Proceedings of the 1996 Winter Simulation Conference
ed. J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain

SELECTION OF INPUT MODELS USING BOOTSTRAP GOODNESS-OF-FIT

Russell C. H. Cheng
Wayne Holland
and
Neil A. Hughes

Institute of Mathematics and Statistics
The University of Kent at Canterbury
Canterbury, Kent CT2 7NF
ENGLAND

ABSTRACT

Bootstrap methods are a natural adjunct of computer simulation experiments; both use resampling techniques to construct the statistical distributions of quantities of interest. In this paper we consider how bootstrap methods can be used in selecting appropriate input models for use in a computer simulation experiment. The proposed method uses a goodness-of-fit statistic to decide on which of several competing input models should be used. We use bootstrapping to find the distribution of the test statistic under different assumptions as to which model is the correct fit. This allows the quality of fit of the different models to be compared.

The bootstrapping process can be extended to the simulation experiment itself, allowing the effect of variability of estimated parameters on the simulation output to be assessed.

The methodology is described and illustrated by application to a queueing example investigating the delays experienced by motorists caused by toll booths at a bridge river crossing.

1 INTRODUCTION

We consider the selection of input models in computer simulation experiments. It is supposed that samples of real data sets exist drawn from input distributions needed in the simulation, but that there is uncertainty concerning the underlying form of these distributions. We suppose however that we are able to draw up a list, for each data set, of candidate distributions that they are likely to have been obtained from. The main task is therefore to fit these possible distributions to the data, and based on the quality of fit, to decide on which distribution is the best fit. In addition to selecting input models, an assessment has to be made concerning the adequacy of the selected models.

Goodness-of-fit procedures are well-known in the statistical literature. A good review of their use in input model selection is discussed by Law and Kelton (1991). See also Banks et al. (1984). A difficulty that arises is that many of the most sensitive tests of fit are based on statistics whose distribution is dependent on both the model being considered and also on whether parameters are estimated or not. These distributions are usually hard to obtain in general. Thus a goodness-of-fit test is limited to only those models for which its own distribution is actually known. In this paper we show that the parametric bootstrap provides a convenient way of applying goodness-of-fit for selecting input models that overcomes this difficulty. We give a method of doing this and illustrate it with an example involving a queueing system.

An important aspect is that the variability of the final simulation output will be influenced by uncertainty in the input model fitting process as well as by uncertainty arising from the (pseudo) random nature of the simulation itself. Both these source of variation need to be taken into account in assessing the overall variability of the simulation output. For a good review of sensitivity analysis see Kleijnen (1995). Cheng and Holland (1995, 1996) have considered this problem using bootstrap techniques to measure the variability of the simulation output arising from estimating unknown parameters when input models are being fitted. The present paper is thus an extension of their technique to include selection as well as fitting of appropriate input models.

For related work see Swain et al. (1988), Barton and Schruben (1993), and Shanker and Kelton (1994).

2 INPUT MODEL SELECTION

We use a framework which highlights the input models used in the simulation. We assume that the simulation uses k univariate input distributions (or models), with distribution functions

$$F_i(x, \theta_i) \;\; i = 1, 2, ..., k \tag{1}$$

where

$$\theta_i = (\theta_{1i}, \theta_{2i}, ..., \theta_{p_i i}) \quad i = 1, 2, ..., k$$

is the vector of p_i unknown parameters on which the ith distribution depends. We assume that each model is selected separately from the others, with the same method of selection used for each model. We can therefore, for simplicity, drop the subscript and denote the typical model being considered by $F(x, \theta)$. We suppose that $F(x, \theta)$ is the unknown true distribution, and that we have narrowed down F to one of m possibilities:

$$G_j(x, \theta_j) \quad j = 1, 2, ..., m$$

In addition we assume that there is available a sample of empirical data for each model:

$$\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{in_i}) \; i = 1, 2, ..., k.$$

The combined samples will be denoted by

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k).$$

Again for simplicity we suppress the subscript and write

$$\mathbf{x} = (x_1, x_2, ..., x_n).$$

for a sample drawn from a typical $F(x, \theta)$. The initial problem is therefore to use $\mathbf{x}$ to to assess the goodness-of-fit of the $G_j's$ and to select one of the G_j to represent F.

