Proceedings of the 2000 Winter Simulation Conference
J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, eds.

VALIDATION OF TRACE-DRIVEN SIMULATION MODELS: MORE ON BOOTSTRAP TESTS

Jack P.C. Kleijnen

Department of Information Systems (BIK)/
Center for Economic Research (CentER)
Tilburg University (KUB)
5000 LE Tilburg, Netherlands

Russell C.H. Cheng

Department of Mathematical Sciences
University of Southampton, Highfield,
Southampton
SO17 1BJ, UK

Bert Bettonvil

Department of Information Systems (BIK)/
Center for Economic Research (CentER)
Tilburg University (KUB)
5000 LE Tilburg, Netherlands

ABSTRACT

'Trace-driven' or 'correlated inspection' simulation means that the simulated and the real systems have some common inputs (say, arrival times) so the two systems' outputs are cross-correlated. To validate such simulation models, this paper formulates six validation statistics, which are inspired by practice and statistical analysis; for example, the simplest statistic is the difference between the average simulated and real responses. To evaluate these validation statistics, the paper develops novel types of bootstrapping based on subruns. Three basic bootstrap procedures are devised, depending on the number of simulation replicates: one, two, or more replicates. Moreover, for the case of more than two replicates the paper considers conditional versus unconditional resampling. These six validation statistics and four bootstrap procedures are evaluated in extensive Monte Carlo experiments with single-server queueing systems. The main conclusion is that bootstrapping of the simplest validation statistic gives the correct type I error probability, and has relatively high power.

1 INTRODUCTION

Validation has many aspects; for a recent review and references see Kleijnen (1999). In this paper, however, we limit ourselves to statistical testing of the validity of trace-driven simulations.

Consider the following 'trace-driven' simulation; also see Table 1. The simulated and the real systems have some common inputs (say) $\boldsymbol{A}$; for example, the same historical sequence of arrival times (we use capital letters for random variables, lower-case letters for realized values, and bold letters for matrices including vectors). The real system generates a time series of outputs $W_{i;\,t}$ whereas the simulation generates outputs $V_{i;\,t}$ with $i = 1, ..., n$ and $t = 1, 2, ..., k$; for example, sojourn time of job t on day i. To evaluate the real system, its manager characterizes the output time series by a single performance measure (response) X_i; for example, average sojourn time on day i. To validate the simulation statistically, this real performance X is compared with the simulated performance (say) Y - for the same situation (same circumstances, same scenario) characterized by the trace $\boldsymbol{A}$. But *how* should we compare X and Y?

Some solutions are presented in Moors and Strijbosch (1998), but we focus on Kleijnen, Bettonvil, and Van Groenendaal (1998), abbreviated here to KLEIJ. Like KLEIJ we assume that all simulation responses Y are identically and independently distributed (i.i.d.). More specifically, each subrun starts in the empty state, and stops after a fixed number k of jobs. The real responses X are also i.i.d. Unlike KLEIJ we do not assume that (X_i, Y_i) are bivariate normal. Indeed, in case of short subruns (say, $k = 10$) the responses are seriously nonnormal. This nonnormality - together with a small n (number of subruns) - is not well handled by conventional non-bootstrap techniques. (Obviously, 'trace-driven' simulation implies that the two members of the pair (X_i, Y_i) are cross-correlated.)

We suppose that the simulation model has at least one more input variable (e.g., service time) not recorded on the trace, so this input is sampled using a pseudorandom number stream $\boldsymbol{R}$. There are s simulation replications (using the same trace $\boldsymbol{A}_i$), which yield $Y_i^{(r)}$ with $r = 1, ..., s$. We distinguish three cases for s, namely 1, 2, or more - namely, five or ten.

To solve this problem, we use *bootstrapping*, which in general samples - randomly with replacement - i.i.d. observations; see the seminal book on bootstrapping (outside simulation), Efron and Tibshirani (1993), here abbreviated to EFRON. (Other monographs on bootstrapping are Davison and Hinkley (19), Mooney and Duval (1993), and Shao and Tu (1995).)

Table 1: Trace-driven Simulation

	Subrun number *i*
Trace:	$\boldsymbol{A}_1 \quad \ldots \boldsymbol{A}_i \ldots \boldsymbol{A}_n$
Real performance:	$X_1 \quad \ldots X_i \ldots \quad X_n$
Simulated performance:	
replicate 1	$Y_1^{(1)} \ldots Y_i^{(1)} \ldots Y_n^{(1)}$
...	
replicate *r*	$Y_1^{(r)} \ldots Y_i^{(r)} \ldots Y_n^{(r)}$
...	
replicate ***s***	$Y_1^{(s)} \ldots Y_i^{(s)} \ldots Y_n^{(s)}$

We wish to test the *hypothesis* that the simulation model is valid. For hypothesis testing through bootstrapping outside simulation we refer to EFRON and also Shao and Tu (1995, pp. 176, 189). Our main discovery will be: *one simulation replicate is certainly a valid model for another simulation replicate*. So if $s \geq 2$ we can obtain the bootstrap distribution of any validation statistic under the null-hypothesis of a valid trace-driven simulation model!

Note that - instead of generating responses through *bootstrapping* - we may generate more *simulation* responses. In practice, however, replicating a simulation generally requires much more computer time than bootstrapping a simulation. We assume that the number of simulation replicates (symbol *s*) is given, and is small compared with the bootstrap sample size *b*. (Breiman 1992, p. 750 also discusses bootstrapping versus replicating, but not in a simulation context.)

