Proceedings of the 2010 Winter Simulation Conference
B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, eds.

A FRAMEWORK FOR INPUT UNCERTAINTY ANALYSIS

Russell R. Barton

Smeal College of Business Administration
The Pennsylvania State University
406 Business Building
University Park, PA 16802, USA

Barry L. Nelson
Wei Xie

Dept. of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208, USA

ABSTRACT

We consider the problem of producing confidence intervals for the mean response of a system represented by a stochastic simulation that is driven by input models that have been estimated from "real-world" data. Therefore, we want the confidence interval to account for both uncertainty about the input models and stochastic noise in the simulation output; standard practice only accounts for the stochastic noise. To achieve this goal we introduce metamodel-assisted bootstrapping, and illustrate its performance relative to other proposals for dealing with input uncertainty on two queueing examples.

1 INTRODUCTION

Discrete-event simulation provides a flexible tool to model the behavior of dynamic stochastic phenomena in manufacturing, transportation, communication, health care, finance and many other areas. Simulation output results are finite samples drawn from a population, and so exhibit what we will call intrinsic output uncertainty. In discrete event dynamic simulation, this uncertainty comes from finite run length of the simulation and the finite number of replications that are conducted. Traditional simulation output analysis methods characterize intrinsic uncertainty in simulation output results by constructing parameter estimates of output probability distributions, typically first and second moments and sometimes quantiles. These are often accompanied by either confidence intervals or moments of the output parameter estimates. This analysis allows decision makers to judge risk in decisions made based on simulation model output.

Discrete event simulation models require input probability distributions, which are then used to generate random variates during a simulation run. One important source of output uncertainty is the imperfect rendition of input probability distributions when finite samples of real-world data are used to fit empirical or parametric input distributions. We call this input-model uncertainty. Without including the uncertainty arising from errors in the input distributions, the simulationist has no guidance on how much input data must be collected to provide model outputs with useful accuracy. Worse, the simulationist may ignore this source of uncertainty, and be overly confident in the results of the usual output analysis, an analysis that is conditional on the set of data used to fit the input probability distributions. The error in such conditional confidence intervals can be quite large, as shown by Barton and Schruben (1993, 2001) and Henderson (2003).

Simulation input model uncertainty has been addressed in the literature in essentially two ways: The first is to perform statistical output characterization conditional on the (correctness of) the input probability models, and separately perform sensitivity or uncertainty or robustness analysis of simulation output to changes in simulation input distributions. Descriptions of this approach can be found in Kleijnen (1987, 1994) and Law and Kelton (2000). As noted by Schmeiser in his discussion in Barton et al. (2002), this method has broad applicability, and can handle qualitative uncertainty in model form as well as situations in which no calibrating real-world data exists. This *separate analysis* approach has a significant drawback: it provides no statistical characterization of the combination of both sources of uncertainty.

The second approach is to combine explicit characterization of the impact of input-model uncertainty on simulation output with the characterization of intrinsic output uncertainty to provide confidence intervals for output parameter values. There have been three primary streams of research in this vein: One stream of research has developed methods for a priori known families of parametric input distributions, using MLE properties of the parameters and Taylor series approximations of the expected simulation response, assuming that the correct parametric families of the input distributions are known. Cheng (1994) and Cheng and Holland (1997) used a delta method approximation

to achieve this, which requires an additional number of simulation runs that is linear in the number of input model parameters. Cheng and Holland (1998, 2004) suggested an alternative method that requires only two runs, but yields more conservative confidence intervals. In many cases, the linear approximation of the impact on output that these methods require cannot be supported. Further, this stream of research requires that the parametric family of all input distributions be known with certainty.

A second stream of research is based on Bayesian resampling methods. The distribution of the expected value of the simulation output is characterized running simulations at each of many repeated samplings from a posterior distribution for the parametric input model parameters. The posterior distributions are determined by applying Bayes rule to a prior distribution and observations of real-world data. This Bayesian model averaging (BMA) strategy has been described and refined by Chick (1997, 1999, 2000, 2001), Chick and Ng (2002), Ng and Chick (2001), and Zouaoui and Wilson (2001a, 2001b, 2003). These methods require normality and homogeneity assumptions on the impact of input-model uncertainty, which may not be reasonable. Zouaoui and Wilson (2004) used a different BMA approach to capture both parameter uncertainty and uncertainty across an a priori determined set of possible parametric families. The assumption of homogeneous variance induced by input-uncertainty variability remains, and the t-type confidence interval they described also depends on approximate normality due to input-model uncertainty.

The third stream of research takes a frequentist approach, characterizing the impact of input-model uncertainty on the simulation output using direct resampling and bootstrap resampling methods. This work includes the papers on nonparametric bootstrap approaches by Barton and Schruben (1993, 2001) and Barton (2007), and a parametric bootstrap approach by Cheng and Holland (1997). The percentile intervals do not require normality or homogeneity due to input-model uncertainty. Unfortunately, since the statistic that is bootstrapped is the output of a simulation run, intrinsic output uncertainty causes the statistic not to be a smooth function of the input data, violating a requirement for the asymptotic validity of the bootstrap.

Further, the resampling methods in this third stream and the percentile method in Zouaoui and Wilson (2004) suffer from a failure to distinguish the behavior of intrinsic output uncertainty—sampling error—and input-model uncertainty—input model error. Schmeiser in his discussion in Barton et al. (2002) noted: "Any reasonable version [of a method capturing both modeling error and sampling error] needs to reflect the fundamental difference that sampling error decreases with additional simulation sampling while additional sampling has no impact on modeling error." This means that confidence intervals based on resampling strategies will have coverage that exceeds the nominal value if the simulation effort for each resample is relatively low. Barton identified this error for nonparametric bootstrap methods in Barton et al. (2002) and proposed simulation effort guidelines in Barton (2007).