Our proposed method is to fit each G_j to the sample $\mathbf{x}$, then assess the goodness-of-fit by calculating the value, T_j, of a goodness-of-fit test statistic, T, under the assumption that G_j is the correct model. In general the distribution of T will depend on the model being fitted and also on whether the parameters have been estimated or not. There is therefore no guarantee that goodness-of-fit of the models will follow the ranked order of the T_j values. We do not therefore compare the T_j values directly. Instead we calculate the probability value of each T_j; i.e.

$$p_j = \Pr\{T \leq T_j | \, A_j\}$$

where A_j is the assumption that G_j is the correct model, that it has been used to estimate θ_j and that $G_j(., \hat{\theta}_j)$ has been used to compute T_j. The model deemed to be the best fit corresponds to that with the smallest p_j-value.

The fact that the distribution of T is altered if parameters are fitted is a difficulty of some awkwardness, as different significance levels are needed for different distributions (See D'Agostino and Stephens, 1986). However the difficulty is overcome by use of the bootstrap to construct an approximation of the required distribution for each T_j. Basically the method is to draw a number of so-called bootstrap samples from $G_j(., \hat{\theta}_j)$. An estimate of θ_j and a corresponding T_j is calculated for each such sample. Then the empirical distribution function (EDF) of these T_j's estimates the true distribution of T_j. For a discussion of bootstrap sampling methods, see Efron (1979, 1987) and Efron and Tibshirani (1994). We use the following result.

Theorem 1. Suppose that $\mathbf{x}$ is a random sample of size n drawn from the distribution with CDF $G(x, \theta)$ and that

$$T = T(\mathbf{x}, \; G(., \theta))$$

is a goodness-of-fit statistic dependent on $\mathbf{x}$ and on $G(., \theta)$. Suppose that (i) the CDF, $F_T(t, \theta)$, of T is a continuous function of θ for each t, (ii) $\hat{\theta}$ is a consistent estimator of θ, calculated from $\mathbf{x}$. Then the distribution of $T^* = T(\mathbf{x}^*, \; G(., \hat{\theta}))$, where $\mathbf{x}^*$ is a random sample of size n drawn from $G(., \hat{\theta})$, tends to that of T, in probability, as $n \to \infty$.

Proof. The distribution of T^* is simply that of T when calculated at $\hat{\theta}$. By assumption (ii) $\hat{\theta}$ is a consistent estimator, that is $\hat{\theta} \to \theta$ in probability. Hence, by (i), $F_{T^*}(t, \hat{\theta}) \to F_T(t, \theta)$ for each t in probability, i.e. the distribution of T^* tends to that of T in probability. □

Theorem 1 shows that, providing n is sufficiently large, we can approximate the distribution of T by bootstrap sampling of T^* from the fitted distribution $G(., \hat{\theta}_j)$. Our model selection method is thus as follows.

For each of the possible models, $j = 1, 2, ..., m$:

1. Fit the model G_j by estimating θ_j from the sample $\mathbf{x}$. Let the estimator be $\hat{\theta}_j$.

2. Calculate an appropriate goodness-of-fit statistic T_j for the fitted model, $G_j(., \hat{\theta}_j)$.

3. Use bootstrap sampling to estimate the distribution of T_j:

   (a) Generate B bootstrap samples $\mathbf{x}^{(i)}$, $i = 1$, 2, ..., B from the fitted model $G_j(., \hat{\theta}_j)$.

   (b) For each sample, $\mathbf{x}^{(i)}$, fit $G_j(x, \theta_j)$ by estimating θ_j, giving estimates $\hat{\theta}_j^{(i)}$ $i = 1$, 2, ..., B.

   (c) Calculate the goodness-of-fit statistic $T_j^{(i)}$ for the fitted model $G_j(., \hat{\theta}_j^{(i)})$ for $i = 1$, 2, ..., B.