To provide some background of our research, we now summarize the literature on bootstrapping in simulation. Friedman and Friedman (1995) provide two academic examples. Kim, Willemain, Haddock, and Runger (1993) formulate their so-called 'threshold' bootstrap for the analysis of autocorrelated simulation outputs. Several authors investigate bootstrapping of empirical input distributions in simulation: Barton and Schruben (1993), Cheng (1995), Cheng and Holland (1997), and Pritsker (1998). Bootstrapping for validation of metamodels is done by Kleijnen, Feelders, and Cheng (1998). A summary of the present paper is Kleijnen, Cheng, and Bettonvil (2000).

Our *main conclusion* will be: if a trace-driven simulation model is run more than twice ($s > 2$), then bootstrapping any statistic gives acceptable (albeit conservative) type I error probability; the simplest statistic (the average deviation) has good power compared with the more complicated statistics.

The remainder of this article is organized as follows. §2 summarizes KLEIJ's F-statistic based on regression analysis, and proposes five more validation statistics. §3 recapitulates EFRON's bootstrapping of time series; EFRON uses 'blocks', which we interpret as terminating subruns. §4 derives three bootstrap procedures for trace-driven simulations, using one, two, or more than two simulation replications per subrun; moreover, in case of more than two replicates the resampling may be either conditional or unconditional. To evaluate these six validation statistics and four bootstrap techniques, §5 designs a Monte Carlo experiment with 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.

2 SIX TESTS FOR VALIDATION

The bootstrap enables estimating the distribution of *any statistic*, provided the statistic is a continuous function of the observations (e.g., the median is not a continuous function). For the validation of trace-driven simulations we investigate six statistics, denoted as T_1 through T_6.

KLEIJ calls a simulation model *valid* if the real and the simulated systems have (i) identical means (say) $\mu_x = \mu_y$ and (ii) identical variances $\sigma_x^2 = \sigma_y^2$. To test this composite hypothesis, KLEIJ computes the differences $D_i = X_i - Y_i$ and the sums $Q_i = X_i + Y_i$, and regresses D on Q: $E(D \mid Q = q) = \gamma_0 + \gamma_1 q$. The null-hypothesis then becomes H_0: $\gamma_0 = 0$ and $\gamma_1 = 0$. To test this H_0, KLEIJ computes the two Sums of Squared Errors or SSEs that correspond with the 'full' and the 'reduced' regression model: $SSE_{full} = \sum (D_i - \hat{D}_i)^2$ with $\hat{D}_i = C_0 + C_1 Q_i$ where C_0 and C_1 are the Ordinary Least Squares (OLS) estimators of γ_0 and γ_1; and $SSE_{reduced} = \sum D_i^2$. These two SSEs give the first validation statistic:

$$T_1 = \frac{[\mathrm{SSE}_{\mathrm{reduced}} - \mathrm{SSE}_{\mathrm{full}}]/2}{\mathrm{SSE}_{\mathrm{full}}/(n-2)}. \qquad (1)$$

If X_i and Y_i are n.i.i.d. (see §1), the statistic in Equation (1) has an F-distribution with 2 and n - 2 degrees of freedom (d.f.). If this statistic is significantly high, then KLEIJ concludes that the simulation model is not valid.

We propose another validation statistic with intuitive appeal to simulation practitioners, namely the *average absolute prediction error*, $T_2 = \sum |D_i| / n$ (also see Kleijnen and Sargent, 1999).

A third statistic related to the two preceding statistics is the *mean squared deviation* (MSE), $T_3 = \sum D_i^2 / n$.

A fourth statistic is the *average deviation*, $T_4 = \sum D_i / n = \bar{X} - \bar{Y}$. A disadvantage of this statistic is that positive model errors may compensate negative errors, and vice versa. (This phenomenon may be ignored if a wrong simulation model always underestimates - or always overestimates - the real response whatever the trace is; moreover, this statistic allows bootstrapping in case of a single simulation run; see §4.1.)

The next statistic is the *average relative error*, $T_5 = \sum (Y_i/X_i)/n$, which is often used in practice. Obviously this statistic assumes that no X_i is zero; actually, the event $X_i = 0$ may occur with non-negligible probability in queueing applications with empty starting states, no excessively saturated traffic rates, and short subruns (see §6).

Finally, T_6 compares $\hat{F}_x$ and $\hat{F}_y$, the estimated distribution function (EDF) computed from the n observations on X and Y respectively:

$$T_6 = \int_{-\infty}^{\infty} |\hat{F}_x(z) - \hat{G}_y(z)| dz \quad . \qquad (2)$$

Note that in Equation (2) we use the L_1 norm, not the L_2 or the L_∞ norms. KLEIJ's statistic T_1 also tests equality of variances, whereas T_2 through T_5 consider only equality of means. More criteria or measures for model selection are examined in detail in the monograph by Linhart and Zucchini (1986).

3 EFRON's BOOTSTRAP FOR TIME SERIES

EFRON (p. 91) assumes a sample of n i.i.d. observations $\boldsymbol{Z}_i$ with $i = 1, ..., n$. (Hence, in our case we define $\boldsymbol{Z} = (X, Y)$; see §1.) EFRON summarizes the sample data through a statistic $T = s(\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n)$. (In our case: $T_j = s_j(\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n)$ with $j = 1, ..., 6$.) *Bootstrapping* means that the original values $\boldsymbol{z}_i$ 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 bootstrap 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 procedure 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. (Alternative confidence intervals are discussed in EFRON and Shao and Tu 1995). This interval can be used for hypothesis testing, as we shall see; also see EFRON (p. 169).

