Proceedings of the 2012 Winter Simulation Conference
C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A.M. Uhrmacher, eds

TUTORIAL: INPUT UNCERTAINTY IN OUTOUT ANALYSIS

Russell R. Barton

The Pennsylvania State University
University Park, PA 16803, USA

ABSTRACT

Simulation output clearly depends on the form of the input distributions used to drive the model. Often these input distributions are fitted using finite samples of real-world data. The finiteness of the samples introduces errors in the input distributions, affecting the output. Yet this propagation of input model uncertainty to output uncertainty is rarely considered in simulation output analysis. This tutorial presents a discussion of input uncertainty issues and recently developed methodological approaches, set in the context of input uncertainty methods proposed over the past twenty years.

1 INTRODUCTION

Discrete event simulation is a powerful tool for gaining insight on the operational behavior of real systems, from call centers to manufacturing lines, from hospitals to ports. The great advantage of simulation is the ability to analyze complex systems using models constructed to capture critical aspects of system behavior with high fidelity.

How is fidelity characterized? Validation of the simulation against actual system behavior is often too costly or practically impossible. Given sufficient resources to conduct a validation, the usual approach would result in failure, as will be seen in an example in a later section.

Generally, when we characterize departures in the behavior of a simulation model from the real system, they fall in two classes: fidelity loss coming from the use of incorrect 'input models' to drive the simulation, and fidelity loss coming from failure of the execution logic of the simulation to match the logic (or lack thereof) of the real-world system. Input models are the probability distributions, univariate or multivariate, used to drive the simulation. These models provide a way to generate random instances of key model elements: entity interarrival times; service times for different resources; breakdown times for resources; routing probabilities for entities; and other entity characteristics such as weight, value, or quantity. Real-world data is important to achieve fidelity in both classes. Sargent (2011) notes: "To build a conceptual model we must have sufficient data on the problem entity to develop theories that can be used to build the model, to develop mathematical and logical relationships for use in the model that will allow the model to adequately represent the problem entity for its intended purpose, and to test the model's underlying assumptions." This advanced tutorial explains ways to characterize the impact on simulation output arising from input model errors. Sargent further describes the need for "data that are appropriate, accurate, and sufficient." But sufficiency connotes a yes/no value relative to the intended purpose of the simulation. It is possible to provide more quantitative characterizations of this source of fidelity loss, which is the value of this tutorial.

Unfortunately, as is often the case with academic pursuits, the research results summarized in this tutorial address a narrower problem. Only errors arising from the finiteness of samples of real-world data are considered. That means we will assume that either we know the correct probability distribution family (or small set of families) for the input probability model, or we use empirical distributions to drive the

simulation. For example, we will be able to say something about the error in output introduced from using a finite set of data to fit the arrival rate λ for an interarrival time distribution that is known to be exponential.

While this may seem a narrow pursuit, we will see that in very reasonable scenarios simulation output analysis can give characterizations quite different from the true system when this error is not considered. This makes the understanding of input model errors due to finite samples a worthwhile study, in spite of Schruben's "five dastardly Ds" of data (Barton et al. 2002). In the sections below this tutorial will explain the nature and seriousness of the problem, then summarize early approaches, now twenty years old, to characterize this source of error. The tutorial then will present recent work in this area, relate some recent research in discrete-event simulation to efforts for deterministic models, and in the final section identify the difficulties that remain with current approaches.

2 UNDERSTANDING INPUT MODEL UNCERTAINTY

2.1 Input Models

Discrete event simulations model stochastic behavior by specifying probabilities or probability distributions for random variables used in the model. These random variables affect the dynamic behavior of the model. They might include interarrival times, service times, times between machine maintenance or failures, travel times, route choice, defect probability, and so forth. Often these probabilities or distributions are estimated from samples of real-world data. The methods that are reviewed in this tutorial can be used for any such estimated probabilities, but for simplicity the focus will be on the probability models used to describe queues.

2.2 An Example of Input Model Uncertainty

Consider a simulation model of a capacitated queue where (surprisingly) we know that both the interarrival times and the service times follow exponential distributions. This $M/M/1/k$ setting allows an easy illustration of the impact of input model error caused by fitting models to finite sets of real-world data. Suppose that the true system has capacity $k = 20$, with interarrival times that are exponentially distributed with mean $1/\lambda = 1.25$ minutes, and that service times are exponentially distributed with mean $\tau = 1$ minute. In this case the true steady-state mean number of customers in the system can be computed using

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

and substituting $\tau = 1$ and $\lambda = .8$ to give $L \approx 3.805$. In this illustrative example, assume that (1) is not known, and that the objective of the simulation model is to estimate L and give an associated confidence interval. Here we let Y_j be the sample average of L that is output by the simulation for the j^{th} replication, and assume that we have m such output replications. Then we can write

$$L(\lambda,\tau) \approx L(\lambda',\tau') + \varepsilon = \frac{1}{m}\sum_{j=1}^{m} Y_j(\lambda',\tau') = \overline{Y} \tag{2}$$

where j indexes the simulation replication. The simulation uses estimates λ' and τ', which can be thought of as observations of the random variables Λ and T, estimators of λ and τ.