(d) Form the EDF of the $T_j^{(i)}$ and hence find the p-value of T_j :

$$p_j = \frac{\#\ of\ T_j^{(i)} \leq T_j}{B}.$$

The model G_j with the smallest p_j value is selected as being the best fit.

Table 1: Service Times (secs) for PCs, LVs and HGs

PC			LV		HG	
5.7	10.9	4.9	4.3	10.9	8.6	7.2
5.4	5.7	4.3	4.7	3.1	32.0	7.7
5.7	6.3	6.9	6.7	4.5	8.7	6.9
3.8	4.9	8.5	7.2	6.6	12.1	8.7
5.3	7.8	5.2	6.3	4.7	25.5	5.1
5.1	7.1	4.4	6.2	4.2	16.7	7.1
7.0	4.9	7.6	4.8	3.3	7.1	9.3
10.5	5.5	7.6	6.3	4.0	37.1	6.4
7.6	5.1	3.5	3.5	7.8	4.6	8.1
4.3	10.1	12.0	5.7	5.8	8.8	11.8
3.2	6.5	3.0	5.2	8.0	22.9	30.4
15.1	3.2	5.9	4.9	6.1	11.0	8.3
7.4	4.4	3.8	7.7	4.3	8.9	6.9
5.6	6.2	5.7	7.9	3.9	8.1	8.3
4.1	6.9	7.2	4.4	6.7	8.7	4.1
5.1	5.1	5.9	6.4	7.2	7.7	12.5
2.7	4.9	6.1	4.7	5.0	8.2	10.6
4.2	4.4	5.6	5.2	6.4	8.8	5.9
2.9	13.1	4.9	3.6	10.5	11.0	6.1
5.7	2.7	6.8	5.8	8.0	11.3	8.1
6.1	3.0	3.2	8.2	12.5	7.4	10.5
5.0	3.8	4.8	4.1	4.0	8.1	14.3
4.2	5.9	7.1	4.6	3.8	7.2	9.7
3.2	5.9	4.5	3.1		7.1	5.0
5.2	3.4	5.9			6.8	12.5
9.2	3.7	3.4			7.7	11.5
4.9	4.6	3.6			7.4	13.4
5.1	3.3	4.9			12.9	9.7
3.9	5.2	5.9			13.4	10.1
4.1	4.6	6.7			9.4	10.3
5.2	6.5	6.1			7.6	18.0
15.6	5.7	5.7			11.8	21.6
5.7	4.2	3.7			9.4	4.5
8.2	9.0	4.7			12.5	6.8
5.3	8.2	6.9			12.2	11.2
6.6	4.6	5.8			8.2	13.0
3.9	3.6	6.4			6.7	25.8
4.7	4.4	7.4			9.4	5.2
					12.6	

3 GOODNESS-OF-FIT STATISTICS

We consider examples of the method of model selection proposed in the previous section. Table 1 gives data of the times in seconds to serve three classes of vehicle: Private Cars (PC), Light Vans (LV) and Heavy Goods Vehicles (HG) at the toll booths of the Severn Bridge River crossing in Britain. This example will be used later to illustrate the complete procedure suggested for carrying out a simulation experiment. Here we consider simply the model selection procedure. There is some evidence that an appropriate distribution should allow for a minimum processing time to handle a vehicle. We thus consider a three parameter versions of the Weibull, gamma and lognormal distributions as possible candidates. An interesting aspect of fitting these models is that they all possess embedded two-parameter special cases, and this possibility must be allowed for; see Cheng and Iles (1990) for a discussion of how this should be done. In the HG data set there is evidence that the service times are a mixture of distributions. We therefore consider, in addition, a five parameter mixture model of two normal models with different means and variances, with an unknown mixing proportion. We use maximum likelihood estimation for the parameters. There are many possible goodness-of-fit test statistics available. For illustration we use the Anderson-Darling test statistic which is known to be particularly sensitive and powerful for detecting differences in tail behaviour. If the ordered sample is $x_1 < x_2 < \ldots < x_n$, and the model under consideration is $G(.,\theta)$ then, writing $Z_i = G(x_i, \hat{\theta})$, the test statistic is

$$A^2 = -n - n^{-1} \sum_{i=1}^{n} (2i-1)\{\ln(Z_i) + \ln(1 - Z_{n-i+1})\} \tag{2}$$

Table 2 gives the values of A^2 and their associated $p - values$ for each of the data sets and each model.