We saw that this bootstrap assumes i.i.d. sample observations $\boldsymbol{Z}_i$, but EFRON (pp. 99-102) also presents a bootstrap for *time series*, called 'moving blocks' (also see Shao and Tu 1995, pp. 387-392, 407-415). In our simulation context we interpret these 'blocks' as *subruns*. So we have n non-overlapping subruns, each starting in the empty state and each of length k; we do not eliminate the transient phase. We shall elaborate our approach in the next section.

4 BOOTSTRAP OF VALIDATION TESTS IN TRACE-DRIVEN SIMULATION

We assume a 'reasonable' number of i.i.d. subruns; more specifically, we use the same numbers as KLEIJ (p. 815): n is either 10 or 25. We distinguish three situations for the number of simulation runs, for which we develop different bootstrapping techniques: s is 1, 2, or more.

4.1 A Single Simulation Run: $s = 1$

By assumption, the n pairs (X_i, Y_i) are mutually independent (as the n subruns are assumed independent). Moreover, these pairs are identically distributed if we do not condition on the trace variable $\boldsymbol{A}_i$; we assumed the latter variable to be i.i.d. So we bootstrap the n original pairs, which gives the n bootstrap pairs (X_i^*, Y_i^*). These bootstrap pairs result in the bootstrap validation statistics T_1^* through T_6^*, *albeit not necessarily under the null-hypothesis of a valid trace-driven simulation model*.

We repeat this bootstrapping b times, to obtain an estimated 1 - α confidence interval for each validation statistic. We have an intuitive *target* or hypothesized value for T_4 ($= \sum D_i/n$), namely zero. The two-sided confidence interval ranges from the lower $\alpha/2$ quantile to the upper 1 - $\alpha/2$ quantile of the bootstrap distribution. If the confidence interval does no cover this target value, then we reject the simulation model.

We follow a similar approach for T_5 ($= \sum (Y_i/X_i)/n$), but now with a target value of one.

We have no target values for the other four statistics. However, we may compare the first statistic, with the tabulated 1 - α quantile of the F-statistic with 2 and n - 2 degrees of freedom, $F_{2,\,n-2}^{1-\alpha}$ (no bootstrapping). Moreover, for this statistic we first apply the normalizing logarithmic transformation: replace x by $\log(x)$ and y by $\log(y)$ provided x_i and y_i are not zero (also see KLEIJ).

4.2 Only Two Simulation Replicates: $s = 2$

When $s = 2$ we bootstrap the two replicates of the simulation model: we replace the pair (X_i, Y_i) by $(Y_i^{(1)}, Y_i^{(2)})$ This yields a bootstrap confidence interval per statistic T, *under the null-hypothesis of a valid trace-driven simulation model*.

We also have two observations on each original validation statistic under the alternative hypothesis, namely $T = s((X_1, Y_1^{(r)}), ..., (X_n, Y_n^{(r)}))$ with $r = 1, 2$. We reject the simulation model if any of these two observations on T falls outside the 1 - α/2 bootstrap confidence interval: we use α/2 instead of α because of Bonferroni's inequality (obviously we may also replace 'any' by 'the maximum').

4.3 More than Two Simulation Replicates: $s > 2$

When $s > 2$ we proceed similarly to the case $s = 2$. However, we now distinguish two approaches: (a) condition on the trace; (b) do not condition on the trace.

(a) *Conditioning*: From each column i of Table 1 we sample two observations $Y_i^{(r)}$ and $Y_i^{(r')}$ with $r \neq r'$ (in the original sample the probability of a pair with identical values is zero in case of continuous X and Y, so we require $r \neq r'$). From these n bootstrap pairs we compute the validation statistic T^*. After b repetitions we compute a $1 - \alpha$ confidence interval for this T^*, as in the case $s = 2$.

(b) *No conditioning*: This approach assumes that the traced variables $\boldsymbol{A}_i$ are i.i.d. So now we resample n pairs from the whole table. More precisely, first we sample one value from the $s \times n$ values of Y; next we sample without replacement a second value from the remaining $sn - 1$ values, giving one bootstrap pair; the next pair is sampled after replacing the preceding pair, etc.

Let us compare approaches (a) and (b), focusing on the simplest validation statistic T_4. Then we see that the expected values of all differences between replicated simulation responses are zero, in both approaches. Their variances, however, are smaller in approach (a): *blocking* is a well-known variance reduction technique in the design of experiments. So we expect conditional resampling to yield more powerful tests (this will turn out to be true: see §6).

Analogous to the $s = 2$ case, we again compare one real response X_i with each of the s simulated responses $Y_i^{(r)}$. We reject the simulation model if any of these s values falls outside the $1 - \alpha/s$ confidence interval (Bonferroni).

4.4 Asymptotic Results: Large n

In the Appendix we derive asymptotic results for the simplest bootstrap validation statistic T_4^* (this statistic will turn out to have the greatest practical relevance; see §6). We can prove that as n tends to infinity, the EDF of T_4^* tends uniformly to the EDF of the original statistic T_4, for all four bootstrap methods defined in §4.1 through §4.3. This uniform convergence is important if confidence intervals with the correct coverage are to be constructed. Of course, this convergence is only asymptotic; our Monte Carlo experiments in §6 estimate small-sample performance.

