Proceedings of the 2001 Winter Simulation Conference
B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, eds.

ACCOUNTING FOR INPUT MODEL AND PARAMETER UNCERTAINTY IN SIMULATION

Faker Zouaoui

Sabre, Inc.
Research Group
Southlake, TX 76092, U.S.A.

James R. Wilson

Department of Industrial Engineering
North Carolina State University
Raleigh, NC 27695-7906, U.S.A.

ABSTRACT

Taking into account input-model, input-parameter, and stochastic uncertainties inherent in many simulations, our Bayesian approach to input modeling yields valid point and confidence-interval estimators for a selected posterior mean response. Exploiting prior information to specify the prior plausibility of each candidate input model and to construct prior distributions on the model's parameters, we combine this information with the likelihood function of sample data to compute posterior model probabilities and parameter distributions. Our Bayesian Simulation Replication Algorithm involves: (a) estimating parameter uncertainty by sampling from the posterior parameter distributions on selected runs; (b) estimating stochastic uncertainty by multiple independent replications of those runs; and (c) estimating model uncertainty by weighting the results of (a) and (b) using the corresponding posterior model probabilities. We allocate runs in (a) and (b) to minimize final estimator variance subject to a computing-budget constraint. An experimental performance evaluation demonstrates the advantages of this approach.

1 INTRODUCTION

The widespread application of stochastic discrete-event simulations is accompanied by a widespread concern about quantifying the uncertainties prevailing in their use. There are three main sources of uncertainty in a simulation experiment (Zouaoui and Wilson 2001a):

1. *Stochastic uncertainty.* This source of variation arises from the dependence of the simulation output on the uniform random variates (random numbers) that are generated within the simulation and then used to sample nonuniform random variates from the input models driving each simulation run.
2. *Model uncertainty.* This source of variation typically occurs when choosing between different input models that adequately fit the available sample data or subjective information.
3. *Parameter uncertainty.* This source of variation arises because the parameters of the selected input model(s) are unknown and must be estimated from available sample data or subjective information.

Using the same symbolism to describe the layout of probabilistic simulation experiments that was introduced in Zouaoui and Wilson (2001a), we initially assume for the sake of simplicity that the target simulation experiment involves a single univariate input process for which the probabilistic input model M (with corresponding c.d.f. $G_M(\cdot, \theta_M)$ and d_M-dimensional parameter vector θ_M) is subject to uncertainty. For example, in a queueing simulation with a prescheduled sequence of customer arrival times, $G_M(\cdot, \theta_M)$ might represent the service-time distribution, which is thought to be either exponential or lognormal; and the parameters of these alternative input models must be estimated from expert opinion and limited sample data. Thus the model and parameter uncertainty are represented by the random variables M and θ_M, respectively, both of which are assumed to depend only on the available subjective information or sample data; and the stochastic uncertainty depends only on the randomness of the vector of uniform variates (random numbers) $\boldsymbol{u}$ that are used to generate samples from $G_M(\cdot, \theta_M)$ during each simulation run. In this situation an output quantity of interest from the simulation run, y, can be regarded as an unknown complicated function of $\boldsymbol{u}$, M, and θ_M,

$$y = y(\boldsymbol{u}, M, \theta_M). \quad (1)$$

Let

$$\eta(M, \theta_M) = \mathrm{E}(y|M, \theta_M) = \int y(\boldsymbol{u}, M, \theta_M)\, d\boldsymbol{u}, \quad (2)$$

be the expected value of y, given the input model M and the corresponding parameter vector θ_M.

The objective of the classical simulation experiment is generally to estimate $\eta(M_0, \theta_0)$, where θ_0 is the true but unknown value of the parameter vector θ_{M_0} under the true model M_0, estimated separately from the simulation experiment using real data. In general this approach fails to assess and propagate model and parameter uncertainty and may lead to miscalibrated uncertainty assessments about y (Draper 1995). The δ-method (Stuart and Ord 1994) and the bootstrap method (Cheng and Holland 1997) are possible ways to account for parameter uncertainty. However, in addition to their failure to incorporate relevant information other than the observed data points, it is unclear that these methods can be extended to account for model uncertainty (Zouaoui and Wilson 2001a).

In this paper, we present the Bayesian Model Averaging (BMA) approach as a coherent mechanism to account for all sources of uncertainty in a simulation experiment (Hoeting et al. 1999). The basic ingredients of the BMA approach for conducting simulation experiments are discussed in Section 2. In Section 3 we develop a "BMA-Based Simulation Replication Algorithm" to estimate the posterior mean response and assess the variability of the resulting estimator. In Section 4 we use the output of the algorithm to estimate the components of this variance that are due to each source of uncertainty. In Section 5 we develop a replication allocation procedure that optimally allocates simulation runs to input models so as to minimize the variance of the estimated posterior mean response subject to a budget constraint on the total amount of simulated experimentation or computer time that is available. Finally in Section 6, we conduct a Monte Carlo experiment on a computer communications network application to evaluate the performance of the BMA approach versus conventional techniques for estimating the posterior mean response and assessing its variability. For a more complete discussion of the results presented in this paper (including proofs of all theorems), see Zouaoui and Wilson (2001b).

2 THE BMA APPROACH

Assume that we have Q random inputs driving our simulation model. We observe the sample data $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_Q)$, where $\boldsymbol{x}_q = (x_{q1}, \ldots, x_{qn_q})$ is the vector of observations based on a random sample of size n_q for the qth random input. Even though stochastic dependencies among simulation inputs will not affect our formulation, we assume for simplicity that these random inputs are independent. Let $M_{\ell,q}$ represent the ℓth adequate model of the qth random input with prior probability $p(M_{\ell,q})$ for $\ell = 1, \ldots, K_q$ so that $\sum_{\ell=1}^{K_q} p(M_{\ell,q}) = 1$ for $q = 1, \ldots, Q$. In practice, most of the K_q's will be one, and only the random inputs with high model uncertainty and enough information to assess such an uncertainty will have K_q's larger than one.