Any simulation run will be of finite length, and the average number of customers in the system over that finite run (or set of runs) will vary randomly from the quantity prescribed by (1) because of the finiteness of the simulation effort. Call this *intrinsic error*, which is represented by ε in (2). Since the simu-

lation is run using λ' and τ', the resulting estimate for L will not be 3.805, generally, even if intrinsic error is near zero. Call this second source of error *extrinsic error*. It is $L(\lambda', \tau') - L(\lambda, \tau)$.

How large might extrinsic error be for this example? Assume for the moment that the simulation is run long enough so that the intrinsic error is near zero; that is, the average number of customers in the system over the duration of the simulation (or set of simulation replications) is well-approximated by (1). Further assume (to simplify notation) that we have the same sample size for each real-world data set used to estimate each input parameter. Specifically in this case, suppose that we estimate $1/\lambda$ using $n = 500$ samples of real-world interarrival times, and τ using $n = 500$ samples of service times. Because these sample sizes are not infinite, the estimates $1/\lambda'$ and τ' will have some random error. Consider $r = 1000$ repeated experiments consisting of collecting 500 interarrival and service times, estimating values λ_i' and τ_i', $i = 1, 2, \ldots, 1000$, and running the simulation long enough to yield a value given by (1) for each set of arguments λ_i' and τ_i'. Figure 1 gives the histogram of $\overline{Y}_i$, each an estimate of L, over 1000 such experiments. The range of L estimates is quite large. No matter how much simulation effort is put into the runs to improve the fidelity of the estimate to (1), this error about the true value 3.805 will remain.

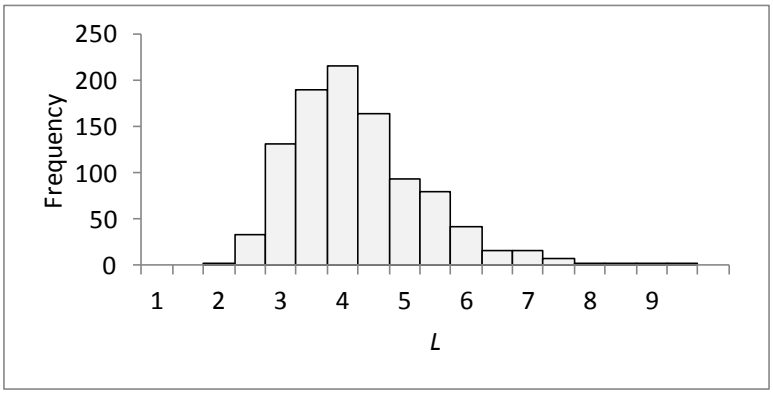

Figure 1: Variation in simulation output for L for an $M/M/1/20$ queue over a set of 1000 experiments, using different samples of size $n = 500$ (to estimate mean interarrival time and mean service time) for each experiment.

Shifting the L axis so that 3.805 becomes zero would yield a histogram approximately centered about zero, and would give a good approximation to the probability distribution of extrinsic error for the case of finite samples of size 500 for this model (where the input distribution families are assumed known). Of course, in practice only one of these experiments would be conducted, so one would not know the distribution of extrinsic error. It seems clear that this error should be included in the characterization of simulation output uncertainty, at least for this case. There is a greater than 25% chance that the one experiment of these 1000 that would be conducted would produce an estimate for average number in the system with error greater than 25%, – either less than 2.8 or greater than 4.8. The problem is that the usual confidence intervals for simulation only capture the error in the simulation output's approximation of (1) conditioned on the fitted values of λ' and τ', but do not capture the uncertainty in output from the errors in the input distributions, in this case arising from $\lambda' \neq \lambda$ and $\tau' \neq \tau$.

2.3 A Problem of General Concern but Little Commercial Action

The example above illustrates the significant risk associated with ignoring input model uncertainty, even when relatively *large* input sample sizes are used to fit parameters for *known* distribution families. The situation is not unique to this example, and has been illustrated in different settings by many researchers

over the past twenty years. Yet simulation software vendors have yet to implement analysis routines that take this error into account. As you will see in the sections below, methods that take this error into account can be complicated, computationally intensive, and overly conservative. So, at least until very recently, vendors had reason to be reluctant to address this issue.

3 THE FIRST TEN YEARS

3.1 Basic Approaches

In discussions during the 1992 Winter Simulation Conference, Lee Schruben introduced the concept of input model uncertainty, and proposed a method for capturing it. The idea was closely related to bootstrap resampling, and was further developed in Barton and Schruben (1993). There was a great deal of work on this topic by many researchers over the next ten years. A panel on input modeling at the 2002 Winter Simulation Conference provides a nice summary of many of basic approaches (Barton et al 2002). The strategies fall into four general categories: direct resampling, bootstrap resampling, methods based on Bayesian Model Average concepts, and approximations based on Taylor's theorem.