There has been related work on allocation-of-effort strategies to minimize the joint impact of input-model uncertainty and intrinsic output uncertainty. Lee and Glynn (2003) developed a framework to estimate the distribution of the conditional expectation of a simulation output statistic in the presence of model and parameter uncertainty, and presented methods to minimize the mean squared error of an estimator of the expected value. Steckley and Henderson (2003) further developed the framework and described a kernel approach for estimating the density of the conditional expectation, depending on a single unknown parameter. Ng and Chick (2001) used the delta method in a Bayesian framework to sequentially choose which real-word data to augment with additional samples to reduce the variance of the simulation output. Freimer and Schruben (2002) described two approaches to determine whether additional real-world data should be collected, given a fixed level of simulation effort. In the first approach, upper and lower confidence limits on model parameters are used as parameter values in a factorial set of simulation experiments. If the parameter effects are statistically significant, more real-world data must be collected. In the second approach, the real-world data are bootstrap resampled to generate random values of the parameters, and a random effects model is used to test for significance of parameter uncertainty.

Henderson (2003) gave a detailed review and critique of these approaches. He found no clearly superior method, and suggested the need for a method that is transparent, statistically valid, implementable and efficient. In this paper we describe a metamodel-assisted bootstrapping method and show its use with some simple parametric input models. The metamodel provides the smooth function necessary for bootstrap convergence, and reduces the computational burden of the bootstrap resamples. The strategy distinguishes input-model and intrinsic simulation uncertainty components of the output, a shortcoming of prior bootstrap methods, and models non-homogeneous and non-normal effects of input-model uncertainty, a limitation of Bayesian methods.

The next section provides a formal description of the problem, and describes the metamodel-assisted bootstrapping method. Section 3 compares the coverage of confidence intervals constructed using our method with conditional intervals and intervals based on direct bootstrapping and Bayesian methods, while Section 4 provides theoretical justification for our method. The final section describes the advantages and limitations of the metamodel-assisted bootstrapping method and areas for future research.

2 A FRAMEWORK FOR CONFIDENCE INTERVALS

To account for input-model uncertainty, we introduce the following representation of a computer simulation.

A simulation is a function $g : \omega \to Y$, where ω is an elementary outcome from a basic probability space (Ω, P); to be concrete, one might think of the intrinsic uncertainty ω as a realization of an infinite sequence of i.i.d. uniform $(0,1)$ random numbers (although only a finite subset will be used in the simulation). The mapping g is also a functional of p input distributions, say $\{F_1, \ldots, F_p\}$. Thus, $g(\omega) = g(\omega|F_1, \ldots, F_p)$. We represent the simulation in this way to indicate that different choices are possible for the driving input processes, and we distinguish p distributions (that need not be univariate) because the typical simulation is driven by i.i.d. samples from one or more distinct distributions, and "input modeling" involves characterizing each of these distributions separately. For instance, in a queueing simulation F_1 might be the distribution of the interarrival times of a stationary arrival process, while F_2 could be the distribution of service times. Since the individual distributions need not be univariate, we will simplify the notation by using F to represent the joint distribution of all input random variables.

In addition, the simulation output often depends on other parameters describing the structure of the model or the experiment, such as the number of servers, the "feeds and speeds" of the equipment, or the simulation stopping time, but this dependence is omitted from the notation and not a part of our framework. We assume that the objective is to characterize system behavior conditional on a given set of these controllable parameters.

The output from a single realization (replication) of a simulation with input distribution F can be written as $Y(F,\omega) = \beta_0(F) + \varepsilon(F,\omega)$ where $\beta_0(F) = \int g(\omega|F)dP$ and $\varepsilon(F,\omega) = g(\omega|F) - \beta_0(F)$. From here on we drop the dependence of these random variables on ω for notational convenience, giving the random functional

$$Y(F) = \beta_0(F) + \varepsilon(F). \tag{1}$$

When independent replications are obtained (from independent realizations ω_j of ω) we append a subscript j to indicate the jth replication as in $Y_j(F) = \beta_0(F) + \varepsilon_j(F)$.

The objective of the simulation study is to characterize (estimate properties of) the distribution of $Y(F^c)$ where F^c is the true (correct) cdf of the real-world stochastic phenomena included in the model. The existence of F^c is, of course, a big assumption, but it is the basis of stochastic simulation and without it there is really no way to proceed.

Assuming the existence of F^c, we formulate the following output analysis problem of interest: obtain (random) bounds $C_L(F^c)$ and $C_U(F^c)$ such that

$$\Pr\{\beta_0(F^c) \in [C_L(F^c), C_U(F^c)]\} \geq 1 - \alpha. \tag{2}$$

Here the randomness in the bounds comes from the intrinsic randomness in the simulation output generated under F^c, that is, through $\varepsilon(F^c,\omega)$. Our reason for not directly addressing the problem of estimating $\beta_0(F^c)$ will become clear shortly.

For many simulation settings the output is itself an average of a large number of more basic outputs, so a version of the Central Limit Theorem implies that the distribution of $\varepsilon(F,\omega)$ (and consequently $Y(F)$) is approximately Gaussian, with mean and standard deviation that depend on F. Simulation outputs in this category include continuous-time-average statistics such as utilization, and discrete-time-average statistics such as average waiting time. If F^c were known then the desired CI could be obtained from classical statistics using sampled outputs $\{Y_j(F^c), j = 1,2,\ldots,n\}$ from a set of n replicated simulation runs.