4.5 Minimal Bootstrap Sample Size

A classic value for b is 1,000; see EFRON (p.275), Andrews and Buchinsky (1996), and also Barton and Schruben (1993) and Shao and Tu (1995, pp. 206-210). We shall use this classic value, but also a much smaller value.

Actually, we are not interested in the whole distribution function (say) g of the bootstrapped statistic T^*, but only in its $\alpha/2$ and $1 - \alpha/2$ quantiles (we reject the null-hypothesis if the value of the original statistic T does not fall between these two quantiles). To estimate this distribution function g, we sort the b observations on T^*, which gives $T^*_{(1)}, ..., T^*_{(b)}$. Hence, $g(T^*_{(1)}), ...,$ $g(T^*_{(i)}), ..., g(T^*_{(b)})$ is an ordered sample from a uniform distribution on [0, 1). The expected value of $g(T^*_{(i)})$ is $i/(b + 1)$. Consequently, if we take the *minimal* bootstrap sample size, then our estimator of the lower $\alpha/2$ quantile is the smallest order statistic, namely $T^*_{(1)}$. Likewise the largest order statistic $T^*_{(b)}$ estimates the upper $1 - \alpha/2$ quantile. It is easy to prove that the *minimum value* for b is

$$b_{\min} = (2/\alpha) - 1. \qquad (3)$$

For example, $\alpha = 0.1$ gives $b = 19$; we shall use this value (besides the classic value of 1,000; see §5).

However, when we have more than one simulation replicate ($s > 1$), then we apply *Bonferroni*'s inequality so α is replaced by α/s. For example, for $\alpha = 0.1$ and $s = 10$ Equation (3) gives 199 (still much smaller than 1,000). Actually, we shall report on $b = 19$ even when $s > 1$: we then avoid Bonferroni's inequality by randomly selecting a single value from the s values for the validation statistic computed from the original (non-bootstrapped) observations on X_i and $Y_i^{(r)}$. We reject the simulation model if this one value lies outside the bootstrap confidence interval.

5 DESIGN OF QEUEING EXPERIMENTS

For the type I error rate of the validation tests we use an α of 0.01, 0.05, and 0.10 respectively. These values determine which quantiles of the bootstrap distribution should be used as thresholds. (Of course, the higher α is, the higher the power is.) We focus on $\alpha = 0.10$ because it gives the smallest relative variance for our Monte Carlo results (see §6); besides, this value is the only one that we can use for $b = 19$.

Following KLEIJ, we start with M/M/1 simulation models, which generate 'real' and simulated individual sojourn times W and V. So these models have Poisson arrival and service parameters (say) $\lambda_a = 1/\mu_a$ and $\lambda_s = 1/\mu_s$ where μ_a and μ_s denote the means of the interarrival and service times. We use a tilde to denote a parameter of the simulation model; for example, $\tilde{\lambda}_s$ refers to the simulation model, whereas λ_s denotes the 'real' parameter.

To study the *type I error* of the validation tests, we use a simulation model and a real system with equal service rates (arrival times are on the trace, so simulated and real arrival times are the same); hence simulated and real traffic rates are the same: $\tilde{\rho} = \rho$. We use an imperfect simulation model: the 'real' and the simulated service times use different pseudorandom numbers.

We examine the following three factors - following KLEIJ (p. 815) - in a 2^3 design: (i) number of jobs per subrun, k: 10 and 1,000 (affects the degree of nonnormality); (ii) number of subruns, n: 10 and 25 (affects the convergence of the bootstrap distribution); (iii) real traffic load, ρ : 0.5 and 1.0 (affects the cross-correlation caused by the common trace).

To study the *type II* error, we use unequal simulated and real rates. For real load $\rho = 0.5$ and number of jobs per subrun $k = 1{,}000$ we use $\tilde{\rho} = 0.46$, 0.48, 0.52, and 0.54; for $k = 10$ we use 0.3, 0.4, 0.6, and 0.7. For $\rho = 1$ and $k = 1{,}000$ we use 0.96, 0.98, 1.02, and 1.04; for $k = 10$ we use 0.8, 0.9, 1.2, and 1.4. (For more extreme values of $\tilde{\rho}$ the estimated power reaches 1.)

Still following KLEIJ (p. 815), we use 1,000 macro-replications; by definition, each macro-replication either rejects or accepts a specific simulation model. (Each macro-replication requires b bootstraps; each bootstrap requires kn observations on the real and the simulated individual outputs.) Because we use many pseudorandom numbers, we select our generator with some care: we use a generator proposed by L'Ecuyer (1999), called MRG32k3a with a cycle length of the order 2^{191}. We select seeds randomly.

All six validation tests use the same data $(X_i, Y_i^{(r)})$, which improves the comparison of these tests. The three values for α also give positively correlated results.

To obtain more general results, we extend KLEIJ: we also use M/G/1 simulation models where we let G stand for service times with a *gamma distribution*. (Cheng 1998 gives generators for this distribution family; the exponential distribution belongs to this family.) The real system remains M/M/1. We limit the design to a single combination of the three factors: traffic load 1.0, number of jobs per subrun 1,000, number of subruns 10.

Finally, we extend our Monte Carlo study to simulations with other priority rules, namely *shortest processing time* (SPT) and *longest processing time* (LPT). We use the same factor combination as for M/G/1.