To simplify the notation, we define the set of candidate models $\mathcal{M}$ to consist of models $\{M_k : k = 1, \ldots, K\}$, where $K = K_1 \times \cdots \times K_Q$ is the total number of different input model combinations each having prior probability of the form $p(M_k) = \prod_{q=1}^{Q} p(M_{\ell_{qk},q})$ for some $\ell_{qk} \in \{1, \ldots, K_q\}$ and for $q = 1, \ldots, Q$ and $k = 1, \ldots, K$. Once $\mathcal{M}$ is chosen, we let θ_k denote the d_k-dimensional vector of parameters under model $M_k \in \mathcal{M}$ with prior distribution $p(\theta_k|M_k)$, where $k = 1, \ldots, K$.

Although the choice of alternative models in the set $\mathcal{M} = \{M_k : k = 1, \ldots, K\}$ is highly dependent on the specific application, Zouaoui and Wilson (2001b) provide general comments in the simulation input modeling context. Theorem 1 establishes the expected value of the target response, y.

Theorem 1. *If the simulation output response has the form (1), then*

$$\mathrm{E}(y|\boldsymbol{X}) = \sum_{k=1}^{K} p(M_k|\boldsymbol{X}) \int \eta(M_k, \theta_k)\, p(\theta_k|\boldsymbol{X}, M_k)\, d\theta_k. \tag{3}$$

There are thus three ingredients for the implementation of the BMA approach in discrete-event simulations:

1. The specification of the prior probabilities $\{p(M_k) : k = 1, \ldots, K\}$ over which model uncertainty is propagated, and the selection of the prior distributions $\{p(\theta_k|M_k) : k = 1, \ldots, K\}$ for the model parameters.
2. The computation of the posterior distributions $\{p(\theta_k|\boldsymbol{X}, M_k) : k = 1, \ldots, K\}$.
3. The computation of the posterior model probabilities $\{p(M_k|\boldsymbol{X}) : k = 1, \ldots, K\}$.

Each of these components is addressed in the subsections that follow.

2.1 Specification of Priors

The specification of the prior model probabilities $\{p(M_k)\}$ is typically context specific. When there is little prior information about the relative plausibility of the models considered, the assumption that all models are equally likely a priori is a reasonable choice (Madigan and Raftery 1994). Different prior probabilities can be viewed as derived from previous data and representing the relative success of the models in predicting those previous data.

The easiest way to deal with the problem of specifying the prior distributions $p(\theta_k|M_k)$ on the model parameters is to ignore them and simply use the Schwarz criterion

(Schwarz 1978). Although this will lead to appropriate conclusions in "sufficiently large" samples, there is not much available guidance on the operational meaning of the qualifying phrase "sufficiently large." Typically, these distributions are specified based on information accumulated from past studies, or from expert opinions. In order to simplify the subsequent computational burden, experimenters often limit this choice somewhat by restricting priors to some familiar distributional family. An even simpler alternative, available in some cases, is to endow the prior distribution with little information content, so that the data from the current study will be the dominant factor in determining the posterior distribution. We address each of these approaches in Zouaoui and Wilson (2001b) and in Zouaoui (2001).

2.2 Computation of Posterior Parameter Distributions

The second ingredient for the implementation of the BMA approach is to compute the posterior distributions $\{p(\theta_k|\boldsymbol{X}, M_k),\ k = 1, \ldots, K\}$. Conditioning on the known value of the data $\boldsymbol{X}$ and using Bayes' rule, we obtain the posterior density,

$$p(\theta_k|\boldsymbol{X}, M_k) \quad = \quad \frac{p(\theta_k|M_k)\ p(\boldsymbol{X}|M_k, \theta_k)}{p(\boldsymbol{X}|M_k)},$$

where $p(\boldsymbol{X}|M_k)$ is the marginal distribution of the data $\boldsymbol{X}$, given model M_k.

For some models, with a specific choice of a prior distribution such as a conjugate prior, the posterior distribution can easily be recognized from the unnormalized posterior density, $p(\theta_k|\boldsymbol{X}, M_k) \propto p(\theta_k|M_k)\ p(\boldsymbol{X}|M_k, \theta_k)$. This removes the burden of computing the normalizing constant $p(\boldsymbol{X}|M_k)$. However, we cannot limit the choices of priors to specific distributional families in all applications, and we will generally have some unnormalized densities that do not belong to any of the well-known distributions. To generate a sample from the posterior distribution, we should compute the exact form of its density. This requires some high-dimensional numerical integrations or asymptotic approximations (Gelman et al. 1995). Markov Chain Monte Carlo (MCMC) methods have been used increasingly for dealing with such problems. The basic philosophy behind MCMC is to take a Bayesian approach and carry out the necessary numerical integrations using Monte Carlo simulation; see Gilks, Richardson, and Spiegelhalter (1996) for background. Instead of calculating exact or approximate estimates of the posterior density, this computer-intensive technique generates a stream of simulated values from the posterior distribution of any parameter or quantity of interest. These computations can be easily coded in the BUGS statistical package (Spiegelhalter et al. 1996) using a small set of BUGS commands.

