvil, and W. Van Groenendaal (1998), Validation of trace driven simulation models: a novel regression test. *Management Science*, 44, pp. 812-819

---, R.C.H. Cheng, and B. Bettonvil (2000), Validation of trace-driven simulation models bootstrap tests. *Proceedings of the 2000 Winter Simulation Conference*, edited by J.A. Joines, R.R. Barton, K Kang, and P.A. Fishwick, pp. 882-892

Law, A.M. and W.D. Kelton (2000), *Simulation Modeling and Analysis*, third edition, McGraw-Hill, Boston

L'Ecuyer, P.L. (1999), Good parameter sets for combined multiple recursive random number generators. *Operations Research*, 47, no. 1, pp. 159-164

Petrov, V.V. (1995), *Limit Theorems of Probability Theory,* Oxford University Press, Oxford

Schmeiser, B. (1982), Batch size effects in the analysis of simulation output. *Operations Research,* 30, no. 3, pp. 556-568

Sherman, M. (1995), On batch means in the simulation and statistics communities. *Proceedings of the 1995 Winter Simulation Conference,* edited by C. Alexopoulos, K. Kang, W.R. Lilegdon, and D. Goldsman, pp. 297-302

Singh, K. (1981) On the asymptotic accuracy of Efron's bootstrap. *Annals of Statistics*, 9, 1187-1195

Acknowledgment

Cheng and Kleijnen received financial support through the NATO Collaborative Research Grants Programme, for their project 'Sensitivity analysis for improved simulation modeling'. Two anonymous referees provided very thoughtful and detailed comments.