3.2 Direct Resampling

One might imagine conducting a process like the one used to construct Figure 1. Barton and Schruben (1993) called this *true resampling* but they used the interarrival and service times that were sampled (from exponential distributions) directly to form empirical distributions. The experiment would consist of r outer replications of m inner simulation replications each, with each outer replication using different fitted values from different real-world samples, each of size n. This paper did not compute confidence intervals for output parameters, but did suggest using a mixed effects Analysis of Variance to determine whether extrinsic variance was important (this implies $m > 1$ for the inner replications). This check was further developed in Freimer and Schruben (2002).

The usual t-based confidence intervals are not appropriate in this setting. First, the extrinsic variability will not generally follow a normal distribution, as can be seen from the skewness in Figure 1 (where there is no intrinsic variance). This skewness is more apparent in Figure 2 of Barton and Schruben (2001) for a capacitated queue with $k = 10$ and $n = 10$. Further, since we are trying to characterize the level of uncertainty in output coming from finite real-world samples of size n, t-based confidence intervals (though robust to some non-normality) are inappropriate, since the uncertainty will decrease to zero as r increases, even for fixed n. If intrinsic error is near zero, confidence intervals for L could be based on percentiles of the $\overline{Y}_i$ output values ($i = 1, 2, \ldots, r$).

Confidence intervals based on direct resampling are problematic for three reasons. First, the impact of intrinsic error on the estimate of L reduces as $1/\sqrt{rm}$ (not just $1/\sqrt{m}$), so confidence intervals based on percentiles of the r output values can significantly over-cover if there is significant intrinsic error. Second, the analysis requires rm simulation runs, compared with m in an analysis that ignores input model uncertainty. Third, it is very expensive and wasteful of real-world data to characterize uncertainty about L due to the finiteness (at level n) of real-world data. It is wasteful because we use rn units of data to characterize an estimate based on n units of data. Each outer replication uses only $1/r$ of the real-world data available to the simulationist. More precise results could be obtained by using all rn values to fit the input models (with reduced error because the finiteness of real-world data would be at level rn).

Methods for finding the distribution of a conditional expectation (Glynn 1986) fall in this direct resampling category. The connection was recognized by Steckley and Henderson (2003), who proposed a kernel method to estimate the conditional density. They built on the work by Lee and Glynn (1999, 2003) who proposed a method to estimate the distribution function (using the usual percentile estimator – even though Lee and Glynn later decompose extrinsic and intrinsic error to find optimal allocation of data collection vs. simulation effort). Examples typically assume that new samples can be generated via computa-

tion, and that resampling vs. simulation costs are known. A common objective is to find values r and m that minimize the MSE of the conditional expectation, subject to a fixed computational budget $c_1 r + c_2 rm = C$. The work on distribution of conditional expectations has focused on the asymptotic behavior of the estimates under conditions that have both r and m grow to infinity. The first problem associated with direct resampling does not apply in the limit, but would apply in practice. In particular, if the simulation effort were small, resulting in significant intrinsic error, the resulting empirical cumulative distribution functions or kernel-smoothed densities would have inflated variance (and generate overly conservative confidence intervals). The latter two problems remain.

3.3 (Direct) Bootstrap Resampling

Direct resampling (and related conditional expectation) methods enable characterization of error due to input models fitted with finite real-world data, subject to the three problems mentioned above. Bootstrapping is a statistical method that was developed specifically for data reuse to estimate the distribution of a statistic of interest (Efron and Tibshirani 1986). The direct bootstrap resampling method is the same as for direct resampling, except that the r resamples of size n are obtained from the (single) real-world sample by resampling from it with replacement (r is often called B in the bootstrap setting). This means that in a bootstrap resample some real-world values will be sampled twice, three times, etc., while others will not be included. The bootstrap resampled input data can be used directly if the simulation is driven by empirical distributions as in Barton and Schruben (1993, 2001) or it can be used to estimate input model parameters such as λ and τ for our example (e.g., Cheng 1995). Note that the implementations in Barton and Schruben (1993, 2001) used $m = 1$.

Instead of bootstrap resampling the original data, the empirical weights (normally $1/n$) on each original data sample can be changed randomly, as Lee Schruben suggested in 1992, using the result that $F_X(X) \sim U(0,1)$. The ordinates of the empirical cumulative distribution function at each of the original samples can be resampled using n (sorted) values from the $U(0,1)$ distribution. This method was called uniform resampling in Barton and Schruben (1993), and is an implementation of the Bayesian bootstrap method of Rubin (1981). Uniform resampling is particularly appropriate if the resampled empirical distribution will be used directly in the simulation, but the uniform resampled empirical distribution can also be used to compute distribution parameter estimates.