2.3 Computation of Posterior Model Probabilities

The posterior model probabilities $p(M_k|\boldsymbol{X})$ are computed as follows:

$$p(M_k|\boldsymbol{X}) = \frac{p(M_k)\ p(\boldsymbol{X}|M_k)}{\sum_{j=1}^{K} p(M_j)\ p(\boldsymbol{X}|M_j)}$$

for $k = 1, \ldots, K$. The evaluation of these probabilities comes down to computing the marginal data density given model M_k,

$$p(\boldsymbol{X}|M_k) \quad = \quad \int p(\boldsymbol{X}|M_k, \theta_k)\ p(\theta_k|M_k)\, d\theta_k. \qquad (4)$$

The integral in (4) may be evaluated analytically for distributions in the regular exponential family with conjugate priors. However, (4) is generally intractable and thus must be computed by numerical methods. In Zouaoui and Wilson (2001b) and Zouaoui (2001), we review various numerical integration strategies and provide good asymptotic approximations.

3 ESTIMATING MEAN RESPONSE

In our inference about the output quantity of interest, y, we focus on estimating its mean response given by equation (3) and assessing its variability. Chick (1999) proposed an algorithm for estimating the output mean response. He suggested that for the rth simulation run, we need to sample a model M^r from its discrete posterior probability mass function $\{p(M_k|\boldsymbol{X}) : k = 1, \ldots, K\}$ and then sample its vector of parameters $\theta^r_{M^r}$ from its posterior distribution $p(\theta_{M^r}|\boldsymbol{X}, M^r)$. The mean response estimate would be the average of all output responses $\{y_r : r = 1, \ldots, R\}$, computed using the randomly sampled input models $\{M^r : r = 1, \ldots, R\}$ and their corresponding randomly sampled parameter vectors $\{\theta^r_{M^r} : r = 1, \ldots, R\}$.

This algorithm gives a good estimate of the mean response for a large number of runs, but we believe that it has several deficiencies that makes it of limited use to simulation practitioners. Most importantly, Chick's algorithm cannot accommodate more models without repeating all the simulation runs. If for some reason we decide to expand our summation in (3) to have more than K models, then we need to repeat all the runs with the new sampled models and their parameters to obtain a new estimate for the posterior mean response.

Figure 1 summarizes an algorithm which implements the BMA approach for designing simulation experiments and overcomes all the above deficiencies. The net effect of the algorithm is to account for the full extent of the model and parameter uncertainty as well as the usual stochastic uncertainty. The innermost loop of the algorithm will be used

for $k = 1, \ldots, K$
  set input model $M \leftarrow M_k$
  for $r = 1, \ldots, R_k$
    generate the rth sample θ^r from $p(\theta|\boldsymbol{X}, M)$
    set the parameter vector $\theta \leftarrow \theta^r$
    for $j = 1, \ldots, m$
      set the random-number input $\boldsymbol{u} \leftarrow \boldsymbol{u}_j$
      perform the jth run using $\boldsymbol{u}$, M, and θ
      calculate the response $y_{krj} = y(\boldsymbol{u}, M, \theta)$
    end loop
    compute $\overline{y}_{kr} = \sum_{j=1}^{m} y_{krj}/m$
  end loop
  compute the model mean $\overline{\overline{y}}_k = \sum_{r=1}^{R_k} \overline{y}_{kr}/R_k$
end loop
compute the weighted mean $\sum_{k=1}^{K} p(M_k|\boldsymbol{X})\overline{\overline{y}}_k$
as an estimate for $E(y|\boldsymbol{X})$

Figure 1: BMA-Based Simulation Replication Algorithm

to generate estimates for the stochastic uncertainty, whereas the middle loop will assess the parameter uncertainty for each input model. Finally, the outermost loop will be used to estimate the model uncertainty.

Every model in the set $\mathcal{M}$ can have a different predictive inference on the output of interest y, and the composite inference will be a weighted average of the predictive distributions $p(y|\boldsymbol{X}, M_k)$ for $k = 1, \ldots, K$. This explains our idea of not resampling the input model M and its associated parameter vector θ_M prior to each replication. If the simulation model is costly in terms of computing time and the number of possible input models is large, then we can eliminate the models which explain the data far less than others using the Occam's window method (Madigan and Raftery 1994).

The total sample sizes $\{R_k : k = 1, \ldots, K\}$ respectively generated from the posterior distributions $\{p(\theta_k|\boldsymbol{X}, M_k) : k = 1, \ldots, K\}$ may not be necessarily the same, because the effect of parameter uncertainty on the variability of the output response, given model M_k, is usually different for different input models. Theoretically, the accuracy of our estimate of the posterior mean response improves as all the R_k's get large. However, we are usually restricted in practice by a fixed number of simulation runs N. In Section 5, we propose a method to specify the $\{R_k\}$ that optimally allocates the total simulation effort to the different models, based on the minimization of the variance of the estimator of the posterior mean response subject to a constraint on the total number of simulation runs N.

4 ASSESSING OUTPUT VARIABILITY

In this section, we seek to assess the variability of the simulation-generated output based on a decomposition of the posterior variance $\mathrm{Var}(y|\boldsymbol{X})$. To simplify the notation, we denote our posterior model probabilities as $p_k = p(M_k|\boldsymbol{X})$ for $k = 1, \ldots, K$. In view of equation (1), we have the following representation for the simulation output response y_{krj}:

$$\begin{aligned} y_{krj} &= y(\boldsymbol{u}_j, M_k, \theta_k^r) \\ &= \eta(M_k, \theta_k^r) + e_j(\boldsymbol{u}_j, M_k, \theta_k^r) \end{aligned} \tag{5}$$

for $k = 1, \ldots, K$; $r = 1, \ldots, R_k$; and $j = 1, \ldots, m$. The error variable e_j is the random difference between the simulation output response y_{krj} and $\eta(M_k, \theta_k^r)$. We generally assume that

$$\mathrm{E}(e_j|M_k, \theta_k^r) = 0 \quad \text{and} \quad \mathrm{Var}(e_j|M_k, \theta_k^r) = \tau_k^2, \tag{6}$$