Of course, F^c is typically unknown and must be estimated using finite samples. In general, the number of samples from each of the p input distributions might differ. However, to simplify notation, we will assume that we have m sample values from each input process, denoted collectively by $\mathbf{X}_m = \{X_1, X_2, \ldots, X_m\}$, that have been obtained from the real-world F^c. This gives the approximation $\widehat{F^c} = F^m(\cdot \mid \mathbf{X}_m)$, or more simply F^m, a (random) fitted distribution based on the random vector $\mathbf{X}_m$. Uniform convergence of empirical distributions in this setting means uniform convergence $F^m \overset{m\to\infty}{\longrightarrow} F^c$, if F^m is an empirical distribution (possibly smoothed). If F^m is a parametric distribution fitted to $\mathbf{X}_m$, no such guarantee exists, except for the case when F^c is a member of the chosen parametric family.

Simulation output analysis is usually based on a particular realization of $\mathbf{X}_m$, say $\mathbf{x}_m^0 = \{x_1^0, x_2^0, \ldots, x_m^0\}$, giving $\widehat{F^c} = F_0^m(\cdot \mid \mathbf{x}_m^0)$, or F_0^m for short. The realization $\mathbf{x}_m^0$ is the actual real-world data used to fit $\widehat{F^c}$. We drop the explicit dependence on $\mathbf{x}_m^0$ in the following and write the sample-based distribution as F_0^m. Abusing notation we will sometimes let F_0^m denote a fixed distribution based on a particular sample $\mathbf{x}_m^0$, but more often think of it as a random function of a random sample $\mathbf{X}_m^0$ from F^c. To describe our framework we will also consider repeated i.i.d. samples of size m of real-world data, $\mathbf{X}_m^0, \mathbf{X}_m^1, \mathbf{X}_m^2, \ldots$, and let F_i^m denote the fitted distribution obtained from the i^{th} sample.

The standard simulation output analysis constructs confidence intervals that are conditioned on F_0^m: obtain (random) bounds $C_L(F_0^m)$ and $C_U(F_0^m)$ such that

$$\Pr\{\beta_0(F_0^m) \in [C_L(F_0^m), C_U(F_0^m)]\} \geq 1 - \alpha. \tag{3}$$

The randomness in the conditional CI comes only from ω; the randomness in $F_0^m(\cdot \mid \mathbf{X}_m^0)$ is ignored. Thus, the bounds are relative to the distribution F_0^m rather than F^c. As has been shown in many papers, this can lead to coverage

probabilities for $\beta_0(F^c)$ that are far from $1-\alpha$. *Our goal is to obtain asymptotically correct confidence intervals for* $\beta_0(F^c)$ *under the assumptions that the real-world distribution* F^c *is stable and the model logic is correct.*

Our focus on a CI for $\beta_0(F^c)$ avoids the problem of correcting for the bias of $\beta_0(F_0^m)$ as an estimator of $\beta_0(F^c)$, a problem that is very difficult to address; instead we bound the value of $\beta_0(F^c)$ with high probability and let this interval account for the bias.

Consider the thought experiment in which additional real-world samples $\mathbf{X}_m^i$ are available, each resulting in a fitted distribution F_i^m. The distribution of the simulation output $Y(F^m)$ therefore has two sources of randomness: the simulation output variability conditional on F^m, and the randomness from the sampling variability of $\mathbf{X}_m$ used in constructing F^m. Specifically,

$$\begin{aligned} Y(F^m) &= \beta_0(F^m)+\varepsilon(F^m) \\ &= \beta_0(F^c)+[\beta_0(F^m)-\beta_0(F^c)]+[Y(F^m)-\beta_0(F^m)] \\ &= \beta_0(F^c)+D+V. \end{aligned} \tag{4}$$

We now make two heroic assumptions that allow us to obtain the CI (2): (a) The functional $\beta_0(\cdot)$ is known, and (b) the distribution of D, denoted F_D, is also known. Let $C_U=\beta_0(F_0^m)-F_D^{-1}(\alpha/2)$ and $C_L=\beta_0(F_0^m)-F_D^{-1}(1-\alpha/2)$. Then

$$\begin{aligned} \Pr\{C_L\le\beta_0(F^c)\le C_U\} &= \Pr\{F_D^{-1}(\alpha/2)\le\beta_0(F_0^m)-\beta_0(F^c)\le F_D^{-1}(1-\alpha/2)\} \\ &= 1-\alpha \end{aligned}$$

since $\beta_0(F_0^m)-\beta_0(F^c)$ has distribution F_D.

Neither $\beta_0(\cdot)$ nor F_D will be known, but these "heroic" assumptions show us one path to obtain an asymptotically valid CI:

- Build a metamodel $M(\cdot)$ for $\beta_0(\cdot)$ that converges, say as the number of design points and number of simulation replications go to infinity, to $\beta_0(\cdot)$.
- Given the metamodel M, use bootstrap samples from F_0^m, say $F_{0i}^m, i=1,2,\ldots,B$, to approximate the distribution F_D by the empirical distribution of $M(F_{0i}^m)-M(F_0^m)$. If the metamodel is exact, then as $B\to\infty$ and $m\to\infty$ this empirical distribution becomes F_D.

Metamodeling the mean simulation response as a function of the input distributions is key to our framework. By running a designed experiment in the F input distribution space, we leverage more information about the response surface $\beta_0(\cdot)$, and the impact of distribution choice on output performance, than we would have from a single simulation at $F=F_0^m$. This effectively eliminates the impact of V. And using a metamodel M also provides the continuity conditions that support the validity of bootstrapping. Of course, the challenge is designing the experiment in F space, and in the choice of metamodel. In our empirical study we tackle the easiest, but still non-trivial case in which the input models are members of a known parametric family.

3 ILLUSTRATION

We present empirical results from two small examples to illustrate the potential of our framework. We compare our metamodel-assisted bootstrap CI to the standard conditional CI that ignores input uncertainty; to the direct and Bayesian bootstrap CIs of Barton and Schruben (2001); and to a Bayesian credible interval that is similar to Zouaoui and Wilson (2004).