While bootstrapping uses data economically, both of the other problems associated with direct resampling remain: the inflation of the intervals by intrinsic variance, and the need to conduct rm simulation runs. The inflation of intervals by intrinsic variance did not have a noticeable impact on coverage probability in Barton and Schruben (2001). Barton et al. (2002) showed that small simulation effort producing large intrinsic variability could result in overcoverage for the bootstrap confidence interval. Barton (2007) showed that with $m > 1$ one could compare the standard deviation of within-resample replications to the width of the percentile-based confidence interval to determine whether intrinsic variance would be likely to cause significant overcoverage.

There is an additional theoretical shortcoming in the application of the bootstrap when the computed statistic is the output of simulation. The asymptotic correctness of coverage for the bootstrap requires that the computed statistic be a smooth function of the (resampled input) data. Since the simulation output can be thought of as stochastic, this condition is violated. Again, if intrinsic variance is small, the impact on coverage is not significant. Interestingly, this condition is violated whenever the bootstrap method is carried out using a digital computer, but this has not seemed to bother statisticians.

3.4 Bayesian Model Average (BMA) Approaches

Chick (1997, 1999, 2000, 2001) employed a Bayesian Model Average strategy to characterize uncertainty in simulation output based on uncertainty about both distribution family as well as distribution parameter values. For an overview of Bayesian Model Averaging, see Hoeting et al. (1999) and the BMA references therein, dating back to 1978. See also Cheng (1998). The BMA approach assumes that the real

world data is generated by one of a known set of parametric distribution families: which family may not be known, and the precise values of the parameters identifying a specific member of the family are assumed unknown. In BMA extrinsic error is referred to as *structural uncertainty* (consisting of *model uncertainty* and *parameter uncertainty*) and intrinsic error is *stochastic uncertainty*.

The BMA approach randomly samples input distributions and parameters before each simulation replication (much like parametric bootstrapping), using a Bayesian posterior distribution for input distributions and parameters, given historical data and priors on distribution families and on parameter values. The slow adoption of this approach (since 1978) is because the Bayesian posterior can be complex to calculate, depending on the form of the prior distributions. When both input model uncertainty and model parameter uncertainty are considered, the process is complicated. Suppose that there are Q input models, $q = 1, 2, \ldots, Q$ and C_q choices for input model form for the q^{th} input. If the Bayesian posterior distribution on input model forms exhibits independence across input models, the BMA process is straightforward to implement: sample from the posterior distribution of model forms independently for each input model, then sample from the chosen form's model parameter values according to the parameter posterior distribution. Even if the input distributions are themselves independent, there may be dependence in the Bayesian probabilities of correct model form (e.g. all inputs are either lognormal or exponential). In such a case one may have to consider all possible combinations of models, each combination with its own posterior probability. That could require estimation of $\Pi\, C_q$ posterior probabilities. An input 'instance' for simulation must be randomly drawn from all such possible combinations of model forms for each input model, using the computed posteriors.

Zouaoui and Wilson (2004) presented an alternative BMA approach requiring simulation of all $\Pi\, C_q$ model combinations, each a sufficient number of times to characterize parameter uncertainty. In their example they resampled parameter values 100 times for each model. Simulating all model combinations can be a distinct disadvantage if there are many input variables, each with many possible distribution families, but there is a silver lining to this cloud: their approach does not require model resampling. Instead, the results of simulation runs (averaged over parameter resamples and simulation replications within these) for each model combination are combined in a weighted average, where the weights are the Bayesian posterior values. This approach has an important advantage: an analysis can be augmented to consider additional candidate input model forms while making full use of the existing simulation runs. Only runs for model combinations that include additional model forms must be computed. The sums will use existing and new run results, weighted using the new model form posterior distribution.

Like Chick's method and all other methods that use percentiles of simulation output to establish confidence intervals, Zouaoui and Wilson (2004) used percentile intervals in their computational example, which showed good results. With simulation runs lengths corresponding to 200,000 customers, intrinsic error was negligible compared with extrinsic error, so this result is not surprising. One can observe perhaps minor overcoverage (93% vs. 90%) in Table 2 of Zouaoui and Wilson (2003), for the input sample size = 50,000 case, where extrinsic uncertainty was small relative to intrinsic uncertainty. Since only 200 macroreplications were performed, the 93% must be interpreted as 93% +/- 4% for 95% confidence. Zouaoui and Wilson (2004) also proposed a confidence interval procedure based on the t distribution and a variance approximation based on Satterthwaite's (1941) formula. Their approach appropriately decomposes the extrinsic and intrinsic error components. Unfortunately, as described earlier, with fixed real-world data size it is quite possible to have a highly skewed distribution for extrinsic error - and there is no Central Limit Theorem effect that will make a t-based interval appropriate in that case. While this interval was not used for the 2004 paper, it was used in an example in Zouaoui and Wilson (2001), with estimated coverage for a 90% interval ranging from 75% - 82%. The example was the same as that in the 2003 paper, but with much smaller real-world data set: a sample size of 1,000 instead of 50,000, with consequent larger extrinsic uncertainty.