so that $\mathrm{E}(y_{krj}|M_k, \theta_k^r) = \eta(M_k, \theta_k^r)$. Here we assume that τ_k^2 does not depend on θ_k^r because we are interested in obtaining a measure of the average variability in the output due to stochastic uncertainty. The effect of the randomness in θ_k^r will be captured instead by the randomness in $\eta(M_k, \theta_k^r)$, which will give a measure of the output variability due to parameter uncertainty. Moreover, given that our main objective is to estimate the overall mean response, we can further assume that

$$\eta(M_k, \theta_k^r) = \beta_k + \delta_{kr}(M_k, \theta_k^r), \tag{7}$$

where $\beta_k = \mathrm{E}_{\theta_k^r}\left[\eta(M_k, \theta_k^r)\right] = \int \eta(M_k, \theta_k) p(\theta_k|\boldsymbol{X}, M_k) d\theta_k$ $= \mathrm{E}(y|\boldsymbol{X}, M_k)$, using Theorem 1, and

$$\mathrm{E}_{\theta_k^r}(\delta_{kr}|M_k) = 0 \quad \text{and} \quad \mathrm{Var}_{\theta_k^r}(\delta_{kr}|M_k) = \sigma_k^2. \tag{8}$$

Based on these assumptions, Theorem 2 shows that the posterior variance is the sum of three variances measuring the model, parameter and stochastic uncertainty.

Theorem 2. *If (5), (6), (7), and (8) hold, then*

$$\mathrm{Var}(y|\boldsymbol{X}) = \sum_{k=1}^{K} p_k(\beta_k - \beta)^2 + \sum_{k=1}^{K} p_k\sigma_k^2 + \sum_{k=1}^{K} p_k\tau_k^2,$$

where $\mathrm{E}(y|\boldsymbol{X}) = \sum_{k=1}^{K} p_k\,\beta_k = \beta$.

The response surface model given by (5)–(8) for each input model M_k is known in the statistical literature as the classical random-effects model (Rao 1997), where one estimates β_k, τ_k^2, and σ_k^2 from the output of the algorithm

in Figure 1 as follows:

$$\widehat{\beta}_k = \overline{\overline{y}}_k,$$

$$\widehat{\tau}_k^2 = \frac{\sum_{r=1}^{R_k}\sum_{j=1}^{m}(y_{krj}-\overline{y}_{kr})^2}{R_k(m-1)}, \tag{9}$$

and

$$\widehat{\sigma}_k^2 = \frac{\sum_{r=1}^{R_k}(\overline{y}_{kr}-\overline{\overline{y}}_k)^2}{(R_k-1)} - \frac{\widehat{\tau}_k^2}{m}. \tag{10}$$

From the above estimates, we can estimate the three variance components of Theorem 2 as

$$\left.\begin{aligned} &\widehat{V}_{\text{mod}} = \sum_{k=1}^{K} p_k(\widehat{\beta}_k - \widehat{\beta})^2, \quad \text{where} \quad \widehat{\beta} = \sum_{k=1}^{K} p_k\widehat{\beta}_k, \\ &\widehat{V}_{\text{par}} = \sum_{k=1}^{K} p_k\widehat{\sigma}_k^2, \qquad \text{and} \qquad \widehat{V}_{\text{sto}} = \sum_{k=1}^{K} p_k\widehat{\tau}_k^2. \end{aligned}\right\}$$

In Subsection 5.3, we present a method for constructing an approximate $100(1-\alpha)\%$ confidence interval for the posterior mean response under any scheme for allocating the sample sizes $\{R_k : k = 1,\ldots,K\}$ among the input models.

5 REPLICATION ALLOCATION PROCEDURES

We describe in this section two methods to determine the sample sizes $\{R_k : k = 1,\ldots,K\}$ that are respectively allocated to the models $\{M_k : k = 1,\ldots,K\}$ based on the practical assumption that the total computational effort is generally limited by a fixed number of simulation replications N. We assume further that the stochastic variability can be assessed by a small number of replications m that are fixed prior to the simulation experiment.

The first replication allocation method is based on minimizing the variance of our posterior mean response estimate. Assuming that all the simulation replications are independent, we have the following result.

Theorem 3. *If (5), (6), (7), and (8) hold, then the BMA-Based Simulation Replication Algorithm of Figure 1 yields*

$$\text{Var}(\widehat{\beta}) = \sum_{k=1}^{K} p_k^2\frac{\sigma_k^2}{R_k} + \sum_{k=1}^{K} p_k^2\frac{\tau_k^2}{mR_k}. \tag{11}$$

5.1 Optimal Allocation Procedure

To minimize the variance (11) subject to a budget constraint on the total number of runs, we must solve the following optimization problem,

$$\left.\begin{aligned} &\min_{\{R_k:1\le k\le K\}} \quad \sum_{k=1}^{K}\frac{p_k^2}{R_k}\left[\sigma_k^2+\tau_k^2/m\right] \\ &\text{subject to:} \quad \sum_{k=1}^{K} R_k = N/m = N'. \end{aligned}\right\} \tag{12}$$