3.1 Examples

Both of the examples are Markovian queues. In this proof-of-concept illustration we assume that the correct parametric family of distributions (exponential) for the interarrival and service processes is known, but the true parameters (arrival rate λ_c and mean service time τ_c) must be estimated from real-world data. To focus on the input uncertainty problem and avoid complications due to initialization bias and autocorrelated output processes, each simulation replication generates a single observation from the steady-state queue-length distribution, which is known for these tractable examples. Let $G(\cdot|\lambda,\tau)$ be the relevant queue-length distribution for generic λ,τ.

Specifically, we consider the following two examples:

M/M/∞ Queue: The goal is to estimate the steady-state expected number of customers in an $M/M/\infty$ queue. For generic arrival rate λ and mean service time τ the expected number in queue is $\beta_0(\lambda,\tau)=\lambda\tau$, and the distribution G is Poisson.

M/M/1/k Queue: The goal is to estimate the steady-state expected number of customers in an $M/M/1/k$ queue. For generic arrival rate λ and mean service time τ the expected number in queue for $\lambda\tau \neq 1$ is

$$\beta_0(\lambda,\tau) = \frac{\lambda\tau}{1-\lambda\tau} - \frac{(k+1)(\lambda\tau)^{k+1}}{1-(\lambda\tau)^{k+1}} \tag{5}$$

and the distribution G is a truncated geometric.

These two examples have distinctly different output distributions (Poisson and truncated geometric), and their expected values, as a function of the inputs parameters, range from simple linear ($M/M/\infty$) to complex nonlinear ($M/M/1/k$). Notice that in both cases G exists for all non-negative λ and τ.

3.2 Procedures

All of the procedures start with a sample of real-world data. In our experiments $\mathbf{S}_m^0 = \{S_1^0, S_2^0, \ldots, S_m^0\}$ are i.i.d. exponential$(1/\tau_c)$ representing observed service times, and $\mathbf{A}_m^0 = \{A_1^0, A_2^0, \ldots, A_m^0\}$ are i.i.d. exponential(λ_c) representing observed interarrival times. Thus, the "real-world" data are $\mathbf{X}_m^0 = \{\mathbf{A}_m^0, \mathbf{S}_m^0\}$. The procedures differ in how they use these data to form a $(1-\alpha)100\%$ CI for $\beta_0(\lambda_c,\mu_c)$.

The conditional CI is what we obtain using sound statistical analysis, but ignoring input uncertainty.

Conditional CI

1. Set $\widehat{\tau}_0 = \bar{S}^0$ and $\widehat{\lambda}_0 = 1/\bar{A}^0$, the sample mean and one over the sample mean of $\mathbf{S}_m^0$ and $\mathbf{A}_m^0$, respectively.
2. Simulate n_c replications $Y_1, Y_2, \ldots, Y_{n_c}$ that are i.i.d. $G(\cdot|\widehat{\lambda}_0, \widehat{\tau}_0)$.
3. Report the CI $\left[\bar{Y} - z_{1-\alpha/2}S_Y/\sqrt{n_c},\ \bar{Y} + z_{1-\alpha/2}S_Y/\sqrt{n_c}\right]$, where $\bar{Y}$ and S_Y^2 and the sample mean and variance, respectively, and z_γ is the γ quantile of the standard normal distribution.

The direct and Bayesian bootstrap CIs, and the Bayesian credible interval, are obtained using a common structure that differs only in the resampling step. Therefore, we present all three of them in the following procedure. When we say "Generate bootstrap samples $\mathbf{S}_m^j \sim \mathbf{S}_m^0$" we mean generate a random sample of size m by sampling with replacement from the real-world data $\mathbf{S}_m^0$.

Direct Bootstrap/Bayesian Bootstrap/Bayesian CI

1. For $j = 1$ to B

   If direct bootstrap then: Generate bootstrap samples $\mathbf{S}_m^j \sim \mathbf{S}_m^0$ and $\mathbf{A}_m^j \sim \mathbf{A}_m^0$ and set $\widehat{\tau}_j = \bar{S}^j$ and $\widehat{\lambda}_j = 1/\bar{A}^j$.

   If Bayesian bootstrap then: Generate $E_1^S, E_2^S, \ldots, E_m^S \sim$ i.i.d. exponential(1), and let $W_i^S = E_i^S / \sum_{\ell=1}^m E_\ell^S$. Similarly, generate $E_1^A, E_2^A, \ldots, E_m^A \sim$ i.i.d. exponential(1), and let $W_i^A = E_i^A / \sum_{\ell=1}^m E_\ell^A$.

   Set $\widehat{\tau}_j = \sum_{\ell=1}^m W_\ell^S S_\ell^0$ and $1/\widehat{\lambda}_j = \sum_{\ell=1}^m W_\ell^A A_\ell^0$.

   Comment: Rather than resampling, the Bayesian bootstrap reweights the data using uniform spacings.

   If Bayesian then: Generate $1/\widehat{\tau}_j \sim \text{gamma}(m, \sum_{\ell=1}^m S_\ell^0)$ and $\widehat{\lambda}_j \sim \text{gamma}(m, \sum_{\ell=1}^m A_\ell^0)$.

   Comment: We sample from the posterior distributions of $1/\tau_c$ and λ_c obtained using the non-informative prior for the exponentially distributed real-world data; see Chick (2001).

   Always: Simulate n_b replications $Y_1, Y_2, \ldots, Y_{n_b}$ that are i.i.d. $G(\cdot|\widehat{\lambda}_j, \widehat{\tau}_j)$ and compute $\bar{Y}_j$, the sample average.

   Next j

2. Report the CI $\left[\bar{Y}_{(\lceil B\alpha/2\rceil)}, \bar{Y}_{(\lceil B(1-\alpha/2)\rceil)}\right]$, where $\bar{Y}_{(j)}$ denotes the jth smallest value (order statistic).