Assuming that real-world processes generate interarrival or service times that follow some known parametric distribution is often unrealistic, but it is important to remember that the usual performance measures for many queueing systems depend at least approximately on only the first two moments of the

input distributions. In many cases, getting the distribution wrong does not have a significant effect on mean system characteristics. For example see Smith (2003) and Whitt (2004), which give two-moment approximate results for some $M/G/c/K$ and $G/G/1/K$ systems. Further, the BMA approach can capture mixture distributions through the posterior on different model types. Lindsay, Pilla and Basak (2000) show that mixtures can provide higher moment-matching ability in fitting distributions. In an example cited in Zouaoui and Wilson (2004), of the 98% of variance due to extrinsic error, 80% was parameter error and only 18% was attributed to error from using the wrong probability model.

3.5 Approximations Based on Taylor's Theorem

In the final section of his WSC State of the Art Review of input modeling in 1994, Russell Cheng identified a means to separate intrinsic and extrinsic error (Cheng 1994). He characterized extrinsic error through a Taylor approximation, which for our example gives:

$$Var(\overline{Y}(\lambda,\tau)) \approx g'Vg + \sigma^2/m \tag{3}$$

where $g = (\partial L/\partial\lambda, \partial L/\partial\tau)'$ and V is the variance-covariance matrix of the random variable vector $(\Lambda, T)'$. The elements of g were estimated by finite differences from simulation runs using common random numbers. This approximation separates the extrinsic error (which decreases as a function of $\sqrt{1/n}$) from the intrinsic error (which decreases as a function of $\sqrt{1/m}$). Cheng (1994) briefly suggested basing confidence intervals for L on this relation, presumably assuming Gaussian distributions for Λ, T. The details were presented in Cheng and Holland (1997). The *δ-method* strategy is superior for small numbers of input distribution parameters, large sample sizes, and system operating conditions where errors in the first-order Taylor approximation will be small. Cheng and Holland (1998) recognized that the δ-method approach could be impractical for many estimated input parameters, and proposed two *δ/two-point* estimation methods requiring only two simulation runs to compute an approximation to (3). The first finds a finite difference approximation using directions corresponding to the product $\hat{V}\hat{g}$, scaled to a vector $(\gamma_1, \gamma_2)'$ in a way that produces for our example:

$$\frac{\overline{Y}(\lambda+\gamma_1,\tau+\gamma_2) - \overline{Y}(\lambda-\gamma_1,\tau-\gamma_2)}{2} \approx \sqrt{g'Vg}. \tag{4}$$

Cheng and Holland find that spending half the computational budget to estimate $\hat{g}$ and half to compute the estimate in (4) produces good interval size and coverage in a number of settings. A second approximation that does not require $\hat{g}$ is possible if the signs of the components of g are known. In that case set $\gamma_i = \sqrt{\hat{V}_{i,i}}$. Evaluation of the performance of the *simplified δ/two-point* method indicated that it is less robust. The performance comparisons were expanded to include a parametric bootstrap approach in Cheng and Holland (2004).

4 STATE OF THE ART

The study of output analysis methods that capture input model uncertainty has continued over the past ten years, though perhaps with less intensity. New results have been published in the streams of direct resampling and Bayesian Model Averaging. In addition, a promising new method has been developed, based on a recently developed stochastic Kriging metamodeling approach and with common features of both the δ-method and bootstrap strategies.

4.1 Direct Resampling: Distributions of Conditional Expectations

Work has continued to find more efficient ways to characterize conditional expectations. For example, Sun, Apley, and Staum (2011) have proposed a new ANOVA-like estimator of the variance of the conditional expectation. The optimal number of inner-samples (m) for their approach remains bounded as the computation effort grows. This simplifies algorithms used to compute such variances, which are important in financial modeling and other applications.

4.2 Bayesian Model Averaging: the Multivariate Case

While BMA approaches over the first ten years addressed multiple input variables, the simplifying assumption was that these inputs were statistically independent. For many important real-world processes this is not the case.

Biller and Corlu (2011) extended the BMA approach to correlated inputs that could be modeled using the normal-to-anything strategy (NORTA, Cario and Nelson 1997). This strategy allows arbitrary continuous marginal distributions with specified correlations. The uncertainty includes uncertainty for the correlation parameters based on a correlation matrix fitted to a finite set of real-world data. Biller and Corlu produced a practical BMA strategy by combining marginal representations using the Johnson (1949) translation system, Sklar's (1959) marginal-copula representation and Cooke's (1997) copula-vine specification for sampling the parameters of the NORTA distribution. The process for estimating confidence intervals was z-based, and coverages were from 6% to 17% below the nominal 95% level, deteriorating as the intrinsic error is reduced (through longer simulation runs).