We reformulate (12) as an unconstrained optimization problem using the method of Lagrange multipliers to show that, modulo rounding, the optimal sample sizes are given by

$$R_k^* = \frac{N' p_k\sqrt{\vartheta_k}}{\sum_{i=1}^{K} p_i\sqrt{\vartheta_i}} \text{ for } k = 1,\ldots,K, \tag{13}$$

where $\vartheta_k = \sigma_k^2 + \tau_k^2/m$ for $k = 1,\ldots,K$. Note that R_k^* depends on the ϑ_k values, which are unknown and usually estimated after observing the actual output responses. We suggest a two-phase replication allocation procedure that exploits the above result. In the first phase, we can make a small, equal number of pilot runs at each model M_k; and then for $k = 1,\ldots,K$, we estimate ϑ_k by $\widehat{\vartheta}_k = \widehat{\sigma}_k^2 + \widehat{\tau}_k^2/m$, where $\widehat{\tau}_k^2$ and $\widehat{\sigma}_k^2$ are estimated using (9) and (10), respectively. In the second phase, we allocate the rest of the runs according to (13). Assuming that the variance estimates are constant from the first phase to the second phase, we see that the two-phase replication allocation procedure delivers a smaller variance for the mean response compared to the equal allocation scheme

$$R_k = \frac{N'}{K} \text{ for } k = 1,\ldots,K. \tag{14}$$

5.2 Proportional Allocation Procedure

One feasible solution to the optimization problem (12) is the proportional allocation procedure

$$R_k = p_k N \text{ for } k = 1,\ldots,K, \tag{15}$$

which is also optimal if all the ϑ_k's in (14) happen to be equal. This allocation scheme can be easily implemented prior to making the simulation runs, and it overcomes the problem of having to estimate the variances in the optimal allocation procedure (13). Moreover, the mean estimator $\widehat{\beta}_{\text{pa}}$, computed from the BMA-Based Simulation Replication Algorithm given in Figure 1 and the allocation scheme (15), has a smaller variance compared to the mean estimator $\widehat{\beta}_{\text{srs}}$ computed using the Simple Random Sampling (SRS)

procedure. The SRS procedure is similar to Chick's (1999) approach described in Section 3, where we randomly sample a new input model and its vector of parameters from their posterior distributions prior to each run, and then perform m independent runs for a total of N' runs. We formally state this result in the following theorem.

Theorem 4. *If (5), (6), (7), and (8) hold, then with the proportional allocation scheme (15) we obtain the following reduction in variance of the posterior mean estimator versus simple random sampling:*

$$\mathrm{Var}(\widehat{\beta}_{\mathrm{srs}}) - \mathrm{Var}(\widehat{\beta}_{\mathrm{pa}}) \quad = \quad \frac{1}{N'} \sum_{k=1}^{K} p_k (\beta_k - \beta)^2 > 0.$$

5.3 Confidence Interval for the Posterior Mean with Any Allocation Procedure

In this subsection we derive a t-type confidence interval for β based on estimating the variance of our posterior mean estimate $\widehat{\beta} = \sum_{k=1}^{K} p_k \overline{\overline{y}}_k$. This interval can be used with any of the allocation schemes described in the above subsections. We proved in Theorem 3 that the variance of our mean estimate is given by

$$\mathrm{Var}(\widehat{\beta}) \quad = \quad \sum_{k=1}^{K} p_k^2 \frac{\sigma_k^2}{R_k} + \sum_{k=1}^{K} p_k^2 \frac{\tau_k^2}{m R_k},$$

so that we have the following estimator for the variance of the posterior mean estimator $\widehat{\beta}$:

$$\begin{aligned} \widehat{\mathrm{Var}}(\widehat{\beta}) \quad &= \quad \sum_{k=1}^{K} p_k^2 \left(\frac{\widehat{\sigma}_k^2}{R_k} + \frac{\widehat{\tau}_k^2}{m R_k} \right) \\ &= \quad \sum_{k=1}^{K} p_k^2 \widehat{\mathbb{V}}_k, \end{aligned} \tag{16}$$

where we combine equations (9) and (10) to obtain the auxiliary variance estimators

$$\widehat{\mathbb{V}}_k \quad = \quad \frac{\sum_{r=1}^{R_k} (\overline{y}_{kr} - \overline{\overline{y}}_k)^2}{R_k (R_k - 1)} \text{ for } k = 1, \ldots, K.$$

Assuming that $\{R_k : k = 1, \ldots, K\}$ are fixed quantities, we can use the approximation of Satterthwaite (1946), who showed that the complex variance estimator (16) has a distribution that is approximately chi-squared with "effective"

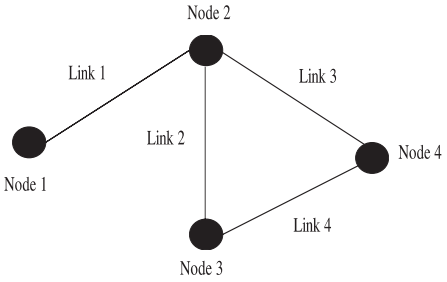

Figure 2: A Communication Network with $\mathbb{Q} = 4$ Nodes and $\mathbb{L} = 4$ Links

degrees of freedom given by

$$f_{\mathrm{eff}} \quad = \quad \left\lfloor \frac{\left(\sum_{k=1}^{K} p_k^2 \widehat{\mathbb{V}}_k \right)^2}{\sum_{k=1}^{K} \frac{p_k^4 \widehat{\mathbb{V}}_k^2}{(R_k - 1)}} \right\rfloor.$$

Thus an approximate $100(1-\alpha)\%$ confidence interval for $\beta = \sum_{k=1}^{K} p_k \beta_k$ is

$$\sum_{k=1}^{K} p_k \overline{\overline{y}}_k \quad \pm \quad t_{1-\alpha/2, f_{\mathrm{eff}}} \left(\sum_{k=1}^{K} p_k^2 \widehat{\mathbb{V}}_k \right)^{1/2}. \tag{17}$$

We will use (17) to evaluate the performance of the proportional allocation procedure (15) and the optimal allocation procedure (13) empirically using a Monte Carlo experiment of a computer communication network.

6 MONTE CARLO EXPERIMENT

6.1 Description

In this example we consider a simulation of a computer communications network (Kleinrock 1976). It is a collection of $\mathbb{Q}$ nodes consisting of computing resources which communicate with each other along a set of $\mathbb{L}$ links (the data communication channels). The aim of the simulation study is to measure the delay in messages transmitted between nodes via the communication channels. Figure 2 illustrates a network with $\mathbb{Q} = 4$ and $\mathbb{L} = 4$.