6 MONTE CARLO RESULTS

Our Monte Carlo experiments with various single-server queues result in estimated type I and II error probabilities of our six validation statistics for five bootstrap procedures.

If this *type I error probability* equals the prespecified (nominal) value α, we call the validation test *acceptable*:

$$H_0: E(\hat{A}) = \alpha \qquad (4)$$

where $\hat{A}$ denotes the Monte Carlo estimator of that probability - with values $\hat{\alpha}$. If no statistic satisfies this condition, we accept a *conservative* validation procedure (Bonferroni's inequality implies such conservatism): in Equation (4) we replace = by $\leq$.

Given Equation (4), this error probability has a binomial distribution with variance $\alpha(1 - \alpha)/1000$ (we have 1,000 macro-replications). For example, $\alpha = 0.10$ gives a standard deviation of 0.0095. We use the normal approximation's factor 1.96 (95% confidence interval) to test the significance of the deviation between observed and nominal type I error probability: we reject H_0 if $|\hat{\alpha} - \alpha| > 0.0186$. In case of a conservative, one-sided test we accept an $\hat{\alpha}$ smaller than 0.1156; see the results printed in bold in the tables below. (There is no need for multiple comparisons or joint inferences, which might use Bonferroni.)

If several statistics have acceptable type I error probabilities, then we compare their *estimated type II error probabilities* (power complement).

How to interpret the massive amount of data generated by our Monte Carlo experiments? We think that the *primary user question* is: which validation statistic should be used, given that it is known how many simulation replicates are available? Remember that when $s = 1$ we should bootstrap only those two statistics that have intuitive target values, namely T_4 and T_5 (for T_1 we use the F table).

The answer may also depend on other known characteristics of the given simulation, namely the number of i.i.d. subruns, n.

If the simulation represents a queueing system, then another known characteristic might be the number of customers per subrun (k), the traffic load (ρ), and the queueing discipline (FIFO, LPT, etc.). Some queueing simulations, however, may be much more complicated than the single-server systems that we study, so these characteristics are of secondary interest.

We start our analysis of all these Monte Carlo results by studying $\hat{\alpha}$ (type I error). Though we have 2^3 combinations of ρ, k, and n (see §5), we present data only for the high ρ and the low n; see Table 2. We do give results for both k values, because this factor may exclude the use of certain validation statistics (namely, T_5) and strongly affect non-normality of the performance measures X and Y. Further, for $s = 1$ we also present the statistic T_1 as applied by KLEIJ using the F-table (instead of bootstrapping) after the normalizing transformation $\log(X)$ and $\log(Y)$. Finally, for $s > 2$ we may condition on the trace or not, but Table 2 shows

results for conditioning only: we found that conditioning does indeed improve the power while maintaining the type I error.

Part A gives results for short subruns ($k = 10$).

Case s = 1: Not applicable (N/A) holds for T_2, T_3, and T_6 because they have no practical thresholds; T_5 has a denominator $X_i = 0$ with high probability so it is also N/A. The table look-up of T_1 gives a worse error probability than bootstrapping the simple statistic T_4; nevertheless, even the latter statistic gives significantly high $\hat{\alpha}$.

Case s = 2: Acceptable - though conservative - results are given by bootstrapping T_6.

Case s = 5: Our bootstrapping gives acceptable - but conservative - $\hat{\alpha}$, except for T_2 and T_3.

Case s = 10: Bootstrapping any statistic gives acceptable $\hat{\alpha}$. This case gives results more conservative than $s = 5$: Bonferroni becomes more conservative as s increases.

We can prove that as s increases for fixed n, then the EDFs of the original statistic T and the bootstrap statistic T^* converge. Because this proof is rather technical we do not give it here.

Part B gives results for long subruns ($k = 1{,}000$).

Case s = 1: KLEIJ's procedure gives an acceptable result; in long runs the nonnormality disappears after the log transformation.

Case s = 2: Acceptable but conservative results are again given by bootstrapping T_6.

Case s = 5: Bootstrapping the simple statistic T_4 gives acceptable $\hat{\alpha}$.

Case s = 10: Our bootstrap gives acceptable - but conservative - $\hat{\alpha}$ for any statistic except T_2 and T_3.

Altogether Table 2 suggests the following conclusions.

Case $s = 1$: All validation statistics give observed type I error probabilities significantly higher than the nominal α, except for KLEIJ's procedure when long subruns are used.

Case $s = 2$: Bootstrapping T_6 gives 'best' conservative results.

Case $s = 5$: Bootstrapping the simple statistic T_4 gives acceptable $\hat{\alpha}$.

Case $s = 10$: Bootstrapping any statistic - except for T_2 and T_3 - gives acceptable $\hat{\alpha}$, albeit rather conservative for short subruns.

Further, these conclusions suggest that - for bootstrapped validation - a trace-driven simulation model be *run more than twice* (using different random numbers).

The next question is: which of the acceptable validation statistics has the *highest power*? Table 3 shows the estimated power for these statistics, for a given combination of s and k. We select four simulated traffic rates $\tilde{\rho}$ that differ from the 'real' rate $\rho = 1$ (see the four rows). Obviously, any statistic has more power as the simulated load deviates more from the real load (read within columns). Further, any statistic can detect smaller deviations between real and simulated traffic rates when k is larger (10 versus 1,000). For $s > 2$ the bootstrapped simple statistic T_4 has good power compared with the more complicated statistics.