4.3 Metamodel-Assisted Bootstrapping

When the simulation response over the range of parameter uncertainty can be captured by a first order Taylor approximation, as in the work by Cheng and co-authors, one might think of this approach as replacing the simulation runs with a (linear) metamodel. For this metamodel (given Gaussian distributions for input parameter uncertainty), the output distribution characterization can be performed analytically. If the response over the range of parameter uncertainty is significantly nonlinear, more complex metamodels might be employed. Barton et al. (1999) used a radial basis function metamodel to propagate uncertainty in product design parameter values to uncertainty in product acceptance by consumers, (or vice versa). This work was focused on deterministic engineering analysis models, and does not have direct applicability for stochastic simulation models.

The stochastic Kriging framework of Ankenman, Nelson and Staum (2010) provides a flexible and well characterized metamodel framework for stochastic simulation. For this metamodel class, computing output distribution characteristics from input model uncertainty characterization analytically is challenging. Instead, one can estimate the distribution of the output by bootstrapping the input data, and using the parameter estimates as arguments for the metamodel. Barton, Nelson and Xie (2010) called this approach *metamodel-assisted bootstrapping*, and used it in the context of known parametric distributions with unknown parameter values. It has three distinct advantages over direct bootstrapping. First, the function that is evaluated for each bootstrap resample is the metamodel so (with the caveat in the next paragraph) bootstrap evaluation does not require computationally expensive simulations. Second, while the stochastic Kriging model allows separate characterization of extrinsic and intrinsic variation, the modeling itself reduces the effect of intrinsic error as a component of the metamodel response variation, so overcoverage risk is greatly reduced. Tables 1 and 2 in Barton, Nelson and Xie (2010) show this advantage compared with direct bootstrap and BMA approaches in limited-data settings. Third, the metamodel, unlike the simulation, is a smooth function of the input resample, thus satisfying the conditions for asymptotic validity of the bootstrap.

This method for input uncertainty characterization has an added stage: before bootstrapping, a set of simulation experimental runs are conducted to fit a metamodel of simulation output as a function of simulation input parameter values. While metamodel-assisted bootstrapping eliminates the rm simulation ef-

fort, it adds an $r_{DOE}m$ simulation effort needed to fit the stochastic Kriging model with high fidelity. When the number of input model parameters is not too large, one would expect $r_{DOE}m << rm$. When the number of input parameters is large, not only would the simulation effort become impractically large; the current technology would not permit fitting stochastic Kriging models with more than a few tens of parameters.

This metamodel-assisted approach could be incorporated into a BMA approach with similar advantages. In fact, this has been done in the deterministic setting by a number of authors; see for example Kennedy and O'Hagan (2001) and Oakley (2004).

5 REMAINING DIFFICULTIES

Figure 2 shows the overall structure of the major approaches covered in this tutorial: direct bootstrap resampling (direct resampling is similar), BMA, δ-method and metamodel-assisted bootstrapping. Each of these approaches provide insight for a previously ignored aspect of model fidelity: input model uncertainty. The first two methods require extensive simulation experiments – a factor of r greater than the naive analysis, with r ranging from hundreds to thousands. The latter two methods require simulation runs only to fit a metamodel, either first order linear models for the δ-method or more extensive space-filling designs for the metamodel-assisted bootstrap method. Large numbers of input parameters can put these latter two methods at a disadvantage in terms of simulation effort. Only the first method has been applied in a in a nonparametric context in which empirical distributions directly drive the simulation.