The $\mathbb{L}$ communication channels are assumed to be noiseless, and have a capacity of C_i bits per second for the ith channel. The $\mathbb{Q}$ nodes carry out the administration tasks such as message reassembly and routing. It is assumed that the nodal processing times are constant with value $\mathcal{T}_i$ for the ith node. In addition there are channel queueing and transmission delays. Traffic entering the network from any

node forms a Poisson process with rate $\gamma(i, j)$ (messages per second) for those messages originating at node i and destined for node j. Each message is assumed to have a length X that is independently sampled from a mixture distribution given by

$$f(x) = \pi_1 f_1(x) + \pi_2 f_2(x) + \pi_3 f_3(x) \text{ for all } x, \quad (18)$$

where the mixture (18) is composed of exponential, lognormal, and uniform probability density functions, respectively:

$$\begin{aligned} f_1(x) &= \lambda e^{-\lambda x} \quad (x \geq 0), \\ f_2(x) &= \frac{1}{\sqrt{2\pi}\sigma x} e^{-(\log x - \mu)^2/(2\sigma^2)} \quad (x \geq 0), \\ f_3(x) &= \frac{1}{b-a} \quad (a \leq x \leq b). \end{aligned}$$

We assume that all nodes have unlimited storage capacity and that all messages are directed through the network on fixed paths. In high speed networks spanning large geographical regions, it may be important to include the propagation time H_i, which is the time required for the energy representing a single bit to propagate down the length of the ith channel. The speed of energy propagation, v miles per second, is a significant fraction of the speed of light depending on the particular type of channel used. If the ith channel has length l_i miles, then $H_i = l_i/v$. Thus if a message has X bits then the time it occupies the ith channel will be $H_i + X/C_i$ seconds.

Some of the parameters in the network were known exactly: $\mathcal{T}_i = 0.001$ seconds ($i = 1, \ldots, \mathbb{Q}$), $C_i = 275,000$ bits/second ($i = 1, \ldots, \mathbb{Q}$), $l_i = i \times 100$ miles ($i = 1, \ldots, \mathbb{L}$), and $v = 150,000$ miles/second. The traffic arrival rates were: $\gamma(1,2) = 60$, $\gamma(1,3) = 40$, $\gamma(1,4) = 50$, $\gamma(2,1) = 80$, $\gamma(2,3) = 65$, $\gamma(2,4) = 20$, $\gamma(3,1) = 100$, $\gamma(3,2) = 22$, $\gamma(3,4) = 26$, $\gamma(4,1) = 40$, $\gamma(4,2) = 50$, $\gamma(4,3) = 60$. The true parameters of the mixture distribution (18) are: $\pi_{0,1} = 0.6$, $\lambda_0 = 1/300$, $\pi_{0,2} = 0.3$, $\mu_0 = 5.46$, $\sigma_0 = 0.7$, $\pi_{0,3} = 0.1$, $a_0 = 290$, $b_0 = 310$.

6.2 BMA Analysis

The true distribution of the message lengths was unknown in the simulation, and only data samples of size $n = 1000$ were observed. We considered three candidate input models for the (assumed independent) message lengths that commonly arise in probabilistic simulation studies—namely model M_1 was the Exponential(λ_1) distribution; model M_2 was the Normal(μ_2, σ_2) distribution; and model M_3 was the Lognormal(μ_3, σ_3) distribution:

$$\begin{aligned} p(x_i|M_1, \lambda_1) &= \lambda_1 e^{-\lambda_1 x_i} \quad (x_i \geq 0), \\ p(x_i|M_2, \mu_2, \sigma_2) &= \frac{e^{-(x_i-\mu_2)^2/(2\sigma_2^2)}}{\sqrt{2\pi}\sigma_2}, \\ p(x_i|M_3, \mu_3, \sigma_3) &= \frac{e^{-(\log x_i - \mu_3)^2/(2\sigma_3^2)}}{\sqrt{2\pi}\sigma_3 x_i} \quad (x_i \geq 0). \end{aligned}$$

We chose the above models because they appear often in simulation applications, and they are available in all input modeling and simulation software systems. Moreover, the posterior model probabilities for these candidate input models can be computed analytically; otherwise estimation of these probabilities may be time-consuming to do using MCMC methods for each data set. Note that these three models $\{M_k : k = 1, 2, 3\}$ are not nested, so that model comparison may not be conclusive in a classical framework. They also can represent very different behavior in terms of message lengths. Finally, we assume that we do not have any prior information to favor one model over the other and assign equal prior probabilities to all candidate models (i.e. $p(M_k) = 1/3$ for $k = 1, 2, 3$).

To construct proper priors from the standard noninformative priors on the model parameters, we generate a *training sample* (Berger and Perricchi 1996) of size $T = 100$ from the true sampling distribution (18). We denote by $\boldsymbol{z} = \{z_1, \ldots, z_T\}$ the observations in the training sample, and $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ the observations in the data sample.