We also obtain results for other systems than M/M/1/FIFO (see §5). However, given the conclusions so far, we focus on T_4 when interpreting these results. Then it suffices to state that the above conclusions also hold for these systems!

Table 4 gives estimated type I error probabilities in case of the *minimum bootstrap sample size* ($b = 19$). These probabilities are similar to Table 2, though less conservative when $k = 10$.

Our results (not displayed to save space) further show that the power is smaller than in case of a large bootstrap sample size (for $s > 1$ we use Bonferroni's inequality in Table 3, whereas we now randomly select one of the s values; which confounds the effects of small b and using only one of the s values).

7 CONCLUSIONS AND FUTURE RESEARCH

In general, bootstrapping is a versatile tool, as it allows the estimation of the distribution of any statistic $T(\boldsymbol{Z})$ for any type of input distribution for $\boldsymbol{Z}$. However, this tool requires mastering the art of modeling: the researchers still have to interpret their problems. Indeed, EFRON (pp. 115, 383) states 'bootstrapping is not a uniquely defined concept ... alternative bootstrap methods may coexist'.

More specifically, for validation in simulation we focused on statistical tests for the validation of trace-driven terminating simulations with i.i.d. response Y. Given the i.i.d. real response X, we proposed six validation statistics $T_j(X, Y)$ ($j = 1, ..., 6$). The pairs (X, Y) are correlated, and may be non-normally distributed.

We developed different bootstrap methods that vary with the number of simulation replicates (symbol s). All these methods use *subruns*. When we have more than two replicates ($s > 2$), we either condition or we do not condition on the trace.

To evaluated and illustrate the resulting tests, we applied them to single-server queueing simulation models with different priority rules. Whether these Monte Carlo results

hold for other applications, requires further research; the current results *might* be seen as rules of thumb. These rules are as follows.

Case s = 1: Most validation statistics give type I error probabilities higher than the nominal α. If a normalizing transformation can be found, then follow KLEIJ; that is, use the F-table without bootstrapping.

Case s = 2: Bootstrapping T_6 gives acceptable - but conservative - results.

Case s > 2: Many statistics give acceptable - possibly conservative - $\hat{\alpha}$. So we recommend to run a trace-driven simulation model more than twice. The simplest statistic, namely the *average deviation* $T_4 = \sum D_i/n$, has good power compared with the more complicated statistics.

A surprisingly small bootstrap sample size might suffice to quickly decide on the validity of a simulation model. Then, little extra computer time is needed for bootstrapping. Nevertheless, if the small bootstrap sample results in a borderline value for the validation statistic, then we recommend a larger bootstrap sample - especially since in practice bootstrapping requires far less computer time than simulation does.

In *future research* we might extend our analysis to other terminating simulations (e.g., queueing networks), and to steady-state and non-stationary simulations. For example, if the trace does not remain stationary over subruns, then we may condition and resample one response from each subrun (column in Table 1; see §4.3).

Whereas we use subruns, EFRON uses overlapping blocks; also see Shao and Tu (1995, pp.391-392). Such a sampling procedure has also been explored in non-terminating, stationary simulation: see Sherman (1995).

We might also study a complication that KLEIJ mentioned but did not solve: a more general null-hypothesis states that the difference between the real and the simulated systems' expected values is smaller than some positive constant δ, not necessarily zero:: $|E(X) - E(Y)| < \delta$.

Since bootstrapping uses simulation (Monte Carlo for resampling the original values z), 'typical' simulation problems may be further explored in a bootstrapping context. For example, the determination of the sample size in quantile estimation is a standard problem in simulation; see Alexopoulos and Seila (1998). We add that computer time may be saved by not taking a fixed sample size b for the bootstrap. Instead, we may use Wald's sequential probability ratio test (SPRT); see Ghosh and Sen (1991). Variance reduction techniques may also be applied to bootstrapping. Indeed, Shao and Tu (1995, pp. 221 - 2228) discuss antithetic and importance sampling in bootstrapping.

We assumed that the number of replicates s is so small that bootstrapping is needed. If, however, (say) $s = 100$, then we can use classic tests such as Student's t test, a distribution-free test (e.g., sign test, rank test), or goodness-of-fit tests (see D'Agostino and Stephens (1986) and Vincent (1998)).

APPENDIX: CONVERGENCE OF EDFs OF T_4^* AND T_4 AS *n* INCREASE

We give a theoretical backing for the conditional sampling bootstrap method described in § 4.3: for T_4 (the statistic we recommend) we show that $T_4^* - E(T_4^*)$ has the same asymptotic distribution as $T_4 - E(T_4)$, as n tends to infinity.

Conditional sampling is both the most interesting and the most difficult case. Here a bootstrap sample has the form

$$\{Z_i^* = Y_i^{U(i)} - Y_i^{V(i)};\ i = 1, \cdots, n\} \qquad \text{(A-1)}$$

where $(U(i), V(i))$ are i.i.d. pairs of random values selected from the $s(s - 1)$ distinct pairs $C = \{r, r'; r, r' = 1, \ldots, s, r \neq r'\}$, with all pairs being equally likely to be selected. This gives

$$\begin{aligned} E(Z_i^*) &= \frac{1}{s(s-1)} \sum_{(u,\ v)\in C} (Y_i^u - Y_i^v), \\ E(Z_i^{*2}) &= \frac{1}{s(s-1)} \sum_{(u,\ v)\in C} (Y_i^u - Y_i^v)^2. \end{aligned} \qquad \text{(A-2)}$$

Elementary considerations show that $E(T_4^*)$ and $Var(T_4^*)$ are exactly the same as in the unconditional case; moreover with probability 1, $E(T_4^*) \to E(T_4)$ and $Var(T_4^*) \to Var(T_4)$. However the form of the moments in Equation (A-2) shows that the Z_i^* are not identically distributed. Thus we need an additional assumption to guarantee that T_4^* is asymptotically normal.