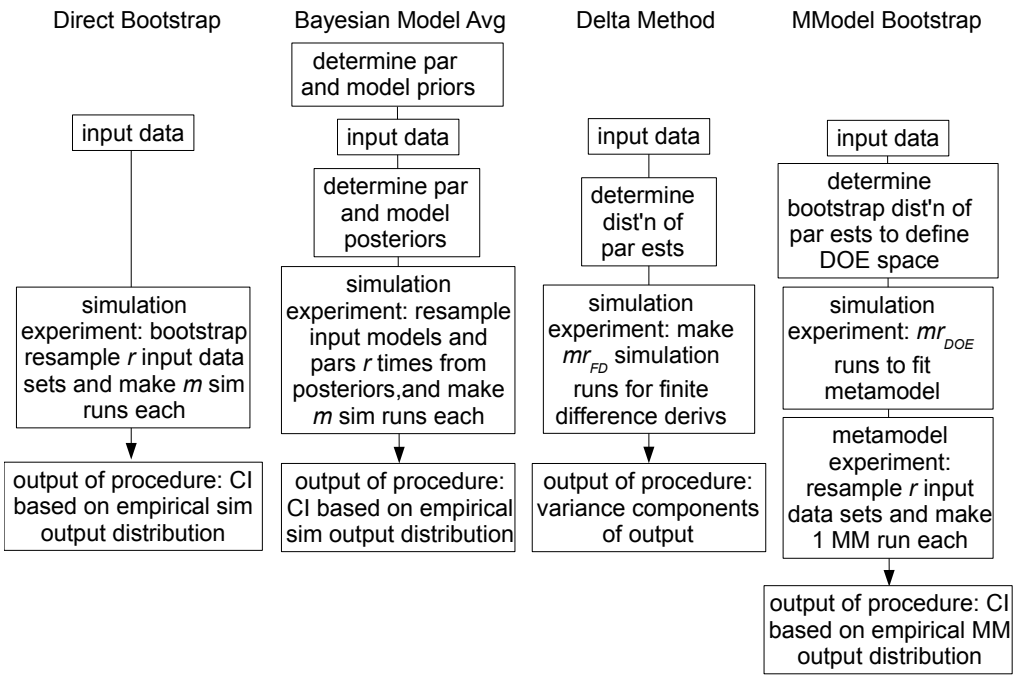

Figure 2. High-level schematic representation of major input uncertainty methods.

Table 1 summarizes the disadvantages for each of the four major approaches. Overall, the direct and BMA approaches avoid issues with metamodel fitting and fidelity, but require extensive computational effort, when compared with the δ-method and metamodel-assisted bootstrap with a small number of parameters. While all but the direct resampling method involve complex computations, the δ -method and metamodel-assisted bootstrap can have these calculations performed in a 'black box' way, so they may still be practical for vendor-provided solutions.

Table 1. Methods for input model uncertainty characterization and their shortcomings.

Problem \ Method	Direct Bootstrap	BMA	δ-Method	MM Bootstrap
Requires rm simulations	X	X		
Requires $r_{DOE}m$ simulations			X (or not)	X
Requires metamodel (or Taylor) fidelity			X	x
Potential overcoverage - intrinsic	X	X		
Potential undercoverage - t		x	X	
Procedurally complex		X	x	x
Violates asymptotic requirements	x			

A small 'x' indicates a problem that is practically small or only occurs for some variants in the literature. From the table, the most promising methods appear to be BMA when intrinsic error is small and percentile intervals are used. When there are relatively few input parameters, the metamodel-assisted bootstrap method is efficient and has robust performance with either small or large intrinsic and/or extrinsic error. For nonparametric applications, the direct bootstrap method has promise, but potential overcoverage must be checked.

ACKNOWLEDGMENTS

The author is indebted to many of those listed in the references below for improving his understanding of input model uncertainty, but particular thanks go to Barry Nelson and Lee Schruben. The author thanks the National Science Foundation for its support of his research while on IPA assignment. The views and conclusions expressed in this paper are those of the author and do not reflect the policies or procedures of the National Science Foundation.

REFERENCES

Ankenman, B., B. L. Nelson, and J. Staum. 2010. "Stochastic Kriging for Simulation Metamodeling." *Operations Research* 58:371–382.

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*, Edited by S. Chick, C.-H. Chen, S.G. Henderson, and E. Yücesan. Fontainebleau, France: 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*, Edited by E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, 353–369. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Barton, R. R., F. Limayem, M. Meckesheimer, and B. Yannou. 1999. "Using Metamodels for Modelling the Propagation of Design Uncertainties." In ICE 99: Proceedings of the 5th International Conference on Concurrent Engineering, 521-528. Nottingham, UK: Centre for Concurrent Enterprising.

Barton, R. R., B. L. Nelson and W. Xie. 2010. "A Framework for Input Uncertainty Analysis." In *Proceedings of the 2010 Winter Simulation Conference*, Edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 1189-1198. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Barton, R. R., and L. W. Schruben. 1993. "Uniform and Bootstrap Resampling of Empirical Distributions." In *Proceedings of the 1993 Winter Simulation Conference*, Edited by G. W. Evans, M. Mol-

laghasemi, W. E. Biles, and E. C. Russell, 503–508. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Barton, R. R., and L. W. Schruben. 2001. "Resampling Methods for Input Modeling." In *Proceedings of the 2001 Winter Simulation Conference*, Edited by B. A. Peters, J. S. Smith, D. J. Medeiros and M. W. Rohrer, 372–378. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Biller, B., and C. Corlu. 2011. "Accounting for Parameter Uncertainty in Large-Scale Stochastic Simulations with Correlated Inputs." *Operations Research* 59:661-673.

Cario, M. C. and Nelson, B. L. 1997. "Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix." Technical Report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois.

Cheng, R. C. H. 1994. "Selecting Input Models." In *Proceedings of the 1994 Winter Simulation Conference*, Edited by J. D. Tew, M. S. Manivannan, D. A. Sadowski and A. F. Seila, 184–191. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Cheng, R. C. H. 1995. "Bootstrap Methods in Computer Simulation Experiments." In *Proceedings of the 1995 Winter Simulation Conference*, Edited by C. Alexopoulos, K. Kang, W. R. Lilegdon and D. Goldsman, 171-177. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Cheng, R. C. H. 1998. "Bayesian Model Selection when the Number of Components is Unknown." In *Proceedings of the 1998 Winter Simulation Conference*, Edited by D. J. Medeiros, E. F. Watson, J. S. Carson and M. S. Manivannan, 653–659. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

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* 14:344–362.