For M_1, the standard noninformative prior is $p(\lambda_1|M_1) = 1/\lambda_1$. This yields a Gamma posterior distribution with shape parameter T and scale parameter $1/(\sum_{t=1}^{T} z_t)$. The marginal density of the data $\boldsymbol{x}$ given the exponential model M_1 is

$$p(\boldsymbol{x}|M_1) = \frac{\Gamma(n+T)}{\Gamma(T)} \frac{(\sum_{t=1}^{T} z_t)^T}{(\sum_{t=1}^{T} z_t + \sum_{i=1}^{n} x_i)^{n+T}}.$$

The standard noninformative prior for M_2 is $p(\mu_2, \sigma_2^2) = 1/\sigma_2^2$. This yields an inverse-gamma posterior distribution for σ_2^2 with shape parameter $(T-1)/2$ and scale parameter $\sum_{t=1}^{T}(z_t - \overline{z})^2/2$, and a generalized student-t distribution for μ_2, having the following density

$$p(\mu_2|\boldsymbol{z}, M_2) = \frac{\Gamma(T/2)\sqrt{T}\left(1 + \frac{T}{T-1}\left(\frac{\mu_2 - \overline{z}}{S_z}\right)^2\right)^{-T/2}}{\Gamma((T-1)/2)\sqrt{(T-1)\pi}S_z},$$

where $\overline{z} = \sum_{t=1}^{T} z_t/T$ and $S_z^2 = \sum_{t=1}^{T}(z_t - \overline{z})^2/(n-1)$. The marginal density of the data $\boldsymbol{x}$ given the normal model M_2 is given in Zouaoui and Wilson (2001b).

The analysis of the lognormal model is similar to that of the normal model, by working with the logarithm of the data observations instead of the original observations.

6.3 Simulation Design

A basic simulation run was of 50 seconds in length and this was repeated $m = 10$ times for each model and its sampled parameters. The sample sizes $\{R_k : k = 1, 2, 3\}$ generated from the posterior distribution of the parameters were taken to be 100. In practice larger values of R_k are recommended, typically 1000. However, $R_k = 100$ is sufficiently large in this case because the observed coverages are stable, indicating the satisfactory behavior of the method. In each case the experiment was repeated 200 times so that 200 confidence intervals were generated.

The "true" value β_0 of the average delay of a message in this communication network cannot be computed analytically. So we used a preliminary Monte Carlo experiment involving direct simulation of the network to estimate β_0 to within $\pm 0.05\%$ of its true value with 99% confidence (Law, Kelton, and Koenig 1981). The final estimate was found to be $\beta_0 = 0.006585$.

6.4 Simulation Results

Table 1 summarizes the results of the BMA analysis. For each candidate model M_k ($k = 1, 2, 3$), we show the average posterior probability $p(M_k|\boldsymbol{x})$ over all the Monte Carlo experiments. We also present for each model M_k the mean estimate $\widehat{\beta}_k$, the stochastic variance estimate $\widehat{\tau}_k^2$, and the parameter variance estimate $\widehat{\sigma}_k^2$ of the average delay of messages in the network. In terms of posterior probabilities, the exponential model is the least favorite, but we cannot really favor the lognormal model over the normal model. Note also that in terms of mean response estimates, the behavior of the lognormal model is completely different from the other two models. This is a situation where model uncertainty is the dominating uncertainty factor since it accounts for about 99% of the overall uncertainty, so that a simulation analyst can have a completely different response choosing a priori one model over the other.

Table 1: Posterior Probability, Mean and Variance Estimates for Each Candidate Model of Message Lengths in the Communication Network of Figure 2

Model	$p(M_k\|\boldsymbol{x})$	$\widehat{\beta}_k$	$\widehat{\tau}_k^2$	$\widehat{\sigma}_k^2$
Exp.	0.123	8.19E-03	6.20E-09	4.58E-08
Norm.	0.396	8.57E-03	7.17E-09	2.30E-08
Logn.	0.481	4.55E-03	1.08E-10	1.08E-11

To study the effect of model uncertainty, we analyzed three different approaches to input-model selection: classical frequentist, partial Bayes, and BMA. In the classical frequentist approach, we made the simulation runs at a fixed model and a fixed parameter estimated using MLE. In the partial Bayes approach, we fixed the model but we accounted for parameter uncertainty by resampling the parameters prior to each set of m simulation runs. Finally in the BMA approach, we used the algorithm of Figure 1 to account for both model and parameter uncertainty. For the BMA approach, we considered three replication allocation procedures. The first procedure allocates the same number of runs to each model; the second procedure uses the Proportional Allocation Procedure (PAP) given in (15); and the third procedure uses the Optimal Allocation Procedure (OAP) given in (13). To find the optimal allocations of the simulation runs to models, we used the final variance estimates of the equal allocation scheme and we limited our computing effort to the total number of replications in a single Monte Carlo experiment.

Table 2 shows the performance of the mean estimate of the message delay in the communication network in terms of the Absolute Percentage Error $100|\widehat{\beta} - \beta_0|/\beta_0$ and the Mean Squared Error $\mathrm{E}[(\widehat{\beta} - \beta_0)^2]$. As expected, the classical frequentist and partial Bayes approaches show almost similar performance because of the small number of unknown parameters in the network and our choice of noninformative priors. However, both of these approaches show extremely poor performance compared to the BMA approach. The mean estimate of the BMA approach is very close to the target mean, having less than 2% absolute percentage error and negligible mean squared error. The optimal allocation procedure delivered the most precise mean estimate, showing almost 50% reduction in mean squared error compared to the equal allocation procedure. However, the performance of the proportional allocation procedure was almost as good as the optimal one. This suggests that the proportional scheme may be more applicable in practice given its simplicity.

Table 2: Absolute Percentage Error (APE) and Mean Squared Error (MSE) for the Estimator of Average Message Delay in the Communication Network of Figure 2

Approach	Model	Mean	APE	MSE
Classical Frequentist	Exp.	8.18E-03	24.16	2.57E-06
	Norm.	8.57E-03	30.04	3.98E-06
	Logn.	4.55E-03	30.98	4.17E-06
Partial Bayes	Exp.	8.19E-03	24.35	2.61E-06
	Norm.	8.57E-03	30.12	4.00E-06
	Logn.	4.55E-03	30.86	4.17E-06
BMA	Mix.	6.60E-03	1.78	2.00E-08
BMA+PAP	Mix.	6.59E-03	1.35	1.20E-08
BMA+OAP	Mix.	6.59E-03	1.35	1.20E-08

In addition to point estimation, we studied the performance of the different approaches in terms of interval estimation. Table 3 summarizes the nominal 90% confidence interval lengths (CIL) and coverage probabilities (CP). Although the classical and partial Bayes approaches have much tighter confidence bands, they have zero coverage probabilities. This shows that their intervals are built around the wrong expected mean response. The BMA approach on the other hand has a much higher coverage probability, at a reasonable length, and centered at the right mean response. The replication allocation procedures deliver intervals with a higher coverage probability compared to the equal allocation BMA approach. This justifies the benefits of running more replications at models with higher posterior probability and smaller output response variance.