   Comment: In the actual implementation we use interpolation between order statistics to improve the estimates of the $\alpha/2$ and $1-\alpha/2$ quantiles.

Finally, we present our procedure for metamodel-assisted bootstrapping. Later we describe the specific metamodel type, experiment design and fitting method that we used in these experiments.

Metamodel-Assisted Bootstrap CI

1. Select experiment design $\{(\lambda_i, \tau_i), i = 1, 2, \ldots, d\}$.
2. For $i = 1, 2, \ldots, d$, do the following: generate $Y_j(\lambda_i, \tau_i) \sim$ i.i.d. $G(\cdot|\lambda_i, \tau_i)$ for $j = 1, 2 \ldots, n_m$, and average them to get $\bar{Y}(\lambda_i, \tau_i)$.

3. Fit a metamodel $M(\lambda,\tau)$ for $\beta_0(\lambda,\tau)$ using $\{\bar{Y}(\lambda_i,\tau_i),\lambda_i,\tau_i,i=1,2,\ldots,d\}$.
4. For $j=1$ to B
   (a) Generate bootstrap samples $\mathbf{S}_m^j \sim \mathbf{S}_m^0$ and $\mathbf{A}_m^j \sim \mathbf{A}_m^0$ and set $\widehat{\tau}_j = \bar{S}^j$ and $\widehat{\lambda}_j = 1/\bar{A}^j$.
   (b) Set $\widehat{\beta}_j = M(\widehat{\lambda}_j,\widehat{\tau}_j)$.

   Next j
5. Report CI $\left[\widehat{\beta}_{(\lceil B\alpha/2\rceil)},\widehat{\beta}_{(\lceil B(1-\alpha/2)\rceil)}\right]$.

Notice that metamodel-assisted bootstrapping lets M stand in for the simulation estimate of $\beta_0(\lambda,\tau)$ used by the other bootstrapping and Bayesian methods.

3.3 Stochastic Kriging

We use stochastic kriging from Ankenman, Nelson, and Staum (2010) to produce the metamodel M. Stochastic kriging is an extension of kriging for deterministic computer experiments to stochastic simulation. In our context, stochastic kriging treats the unknown response $\beta_0(\lambda,\tau)$ as a realization of a Gaussian random field that exhibits spatial correlation, meaning that the values $\beta_0(\lambda,\tau)$ and $\beta_0(\lambda',\tau')$ will tend to be similar when (λ,τ) and (λ',τ') are "close" in some spatial sense. In our experiments we used the well-known Gaussian correlation function $\exp\{-\theta_1(\lambda-\lambda')^2-\theta_2(\tau-\tau')^2\}$ where θ_1 and θ_2 are estimated as part of the fitting process. See Ankenman, Nelson, and Staum (2010) for complete details.

Stochastic kriging is particularly well suited to metamodel-assisted bootstrapping because it does not make strong assumptions about the form of the response surface, but rather predicts the surface as a weighted average of observed data that gives more weight to observations that are closer to the prediction point. When we use parametric input models, we typically expect that input models with nearly the same parameters will produce similar simulation results.

3.4 Testing Protocol

We evaluate the procedures by estimating the coverages of their CIs. This is not entirely fair to the Bayesian procedure which actually produces a credible interval that is intended to provide a posterior assessment of the user's remaining uncertainty about β_0 rather than cover the true value with a given confidence. Nevertheless, the coverage provides a common measure of performance.

We also want to compare the procedures given approximately equal computational effort. We take the position that in practical simulations the computational effort to produce simulation output is significantly larger than that required to generate bootstrap or posterior samples (at least when conjugate or non-informative priors are used for the input parameters) or to fit or evaluate a metamodel. Therefore, in our experiments we set $n_c = Bn_b = dn_m$, so that all five methods generate the same number of simulation replications. This means that whatever number of replications we would spend on the standard simulation experiment that ignores input uncertainty, we spread evenly among the B bootstrap or posterior resamples for the direct and Bayesian bootstrap, and the Bayesian method; and spread them evenly among the d design points for our metamodel-assisted bootstrap. This slightly favors our method because we require fitting a metamodel, generating bootstrap samples, and evaluating the metamodel for each bootstrap sample, which does take some time. The direct and Bayesian bootstrap, and the Bayesian method, also require resampling.

To estimate coverage we make macroreplications. On each macroreplication, we generate an independent sample of real-world data $\mathbf{S}_m^0 = \{S_1^0,S_2^0,\ldots,S_m^0\}$ and $\mathbf{A}_m^0 = \{A_1^0,A_2^0,\ldots,A_m^0\}$, and apply all five procedures to the same sample. Since β_0 is known in our examples, we estimate coverage by the fraction of macroreplications on which each procedure's CI contained it.

Specifically, we consider the M/M/∞ queue with $\lambda_c = 1,\tau_c = 5$, implying $\beta_0(\lambda_c,\tau_c) = 5$, and the M/M/1/$k$ queue with $\lambda_c = 0.25,\tau_c = 5,k = 20$, implying $\beta_0(\lambda_c,\tau_c) = 16.195$. We make 1000 macroreplications, each producing a 95% CI. Thus, a confidence interval on the coverage estimates is approximately ± 0.014. The factors we vary include the total replication budget for each procedure, n_c, and the number of resamples B.

3.5 Results

Tables 1 and 2 summarize the results. Not surprisingly, the conditional confidence interval provides extremely poor coverage when there are only $m = 100$ observations of interarrival and services times from which to estimate the arrival rate and mean service time. Metamodel-assisted bootstrapping attains the nominal coverage in all cases considered. The direct and Bayesian bootstrap, and the Bayesian method, tend to be quite conservative when there are few simulation replications per bootstrap or posterior resample (each resample receives n_c/B replications per resample). Metamodel-assisted bootstrap invests n_c/d replications per design point, and for these examples fewer design points are needed to develop well-fitting metamodels than bootstrap/posterior resamples ($d = 20$ design points vs. $B = 100,1000$ resamples).