It will be seen that the best fit is the mixture model for the PC and HG data sets, and the lognormal model for the LV data set.

It is of interest to see how sensitive the test is in distinguishing between models. To illustrate this we consider the lognormal model and mixture models

Table 2: Anderson-Darling Statistic and p-Values

Model		PC Data	LV Data	HG Data
Weibull	A^2	1.269	0.277	2.282
	$P(A^2)$	0.999	0.385	> 0.999
Lognormal	A^2	0.544	0.227	1.041
	$P(A^2)$	0.902	0.229	0.998
Gamma	A^2	0.798	0.238	1.833
	$P(A^2)$	0.968	0.237	> 0.999
Mixture	A^2	0.256	0.541	0.423
	$P(A^2)$	0.544	0.897	0.870

fitted to the HG data. Figure 1 shows ten separate EDF's of A^2, where each EDF is constructed as follows. The fitted lognormal model is treated as the stage-1 model and a bootstrap sample of the same size as the original sample, is obtained from it. The lognormal model (stage-2) is fitted to this bootstrap sample. 1000 secondary bootstrap samples, each of size 20, are then drawn from this stage-2 model, and the lognormal model (stage-3) is then fitted to each sample. The corresponding A^2 is calculated as in (2) for this stage-3 fit. The variability between EDF's is due to the bootstrap sampling of the stage-1 model, and gives an indication of the variabiity due to using fitted parameters instead of the unknown "true" parameter values. For comparison Figure 1 also shows ten EDF's each formed from 1000 A^2 values calculated in exactly the same way except that the lognormal model of stages 1 and 2 is replaced by the normal mixture model; however the lognormal is still the fitted model at stage 3, and each A^2 is calculated for this fitted lognormal model. These latter EDF's therefore indicate how the distribution of A^2 when the incorrect (lognormal) model is fitted to data drawn from the mixture model. There is significantly more variability in the EDF when the incorrect model is fitted. This is due to the nature of the mixture distribution with its far greater inherent variability. With data generated from simpler alternative distributions like the gamma, there should be considerably less variation, though this has yet to be investigated. Figure 1 does however give an indication of the reasonable power of the test in rejecting the lognormal model when the alternative mixture model is actually the correct one.

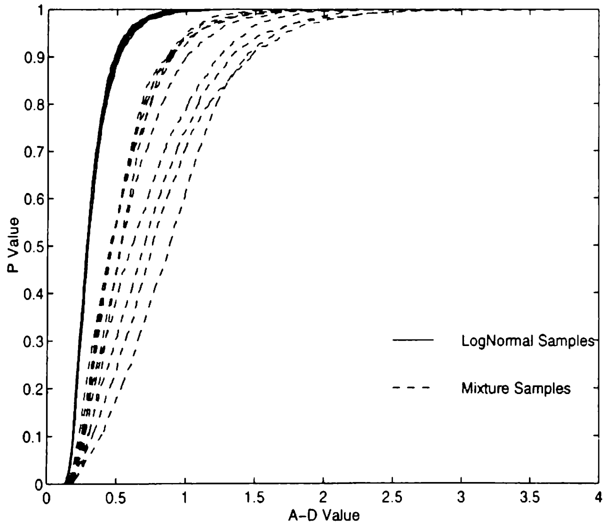

Figure 1: Comparison of A^2 EDFs for Lognormal Model Fitted to Lognormal and Mixture Samples

4 SENSITIVITY ANALYSIS

Cheng and Holland (1995, 1996) show how bootstrap sampling can be applied to the simulation experiment itself. We follow their terminology. The simulation study is assumed to consist of making a number of runs of a computer simulation model and observing the output of interest, y, from each run. Let the length of each run be l (measured in simulation time, say). In each run, the input models (1) are used to generate k streams of random variate. The variates used in one simulation run will be denoted by

$$\xi_i = (\xi_{i1}, \xi_{i2}, ..., \xi_{im_i}) \; i = 1, 2, ..., k.$$

Each ξ_i is assumed to be a random sample with the individual observations, ξ_{ij}. The number of variates, m_i, used is different in each stream. In what follows the m_i can be variable, but to simplify the discussion the run length, l, can be regarded as fixed, with the m_i also fixed. The F_i are assumed to have been selected using the method of the previous section.