Chick, S. E. 1997. "Bayesian Analysis for Simulation Input and Output." In *Proceedings of the 1997 Winter Simulation Conference*, Edited by S. Andradottir, K. J. Healy, D. H. Withers, and B. L. Nelson, 253–260. Piscataway, New Jersey: Institute of Electronic and Electrical Engineers, Inc.

Chick, S. E. 1999. "Steps to Implement Bayesian Input Distribution Selection." In *Proceedings of the 1999 Winter Simulation Conference*, Edited by P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, 317–324. Piscataway, New Jersey: Institute of Electronic and Electrical Engineers, Inc.

Chick, S. E. 2000. "Bayesian Methods for Simulation." In *Proceedings of the 2000 Winter Simulation Conference*, Edited by J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, 109–118. Piscataway, New Jersey: Institute of Electronic and Electrical Engineers, Inc.

Chick, S. E. 2001. "Input Distribution Selection for Simulation Experiments: Accounting for Input Uncertainty." *Operations Research* 49:744-758.

Cooke, R. M. 1997. "Uncertainty Modeling: Examples and Issues." *Safety Science* 26:49–60.

Efron, B., and R. Tibshirani. 1986. "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy." Statistical Science 1:54-77.

Freimer, M., and L. W. Schruben. 2002. "Collecting Data and Estimating Parameters for Input Distributions." In *Proceedings of the 2002 Winter Simulation Conference*, Edited by E. Yücesan, C. Chen, J. L. Snowdon, and J. M. Charnes, 392–399. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Glynn, P. 1986. "Problems in Bayesian Analysis of Stochastic Simulation." In *Proceedings of the 1986 Winter Simulation Conference*, Edited by J. R. Wilson, J. O. Henriksen, and S. D. Roberts, 376–383. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Hoeting, J. A., D. Madigan, A. D. Raftery and C. T. Volinsky. 1999. "Bayesian Model Averaging: A Tutorial (with discussion)." *Statistical Science* 14:382-417.

Johnson, N. L. 1949. "Systems of Frequency Curves Generated by Methods of Translation." *Biometrika* 36:149-176.

Kennedy, M. C. and A. O'Hagan. 2001. "Bayesian Calibration of Computer Models (with discussion)." *Journal of the Royal Statistical Society Series B*, 63:425–464.

Lee, S., and P. W. Glynn. 1999. "Computing the Distribution Function of a Conditional Expectation via Monte Carlo: Discrete Conditioning Spaces." In *Proceedings of the 1999 Winter Simulation Conference*, Edited by P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, 1654-1663. Piscataway, New Jersey: Institute of Electronic and Electrical Engineers, Inc.

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* 13:238–258.

Lindsay, B. G., R. S. Pilla and P. Basak. 2000. "Moment-Based Approximations of Distributions Using Mixtures: Theory and Applications." *Annals of the Institute of Statistical Mathematics* 52:215-230.

Oakley, J. 2004. "Estimating Percentiles of Uncertain Computer Code Outputs." *Applied Statistics* 53:83-93.

Rubin, D. B. 1981. "The Bayesian Bootstrap." *Annals of Statistics* 9:130-134.

Sargent, R. G. 2011. "Verification and Validation of Simulation Models." In *Proceedings of the 2011 Winter Simulation Conference*, Edited by S. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu, 183–198. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Satterthwaite, F. E. 1941. "Synthesis of Variance." *Psychometrika* 6:309–316.

Sklar, A. 1959. "Fonctions de Répartition à n Dimensions et Leurs Marges." *Publications de l'Institut de Statistique de l'Université de Paris* 8: 229–231.

Smith, J. M. 2003. "M/G/c/K Blocking Probability Models and System Performance." *Performance Evaluation* 52:237-267.

Steckley, S., and S. G. Henderson. 2003. "A Kernel Approach to Estimating the Density of a Conditional Expectation." In *Proceedings of the 2003 Winter Simulation Conference*, Edited by S. E. Chick, P. J. Sanchez, D. Ferrin, and D. J. Morrice (Eds.), 383–391. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Sun, Y., D. W. Apley, and J. Staum. 2011. "Efficient Nested Simulation for Estimating the Variance of a Conditional Expectation." *Operations Research* 59:998-1007.

Whitt, W. 2004. "Heavy-Traffic Limits for Loss Proportions in Single Server Queues." *Queueing Systems* 46:507-536.

Zouaoui, F., and J. R. Wilson. 2001. "Accounting for Input Model and Parameter Uncertainty in Simulation." *In Proceedings of the 2001 Winter Simulation Conference*, Edited by B. A. Peters, J. S. Smith, D. J. Medeiros and M. W. Rohrer, 290-299. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

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 BIOGRAPHY

RUSSELL R. BARTON is a professor in the Department of Supply Chain and Information Systems at the Pennsylvania State 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.