Table 1: Coverage results for the $M/M/\infty$ experiment based on 1000 macroreplications.

					Estimated Coverage					
n_c	d	B	m	$1-\alpha$	Conditional	Direct Boot	Bayes Boot	Bayesian	M-A Boot	R
4000	20	100	100	0.95	0.080	0.972	0.952	0.957	0.944	0.12
4000	20	1000	100	0.95	0.084	0.998	0.998	0.998	0.949	0.40
2000	20	100	100	0.95	0.119	0.976	0.976	0.978	0.943	0.18
2000	20	1000	100	0.95	0.120	1.000	1.000	1.000	0.956	0.57

Table 2: Coverage results for the $M/M/1/k$ experiment based on 1000 macroreplications.

					Estimated Coverage				
n_c	d	B	m	$1-\alpha$	Conditional	Direct Boot	Bayes Boot	Bayesian	M-A Boot
4000	20	100	100	0.95	0.045	0.963	0.943	0.939	0.946
4000	20	1000	100	0.95	0.056	0.987	0.986	0.990	0.954
2000	20	100	100	0.95	0.069	0.960	0.945	0.905	0.952
2000	20	1000	100	0.95	0.062	1.000	0.999	1.000	0.947

In a limited number of experiments with only $m = 50$ real-world samples (not reported here), metamodel-assisted bootstrapping still achieved the nominal coverage.

The over coverage of the bootstrapping approaches was described and explained in Barton et al. (2002) and Barton (2007): formal validity of the bootstrap depends on the statistic of interest being a well-behaved (and in particular, deterministic) function of the data. However, in stochastic simulation—without the metamodel assist—β_0 is estimated with noise on each bootstrap resample. This additional variability is reflected in the bootstrap distribution, and if too large causes significant over coverage. In our results, when we use $B = 1000$ bootstrap samples, there are only a small number of simulation replications per sample, making them noisy estimates of β_0.

Barton (2007) proposed looking at a measure of stochastic noise relative to confidence-interval width as an indicator of when direct bootstrap confidence intervals will be overly conservative. In particular, he showed empirically in an example like our $M/M/1/k$ queue that if this ratio is greater than about 0.15 then over coverage occurs. For the $M/M/\infty$ queue, using the direct bootstrap and Bayesian methods, we can approximate this measure in closed form.

For the direct bootstrap and Bayesian methods, let $\bar{Y} = B^{-1}\sum_{j=1}^{B}\bar{Y}_j$, the overall average of the bootstrap and posterior sample estimators, respectively. Then we can show that

$$\mathrm{Var}(\bar{Y}) \approx \frac{\lambda_c\tau_c}{Bn_b} + \frac{2}{m}\left(\frac{B+1}{B}\right)(\lambda_c\tau_c)^2.$$

This expression shows the impact of simulation replications per bootstrap or posterior resample, n_b, number of resamples B, and quantity of real-world data m. Recall that in our examples $Bn_b = n_c$. If we approximate the width of the 95% confidence intervals by $2 \times 1.96\sqrt{\mathrm{Var}(\bar{Y})}$, then Barton's measure is

$$R = \frac{\sqrt{\lambda_c\tau_c/n_b}}{2 \times 1.96\sqrt{\mathrm{Var}(\bar{Y})}}. \tag{6}$$

The last column of Table 1 displays the R ratios for the $M/M/\infty$ example. Notice the substantial over coverage of the direct and Bayesian bootstrap, and the Bayesian method, when $R = 0.40$ and 0.57, and slight over coverage when $R = 0.18$. This provides further evidence that Barton's ratio is a useful indicator. We do not expect metamodel-assisted bootstrapping to need such a check, and that expectation is supported by our simple test cases.

4 WHY METAMODEL-ASSISTED BOOTSTRAPPING WORKS

In this section we provide some support for the metamodel-assisted bootstrapping approach, focusing on the examples in this paper and indicating what needs to be done to establish its validity more generally.

Suppose that, in our two examples, the metamodel M is exact, which means that $M(\lambda,\tau)=\beta_0(\lambda,\tau)$. Then we will prove that

$$\lim_{m,B\to\infty}\Pr\left\{\beta_0(\lambda_c,\tau_c)\in\left[\widehat{\beta}_{(\lceil B\alpha/2\rceil)},\widehat{\beta}_{(\lceil B(1-\alpha/2)\rceil)}\right]\right\}=1-\alpha.$$

In other words, the confidence interval is asymptotically consistent.

In both examples we assume that $0<\lambda_c,\tau_c<\infty$, and for the $M/M/1/k$ example we also assume that $\lambda_c\tau_c\neq 1$. We drop the explicit dependence on $B\to\infty$ from here on since resampling it is just a tool to evaluate the bootstrap distribution and is not fundamental to the following argument.