Though they may be generated in various ways, it will be convenient to regard the input variates as generated by the inverse transform method (see Law and Kelton, 1991, for example):

$$\xi_{ij} = F_i^{-1}(u_{ij}, \theta), \quad j = 1, 2, ..., m_i$$

where F^{-1} is the inverse of F, and the u_{ij} are independent uniform $U(0, 1)$ variates. We write

$$\mathbf{u}_i = (u_{i1}, u_{i2}, ..., u_{im_i}) \; i = 1, 2, ..., k$$

and

$$\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_k).$$

Helton (1993, 1994) distinguishes two types of uncertainty: *parameter uncertainty* and *simulation uncertainty*. The parameter uncertainty is due to the θ_j

being unknown and having to be estimated, whilst the simulation uncertainty occurs because of the randomness of $\mathbf{U}$. Thus the output of interest from the simulation run, y, has random variability because it depends on $\mathbf{U}$, and θ:

$$y = y(\mathbf{U}, \theta).$$

Let

$$\eta(\theta) = E(y, \theta) = \int y(\mathbf{U}, \theta)\mathbf{dU}$$

be the expected value of y, and let θ^0 denote the true parameter value. We assume that the objective is to estimate $\eta(\theta^0)$ in the case where θ^0 is not known.

We consider the overall simulation experiment as being made up of r runs. Assuming, for the moment, that the runs are made at some given θ, the responses or outputs from these runs can be written as:

$$y_j(\mathbf{U}_j, \theta) = \eta(\theta) + e_j(\mathbf{U}_j, \theta), \ j = 1, 2, ..., r. \tag{3}$$

The 'error' variable e_j is the random difference between the jth simulation run output and $\eta(\theta)$. We shall assume $E(e_j|\theta) = 0$ and $Var(e_j|\theta) = \tau^2(\theta)/l$ for $j = 1, 2, ..., r$. Thus

$$E[y_j(\mathbf{U}_j, \theta)|\theta] = \eta(\theta),$$

and the mean of the outputs

$$\bar{y} = \sum_{j=1}^{r} y_j(\mathbf{U}_j, \theta)/r,$$

is an unbiased estimator of $\eta(\theta)$ with

$$Var(\bar{y}|\theta) = \tau^2(\theta)/rl. \tag{4}$$

This variance measures the simulation uncertainty within the model. It is the total variance of the response only when θ is fixed and known.

When θ is unknown, it has to be estimated. If we have estimates $\hat{\theta}_j$ that are $\sqrt{n}$-consistent (where n is the typical size of real data samples), as would be the case if we use maximum likelihood estimation, then the total variation of $\bar{y}$ is given by

$$\begin{aligned} Var(\bar{y}) &= \underset{\hat{\theta}}{Var}\,(\underset{y}{E}\,(\bar{y}|\hat{\theta})) + \underset{\hat{\theta}}{E}\,(\underset{y}{Var}\,(\bar{y}|\hat{\theta})) \\ &= \sigma^2(\theta^0)/n + \tau^2(\theta^0)/rl \ + \ R \end{aligned} \tag{5}$$

where θ^0 is the unknown true parameter value, $\sigma^2(\theta^0)/n$ is the parameter uncertainty, $\tau^2(\theta^0)/rl$ is the simulation certainty, and the remainder term, R, involves only terms of order $O(n^{-3/2})$ and $O[(nrl)^{-1}]$.