Table 3: Performance of Nominal 90% Confidence Interval for the Average Message Delay in the Communication Network of Figure 2

Approach	Model	CIL	CV(CIL)	CP
Classical Frequentist	Exp.	8.32E-05	2.62E-01	0%
	Norm.	9.40E-05	2.93E-01	0%
	Logn.	1.21E-05	2.29E-01	0%
Partial Bayes	Exp.	6.90E-04	1.19E-01	0%
	Norm.	4.89E-04	1.48E-01	0%
	Logn.	1.05E-05	8.32E-02	0%
BMA	Mix.	2.57E-04	1.31E-01	75%
BMA+PAP	Mix.	2.89E-04	1.16E-01	81%
BMA+OAP	Mix.	2.87E-04	1.13E-01	82%

7 CONCLUSIONS

In this paper we have proposed a framework based on a Bayesian Model Averaging (BMA) approach to account for model and parameter uncertainty as well as the conventional stochastic uncertainty in discrete-event stochastic simulation. Our computational experience with this approach has been very promising.

With the recent advances in Bayesian computations, we believe that the introduction of Bayesian techniques into routine practice is a much more attainable goal than previously thought. However, much more work remains to be done in a number of areas to reach this goal. Correlated sampling schemes for improving the efficiency of the Simulation Replication Algorithm needs further exploration. A more comprehensive experimental performance evaluation and the implementation of a user-friendly software tool are also important steps for the widespread use of the Bayesian techniques in simulation input modeling.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation under grant number DMI-9900164. The authors thank Bibhuti B. Bhattacharyya, Sujit K. Ghosh, and Stephen D. Roberts for their enlightening discussions on this paper.

REFERENCES

Berger, J. O., and L. R. Perrichi. 1996. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* 91 (433): 109–122.

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.

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, New Jersey: Institute of Electrical and Electronics Engineers. Available on-line via `http://www.informs-cs.org/wsc99papers/044.PDF` [accessed June 18, 2000].

Draper, D. 1995. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society* B57 (1): 45–97.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 1995. *Bayesian data analysis*. London: Chapman & Hall.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. Bayesian model averaging: A tutorial. Technical Report 9814, Department of Statistics, Colorado State University, Fort Collins, Colorado.

Kleinrock, L. 1976. *Queueing systems, volume 2: Computer applications*. New York: John Wiley & Sons.

Law, A. M., W. D. Kelton, and L. W. Koenig. 1981. Relative width sequential confidence intervals for the mean. *Communications in Statistics: Simulation and Computation* B10 (1): 29–39.

Madigan, D., and A. E. Raftery. 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* 89 (428): 1535–1546.

Rao, P. 1997. *Variance components estimation: Mixed models, methodologies and applications*. London: Chapman & Hall.

Satterthwaite, F. E. 1946. An approximate distribution of estimates of variance components. *Biometrics* 2:110–114.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6:461–464.

Spiegelhalter, D., A. Thomas, N. Best, and W. Gilks. 1996. *BUGS 0.5: Bayesian inference using Gibbs sampling manual (Version ii)*. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge, UK.

Stuart, A., and J. K. Ord. 1994. *Kendall's advanced theory of statistics, volume 1: Distribution theory*. 6th ed. London: Edward Arnold.

Zouaoui, F. 2001. Accounting for input uncertainty in discrete-event simulation. Doctoral dissertation, Graduate Program in Operations Research, North Carolina State University, Raleigh, NC. Available on-line via `http://www.lib.ncsu.edu/etd/public/etd-3949188410121271/etd.pdf` [accessed June 16, 2001].

Zouaoui, F., and J. R. Wilson. 2001a. Accounting for parameter uncertainty in simulation input modeling. In *Proceedings of the 2001 Winter Simulation Conference*, ed. B. A. Peters, J. S. Smith, M. W. Rohrer, and D. J. Medeiros, to appear. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available on-line with a more detailed development via `ftp://ftp.ncsu.edu/pub/eos/pub/jwilson/msc1.pdf` [accessed June 16, 2001].

Zouaoui, F., and J. R. Wilson. 2001b. Accounting for input model and parameter uncertainty in simulation. Technical Report, Department of Industrial Engineering, North Carolina State University, Raleigh, North Carolina. Available on-line via `ftp://ftp.ncsu.edu/pub/eos/pub/jwilson/msc2.pdf` [accessed June 16, 2001].

AUTHOR BIOGRAPHIES

FAKER ZOUAOUI is an operations research consultant in the Research Group at Sabre, Inc. He received his bachelor's and master's degrees in industrial engineering from Bilkent University, and he received his Ph.D. degree in operations research from North Carolina State University. He is a member of INFORMS. His e-mail address is <faker.zouaoui@sabre.com>.

JAMES R. WILSON is Professor and Head of the Department of Industrial Engineering at North Carolina State University. He currently serves as the corepresentative of INFORMS–College on Simulation to the Board of Directors of the Winter Simulation Conference. He is a member of ASA, ACM, IIE, and INFORMS. His e-mail address is <jwilson@eos.ncsu.edu>, and his web address is <www.ie.ncsu.edu/jwilson>.