1. First redefine the mean response to be a function of $\eta=1/\lambda$ rather than λ; that is, the mean steady-state queue length is $\beta_0(\eta,\tau)$. Notice that β_0 is continuously differentiable and has non-zero gradient in a neighborhood of (η_c,τ_c) for both the $M/M/\infty$ and $M/M/1/k$ queues.
2. As a consequence of Step 1, the distribution of $\sqrt{m}\left[\beta_0(\bar{A}^j,\bar{S}^j)-\beta_0(\bar{A}^0,\bar{S}^0)\right]$ is strongly consistent for the distribution of $\sqrt{m}\left[\beta_0(\bar{A}^0,\bar{S}^0)-\beta_0(\eta_c,\tau_c)\right]$ as $m\to\infty$ by Theorem 3.1 of Shao and Tu (1995).
3. As a further consequence of Step 1, $\sqrt{m}\left[\beta_0(\bar{A}^0,\bar{S}^0)-\beta_0(\eta_c,\tau_c)\right]$ is asymptotically normal with mean 0 as $m\to\infty$ using standard delta-method arguments.
4. Steps 2 and 3 establish that the percentile interval $\left[\widehat{\beta}_{(\lceil B\alpha/2\rceil)},\widehat{\beta}_{(\lceil B(1-\alpha/2)\rceil)}\right]$ satisfies the conditions of Theorem 4.1 of Shao and Tu (1995), and is therefore asymptotically consistent as $m\to\infty$.

This proves that the metamodel-assisted bootstrap does exactly as one would hope as more and more real-world data are obtained. A key to this proof is assuming that the metamodel M is exact. Clearly a complete proof also needs to establish convergence of M to β_0 as more and more simulation effort is expended on building the metamodel. If M were of the correct parametric form, but with unknown parameters, then the availability of consistent parameter estimators would be sufficient. For the $M/M/\infty$ example this would mean knowing that $\beta_0(\lambda,\tau)=\beta\lambda\tau$, but not knowing the value of β. Nevertheless, we prefer interpolation approaches like stochastic kriging because we believe that practical situations in which the true parametric form of the model is known are rare. However, this belief makes the asymptotic analysis more difficult.

5 CONCLUSIONS

We have presented a method to capture "input uncertainty" that appears to be more robust than previous proposals. The expected over coverage by the bootstrapping and Bayesian approaches was supported by our two test cases, and the severity of the over coverage was related to Barton's R for the $M/M/\infty$ example.

At the core of our approach is a metamodel that links the expected value of the simulation response to the input distributions that drive the simulation model; we proved that this approach is asymptotically valid under the assumption that the metamodel is correct: $M=\beta_0$. It is important to note that having an exact metamodel is a stronger condition than what is really needed for practical usefulness. The validity of the CI depends on approximating the distribution of $\beta_0(F^m)-\beta_0(F^c)$ by the bootstrap distribution of $M(F_{0i}^m)-M(F_0^m)$. This distribution will be a good approximation if M is such that the distribution of the deviation of $M(F_{0i}^m)$ from $M(F_0^m)$ is similar to the distribution of the deviation of $\beta_0(F^m)$ from $\beta_0(F^c)$. This is more likely to be approximately true than that $M=\beta_0$, and seems achievable with sound experimental design for metamodel estimation.

Our proof also made use of the well-developed theory of bootstrapping for estimators that are functions of sample means. This approach is easily extended to metamodels that are functions of sample moments beyond just the first, but that still implies assuming that the correct parametric input-model families are known. Zouaoui and Wilson (2004) accounted for lack of knowledge of the correct model family by using Bayesian model averaging that employs a collection of plausible input-model families that are weighted by their posterior likelihood of being correct. An approach that is more consistent with our framework is to formulate metamodels that are functionals of the entire input distribution, and drive the simulation with the empirical cdfs rather than any fitted parametric input distributions.

A challenge for the metamodel-assisted bootstrap method will be efficient fitting of metamodels when the number of distributions or distribution parameters grows. As this number grows, the number of design points needed in the experiment can be expected to grow in a nonlinear fashion. Many distributions will be needed for complex simulation models. This is a topic of current research.

ACKNOWLEDGMENTS

This paper is based upon work supported by the National Science Foundation under Grant No. CMMI-0900354.

REFERENCES

Ankenman, B., B. L. Nelson, and J. Staum. 2010. Stochastic kriging for simulation metamodeling. *Operations Research* 58 (2): 371–382.

Barton, R. R. 2007. Presenting a more complete characterization of uncertainty: Can it be done? In *Proceedings of the 2007 INFORMS Simulation Society Research Workshop*. Fontainebleau: INFORMS Simulation Society.

Barton, R. R., R. C. H. Cheng, S. E. Chick, S. G. Henderson, A. M. Law, L. M. Leemis, B. W. Schmeiser, L. W. Schruben, and J. R. Wilson. 2002. Panel on current issues in simulation input modeling. In *Proceedings of the 2002 Winter Simulation Conference*, ed. E. Yücesan, C. Chen, J. L. Snowdon, and J. M. Charnes, 353–369: Piscataway, N.J.: Institute of Electronic and Electrical Engineers.

Barton, R. R., and L. W. Schruben. 1993. Uniform and bootstrap resampling of input distributions. In *Proceedings of the 1993 Winter Simulation Conference*, ed. G. W. Evans, M. Mollaghasemi, W. E. Biles, and E. C. Russell, 503–508: Piscataway, N.J.: Institute of Electrical and Electronics Engineers.

Barton, R. R., and L. W. Schruben. 2001. Resampling methods for input modeling. In *Proceedings of the 2001 Winter Simulation Conference*, ed. B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, 372–378: Piscataway, N.J.: Institute of Electronic and Electrical Engineers.

Cheng, R. C. H. 1994. Selecting input models. In *Proceedings of the 1994 Winter Simulation Conference*, ed. J. D. Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila, 184–191: Piscataway, N.J.: Institute of Electronic and Electrical Engineers.

Cheng, R. C. H., and W. Holland. 1997. Sensitivity of computer simulation experiments to errors in input data. *Journal of Statistical Computation and Simulation* 57:219–241.

Cheng, R. C. H., and W. Holland. 1998. Two-point methods for assessing variability in simulation output. *Journal of Statistical Computation and Simulation* 60:183–205.

Cheng, R. C. H., and W. Holland. 2004. Calculation of confidence intervals for simulation output. *ACM Transactions on Modeling and Computer Simulation* 14 (4): 344–362.