Theorem: Let T_4^* be calculated from the conditional bootstrap sample in Equation (A-1) 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 $\boldsymbol{Y} = (Y^{(1)},..., Y^{(s)})$, the s observations simulated. Further, let

$$c(z) = Pr[\sqrt{n}\ (T_4 - E(T_4)) \le z]$$
$$c^*(z) = Pr[\sqrt{n}\ (T_4^* - T_4) \le z]$$

Then with probability 1 we have

$$\sup_z \left| c(z) - c^*(z) \right| \to 0. \qquad \text{(A-3)}$$

Proof: Let

$$B_n = \sum Var(Z_i^*),$$
$$C_n = \sum E(|Z_i^* - E(Z_i^*)|^3).$$

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 → ∞. Thus

$$B_n^{-3/2} C_n \to 0$$

with probability 1 as n → ∞. It follows by Lyapunov's Theorem (given in e.g. Petrov (1995) as Theorem 4.9) that T_4^* is asymptotically normally distributed with probability 1.

With probability 1 we have $E(T_4^*) \to E(T_4)$ and $Var(T_4^*) \to Var(T_4)$ so we can apply Theorem 6.7 in Hjorth (1994) (see also Singh (1981) and Bickel and Freedman (1981), to show that Equation (A-3) holds.

REFERENCES

Alexopoulos, C. and A.F. Seila 1998, Output Data Analysis. *Handbook of Simulation*, edited by Jerry Banks, Wiley, New York

Andrews, D.W.K. and M. Buchinsky 1996, On the number of bootstrap repetitions for bootstrap standard error estimates. Cowles Foundation Discussion Paper no. 1141, Yale University, PO Box 208281, New Haven, Connecticut 06520-8281

Barton, R.R. and L.W. Schruben 1993, Uniform and bootstrap resampling of empirical distributions. In *Proceedings of the 1993 Winter Simulation Conference*, 503-508. ed. G.W. Evans et al., IEEE, Piscataway, N.J.

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

Breiman, L. 1992, The little bootstrap and other methods for dimensionality selection in regression: x-fixed prediction error. *Journal American Statistical Association*, 87, no. 419, pp. 738-754.

Cheng, R.C.H. 1995, Bootstrap methods for computer simulation experiments. *Proceedings of the 1995 Winter Simulation Conference*, 171-177. ed. C. Alexopoulos, K. Kang, W.R. Lilegdon, and D. Goldsman.

--- 1998, Random variate generation. *Handbook of Simulation*, edited by J. Banks, Wiley, New York.

--- and W. Holland 1997, Sensitivity of computer simulation experiments to errors in input data *Journal Statistical Computation and Simulation*, 57(1-4): 219-241.

D' Agostino, R.D. and H.A. Stephens, editors 1986, *Goodness-of-fit distributions*. Marcel Dekker, New York

Davison, A.C. and D.V.Hinkley, Bootstrap methods and their application, CUP

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

Friedman, L.W. and H.H. Friedman (1995), Analyzing simulation output using the bootstrap method. *Simulation*, 64(2): 95-100.

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

Hjorth, J.S.U. (1994) *Computer intensive statistical methods*, Chapman & Hall, London.

Kim, Y.B., T.R. Willemain, J. Haddock, and G.C. Runger (1993), The threshold bootstrap: a new approach to simulation output analysis. In *Proceedings of the 1993 Winter Simulation Conference*, 498-502. ed. G.W. Evans, M. Mollaghasemi, E.C. Russell, and W.E. Biles.

Kleijnen (1999). Validation of models: statistical techniques and data availability. *Proceedings of the 1999 Winter Simulation Conference*, 647-654. (ed. by P.A. Farrington, H.B. Nembhard, D.T.Sturrock, and G.W.Evans).

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

---, R.C.H. Cheng, and B. Bettonvil (2000), Validation of trace-driven simulation models: bootstrapped tests. *Management Science*(under review).

--- , A.J. Feelders, and R.C.H. Cheng (1998).Bootstrapping and validation of metamodels in simulation. *Proceedings of the 1998 Winter Simulation Conference*.

--- and R.G. Sargent (1999), A methodology for fitting and validating metamodels in simulation. *European Journal Operational Research* (accepted).

L'Ecuyer, P.L. (1999), Good parameter sets for combined multiple recursive random number generators. *Operations Research*, 47(1).

Linhart, H. and W. Zucchini (1986), *Model selection*. Wiley, New York.

Mooney, C.Z. and R.D. Duval (1993), *Bootstrapping: a nonparametric approach to statistical inference*. Sage Publications, Newbury Park, California 91320.

Moors, J.J.A. and L.W.G. Strijbosch (1998), New proposals for the validation of trace-driven simulations. *Communications in Statistics: Simulation and Computation*, 27(4): 1051-1073.

Petrov, V.V. (1995), *Limit theorems of probability theory*, Oxford University Press, Oxford

Pritsker, A.A. (1998), Life & death decisions. *OR/MS Today*, 25(4): 22-28.