Cheng and Holland (1996) show that the variance for $\bar{y}$ given $\hat{\theta}$ is:

$$Var(\bar{y}|\hat{\theta}) = \tau^2(\hat{\theta})/(rl) = [\tau^2(\theta^0) + O_p(n^{-1/2})]/(rl),$$

This means that we can use

$$t_0^2 = \sum_{j=1}^{r} (y_j - \bar{y})^2/(r-1), \tag{6}$$

to estimate the stochastic uncertainty, $\tau^2(\theta^0)/l$.

Assessing the parameter uncertainty is more difficult. Cheng and Holland give two methods for doing this. We consider here the bootstrap method that they propose.

A sample $\mathbf{X}^1 = (\mathbf{x}_1^1, \mathbf{x}_2^1, ..., \mathbf{x}_k^1)$ is obtained with the $\mathbf{x}_i^1$ drawn from the selected distributions $G_j(x, \hat{\theta})$ $(j = 1, ..., k)$. This sample, $\mathbf{X}^1$, can be used to produce an estimate $\hat{\theta}^1$,in exactly the same way that $\hat{\theta}$ was obtained from $\mathbf{X}$. Repeating this B times yields B such estimates: $\hat{\theta}^1$, $\hat{\theta}^2$, ..., $\hat{\theta}^B$. We can then carry out B bootstrap simulation experiments, one for each $\hat{\theta}^i$, with all runs of length l as in the original experiment, but with the bootstrap experiment containing r' runs, where r' may be different from r. (The run lengths can be different from l but there is actually no loss of generality in assuming them to be the same.) This yields B sets of responses, with r' responses in each set:

$$y_1^i, y_2^i, ..., y_{r'}^i, \quad i = 1, 2, ..., B. \tag{7}$$

Let the means of each set be

$$\bar{y}^i, \quad i = 1, 2, ..., B. \tag{8}$$

Cheng and Holland (1996) show how the variability of these $y's$ depends on both the simulation variance τ^2 and on the parameter variance σ^2.

Theorem 2.

$$Var(\bar{y}^i|\hat{\theta}) = \sigma^2(\theta^0)/n + \tau^2(\theta^0)/r'l + R. \tag{9}$$

where the remainder, R, is of order

$$R = O_p(n^{-3/2}) + O_p[(n^{1/2}r'l)^{-1}].$$

Proof: See Cheng and Holland (1996). □

An estimate of this variance, (9), is the sample variance of the $\{\bar{y}^i\}$ (8):

$$s_B^2 = \sum_{i=1}^{B} (\bar{y}^i - \bar{\bar{y}})^2/(B-1). \tag{10}$$

As t_0^2 (6) estimates $\tau^2(\theta^0)/l$, this can then used to adjust s_B^2 to give the following estimate of the total variance of $\bar{y}$ (5):

$$\widehat{Var}(\bar{y}) = s_B^2 + (\frac{1}{r} - \frac{1}{r'})t_0^2. \qquad (11)$$

When $r' = r$, i.e. when each bootstrap is an exact replica of the original experiment, then $r = r'$, and the total variance is estimated by s_B^2 on its own.

5 APPLICATION TO A QUEUEING SYSTEM

To illustrate the applicability of the above, we consider the simulation of a queueing situation, investigating the delays experienced by motorists caused by toll booths at a bridge river crossing. The bridge considered crosses the River Severn in the United Kingdom. The bridge is approached by the M4 motorway, consisting of three carriageways. The approach widens to accommodate eight toll booths.

Data for service times at toll booths for three types of vehicles - private cars (PC), light vans (LV) and heavy goods vehicles (HG) have already been given in Table 1. Preliminary investigation of the three data sets indicated that the service time distribution was different for each type of vehicle. The mean and standard deviation of service time (in seconds) was 5.73 and 2.2 for PCs, 5.80 and 2.03 for LVs and 10.88 and 6.20 for HGs. These were based on 114, 47 and 77 observations respectively. No data was available for inter-arrival times. However, Griffiths and Williams (1984), in their study of the Severn Bridge, conclude that it may be reasonably assumed that inter-arrival times follow a negative exponential distribution for each type of vehicle. Further, they state that past records indicate that the composition of traffic types is 75.4% PCs, 19.4% HGs and 5.4% LVs. It was further predicted that traffic flow by 1995 would reach 51,051 vehicles per day. This leads to estimates for mean arrival rates per hour of 1604 for PCs, 413 for HGs and 111 for LVs.