Chick, S. E. 1997. Bayesian analysis for simulation input and output. In *Proceedings of the 1997 Winter Simulation Conference*, ed. S. Andradottir, K. J. Healy, D. H. Withers, and B. L. Nelson, 253–260: Piscataway, N.J.: Institute of Electronic and Electrical Engineers.

Chick, S. E. 1999. Steps to implement Bayesian input distribution selection. In *Proceedings of the 1999 Winter Simulation Conference*, ed. P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, 317–324: Piscataway, N.J.: Institute of Electronic and Electrical Engineers.

Chick, S. E. 2000. Bayesian methods for simulation. In *Proceedings of the 2000 Winter Simulation Conference*, ed. J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, 109–118: Piscataway, N.J.: Institute of Electronic and Electrical Engineers.

Chick, S. E. 2001. Input distribution selection for simulation experiments: Accounting for input uncertainty. *Operations Research* 49 (5): 744–758.

Chick, S. E., and S. H. Ng. 2002. Joint criterion for factor identification and parameter estimation. In *Proceedings of the 2002 Winter Simulation Conference*, ed. E. Yücesan, C. Chen, J. L. Snowdon, and J. M. Charnes, 400–406: Piscataway, N.J.: Institute of Electrical and Electronics Engineers.

Freimer, M., and L. W. Schruben. 2002. Collecting data and estimating parameters for input distributions. In *Proceedings of the 2002 Winter Simulation Conference*, ed. E. Yücesan, C. Chen, J. L. Snowdon, and J. M. Charnes, 392–399: Piscataway, N.J.: Institute of Electrical and Electronics Engineers.

Henderson, S. 2003. Input model uncertainty: Why do we care and what should we do about it? In *Proceedings of the 2003 Winter Simulation Conference*, ed. S. E. Chick, P. J. Sanchez, D. Ferrin, and D. J. Morrice, 90–100: Piscataway, N.J.: Institute of Electronic and Electrical Engineers.

Kleijnen, J. P. C. 1987. *Statistical tools for simulation practitioners*. Marcel Dekker Inc.

Kleijnen, J. P. C. 1994. Sensitivity analysis versus uncertainty analysis: When to use what? In *Predictability and Nonlinear Modelling in Natural Sciences and Economics*, ed. ed. J. Grasman and G. van Straten, 322–333. Dordrecht: Kluwer.

Law, A. M., and W. Kelton. 2000. *Simulation modelling and analysis*. 3rd ed. New York: McGraw Hill.

Lee, S., and P. W. Glynn. 2003. Computing the distribution function of a conditional expectation via Monte Carlo: Discrete conditioning spaces. *ACM Transactions on Modeling and Computer Simulation* 13 (3): 238–258.

Ng, S. H., and S. E. Chick. 2001. Reducing input parameter uncertainty for simulations. In *Proceedings of the 2001 Winter Simulation Conference*, ed. B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, 364–371: Piscataway, N.J.: Institute of Electrical and Electronics Engineers.

Shao, J., and D. Tu. 1995. *The jackknife and bootstrap*. New York: Springer.

Steckley, S., and S. Henderson. 2003. A kernel approach to estimating the density of a conditional expectation. In *Proceedings of the 2003 Winter Simulation Conference*, ed. S. E. Chick, P. J. Sanchez, D. Ferrin, and D. J. Morrice, 383–391: Piscataway, N.J.: Institute of Electrical and Electronics Engineers, Inc.

Zouaoui, F., and J. R. Wilson. 2001a. Accounting for input model and parameter uncertainty in simulation. In *Proceedings of the 2001 Winter Simulation Conference*, ed. B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, 290–299: Piscataway, N.J.: Institute of Electrical and Electronics Engineers, Inc.

Zouaoui, F., and J. R. Wilson. 2001b. Accounting for parameter uncertainty in simulation input modeling. In *Proceedings of the 2001 Winter Simulation Conference*, ed. B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, 354–363: Piscataway, N.J.: Institute of Electronic and Electrical Engineers.

Zouaoui, F., and J. R. Wilson. 2003. Accounting for parameter uncertainty in simulation input modeling. *IIE Transactions* 35:781–792.

Zouaoui, F., and J. R. Wilson. 2004. Accounting for input-model and input-parameter uncertainties in simulation. *IIE Transactions* 36:1135–1151.

AUTHOR BIOGRAPHIES

RUSSELL R. BARTON is a professor in the Department of Supply Chain and Information Systems at the Pennsylvania State University. He is co-director of the Penn State Master of Manufacturing Management degree program and associate director of the Center for the Management of Technological and Organizational Change. He received a B.S. degree in Electrical Engineering from Princeton University and M.S. and Ph.D. degrees in Operations Research from Cornell University. Before entering academia, he spent twelve years in industry. He is a past president of the INFORMS Simulation Society and serves on the advisory board for the INFORMS Quality Statistics and Reliability section. He is a senior member of IIE and IEEE. His research interests include applications of statistical and simulation methods to system design and to product design, manufacturing and delivery. His email address is ⟨rbarton@psu.edu⟩.

BARRY L. NELSON is the Charles Deering McCormick Professor and Chair of the Department of Industrial Engineering and Management Sciences at Northwestern University, and a Fellow of INFORMS. His research centers on the design and analysis of computer simulation experiments on models of stochastic systems. His e-mail and web addresses are ⟨nelsonb@northwestern.edu⟩ and ⟨www.iems.northwestern.edu/~nelsonb/⟩.

WEI XIE is a Ph.D. student in the Department of Industrial Engineering and Management Sciences at Northwestern University. Her research interests are in simulation metamodeling and input uncertainty. Her e-mail address is ⟨WeiXie2013@u.northwestern.edu⟩.