Shao, J and D. Tu (1995), *The jackknife and bootstrap*. Springer-Verlag, New York.

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

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

Vincent, S. (1998), Input data analysis. *Handbook of simulation*, edited by J. Banks, Wiley, New York.

ACKNOWLEDGMENT

Cheng and Kleijnen acknowledge the 'NATO Collaborative Research Grants Programme''s financial support for their project on 'Sensitivity analysis for improved simulation modeling'.

Table 2: Estimated Type I Error Probability of Validation Statistic (T) for Varying Number of Simulation Replicates (s) of M/M/1/FIFO with Number of Customers per Subrun k, Traffic Rate $\rho = 1$; Number of Subruns $n = 10$; Nominal $\alpha = 0.10$; Bootstrap Sample Size b = 1,000; Bold Numbers Denote Acceptable Results

[1)] F-table used (instead of bootstrap) after normalizing transformation log(X) and log (Y)

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

s	T_1	T_2	T_3	T_4	T_5	T_6
1	.212[1)]	N/A	N/A	.174	N/A	N/A
2	**.021**	.142	.180	.172	.180	**.044**
5	**.055**	.127	.142	**.063**	**.046**	**.068**
10	**.024**	**.046**	**.059**	**.028**	**.023**	**.033**

(B) Number of Customers per Subrun k = 1,000

s	T_1	T_2	T_3	T_4	T_5	T_6
1	**.098**[1)]	N/A	N/A	.167	.235	N/A
2	**.027**	.196	.265	.364	.358	**.050**
5	.124	.252	.301	**.107**	.118	.122
10	**.096**	.126	.146	**.088**	**.095**	**.080**

Table 3: Estimated Power of Acceptable Statistics T for Varying Simulated Traffic Rates $\tilde{\rho}$ and Fixed Real Traffic Rate $\rho = 1$ (for Remaining Symbols See Table 2)

$s = 1$; $k = 1{,}000$

$\tilde{\rho}$	T_1
.96	.622
.98	.276
1.02	.264
1.04	.618

$s = 2$; $k = 10$

$\tilde{\rho}$	T_1	T_6
.8	.098	.401
.9	.039	.161
1.2	.172	.045
1.4	.428	.272

$s = 2$; $k = 1{,}000$

$\tilde{\rho}$	T_1	T_6
.96	.249	.661
.98	.086	.265
1.02	.088	.098
1.04	.250	.419

$s = 5$; $k = 10$

$\tilde{\rho}$	T_1	T_4	T_5	T_6
.8	.186	.444	.148	.453
.9	.068	.204	.072	.219
1.2	.220	.142	.175	.108
1.4	.490	.434	.511	.358

$s = 5$; $k = 1{,}000$

$\tilde{\rho}$	T_4
.96	.874
.98	.434
1.02	.335
1.04	.782

$s = 10$; $k = 10$

$\tilde{\rho}$	T_1	T_2	T_3	T_4	T_5	T_6
.8	.098	.377	.350	.394	.088	.401
.9	.039	.165	.185	.149	.025	.161
1.2	.172	.0	.003	.072	.119	.045
1.4	.428	.002	.002	.353	.424	.272

$s = 10$; $k = 1{,}000$

$\tilde{\rho}$	T_1	T_4	T_5	T_6
.96	.534	.874	.873	.865
.98	.239	.404	.415	.391
1.02	.236	.338	.350	.299
1.04	.565	.808	.831	.782

Table 4: Estimated Type I Error Probability Using Small Bootstrap Sample Size $b = 19$ (Remaining Symbols Defined in Table 2)

$k = 10$

s	T_1	T_2	T_3	T_4	T_5	T_6
1	.212[1)]	N/A	N/A	.160	N/A	N/A
2	**.046**	.198	.256	.179	.245	**.061**
5	.118	.185	.210	.118	.173	.139
10	**.100**	.137	.131	**.112**	**.115**	**.105**

$k = 1{,}000$

s	T_1	T_2	T_3	T_4	T_5	T_6
1	**.098**[1)]	N/A	N/A	.150	.210	N/A
2	**.034**	.183	.223	.160	.172	**.058**
5	.121	.178	.199	.120	.126	.124
10	**.096**	.121	.132	**.113**	**.108**	.118

AUTHOR BIOGRAPHIES

JACK P.C. KLEIJNEN is a Professor of Simulation and Information Systems. His research concerns simulation, mathematical statistics, information systems, and logistics; this research resulted in six books and nearly 160 articles. He has been a consultant for several organizations in the USA and Europe, and has served on many international editorial boards and scientific committees. He spent several years in the USA, at both universities and companies, and received a number of international fellowships and awards. More information is provided on his web page: <http://center.kub.nl/staff/kleijnen>.

RUSSELL C.H. CHENG is Professor of Operational Research at the University of Southampton. He has an M.A. and the Diploma in Mathematical Statistics from Cambridge University, England. He obtained his Ph.D. from Bath University. He is Chairman of the U.K. Simulation Society, a Fellow of the Royal Statistical Society, Member of the Operational Research Society. His research interests include: variance reduction methods and parametric estimation methods. He is Joint Editor of the *IMA Journal on Mathematics Applied to Business and Industry.*

BERT BETTONVIL is Associate Professor at the Department of Information Systems of Tilburg University. Educated as a mathematical statistician, his research interests are in the field of statistical aspects of simulation and in research methodology.