Within the simulation, vehicles are assumed to choose an approach lane at random, unless there are queues in which case the carriageway with the shortest queue is selected. Driver choice of toll booth is selected by conditional probability distribution (conditional upon the lane of approach). The actual probabilities used are given in Table 3. Thus three probability distributions are constructed for choice of toll booth (out of eight) dependent upon lane of approach (choice of three). These distributions are then modified by dividing by the queue length at each booth, and then normalised so that the probabilities summed to unity. This means that each driver has an individual probability distribution for choice of booth dependent on lane of approach and queue length at each toll booth at the moment of selection. The maximum queue that can develop at a toll booth is ten, at which point the traffic commences queueing on the approach road. Once vehicles are queueing it is assumed that PCs and LVs will take two seconds to move up one place in the queue; HGs are assumed to take four seconds.

Table 3: Distribution of Driver Choice of Toll Booth

Lane	1	2	3
Booth			
1	0.20	0.05	0.01
2	0.25	0.25	0.09
3	0.25	0.25	0.10
4	0.10	0.15	0.15
5	0.09	0.12	0.15
6	0.07	0.08	0.15
7	0.03	0.06	0.15
8	0.01	0.04	0.20

The simulation procedure may be outlined in the following manner. The best distributions are fitted to the service time data sets and a simulation run of length l = 36,000 simulation seconds is repeated r = 250 times. The output from the simulation experiment consists of the mean delay for PCs, LVs, HGs and a combined overall mean for each simulation run. Considering just the overall mean for a moment, these may be thought of as (y_1, y_2, ..., y_{250}) from (3). These may be used to obtain an estimate of the simulation variance (4). Following this, bootstrap samples are generated from each of the three specified service time distributions. Bootstrapping is not performed on the inter-arrival time distributions because we are assuming that they are known to be negative exponential. As discussed in the previous section, the purpose of bootstrapping here is to obtain estimates of variability caused by fitting distributions where there is uncertainty. Each of the bootstrap samples is used to fit parameter estimates for the appropriate distribution and the simulation then proceeds with these estimates. This time, a single simulation run is performed ($r' = 1$). This is of length 36,000 simulation seconds, as before. The reason for this is that we have our estimate of simulation variability from the first experiment. We are now interested in variability between bootstrap experiments to give us the variability caused by unknown input parameters. Thus this bootstrapping procedure is repeated B = 250

Table 4: Results for Toll Booth Experiment

	Mean Delay	**Para Var**	**Sim Var**	**Total Var**
Best Fit	15.66	1.0400	2.5 x 10^{-4}	1.0402
Worst Fit	15.70	1.0532	2.1 x 10^{-4}	1.0532
Results for PCs				

	Mean Delay	**Para Var**	**Sim Var**	**Total Var**
Best Fit	15.78	0.9428	3.2 x 10^{-4}	0.9460
Worst Fit	15.78	0.9581	3.2 x 10^{-4}	0.9584
Results for LVs				

	Mean Delay	**Para Var**	**Sim Var**	**Total Var**
Best Fit	24.56	2.2739	1.1 x 10^{-3}	2.2751
Worst Fit	24.74	1.8377	9.2 x 10^{-4}	1.8386
Results for HGs				

	Mean Delay	**Para Var**	**Sim Var**	**Total Var**
Best Fit	16.14	1.0284	2.6 x 10^{-4}	1.0287
Worst Fit	16.18	1.0258	2.3 x 10^{-4}	1.0260
Results for All Vehicles Combined				

times. The output provided by each bootstrap simulation is the mean delay for PCs, LVs, HGs and the overall mean delay. For mean overall delay, say, we may regard the output as that described in (7). The overall variance (5) may be calculated using (10) and (11). Table 4 summarises the results for delays to PCs, LVs, HGs and overall delay respectively, when the best distributions have been used, and for comparison when the worst distributions have been used. It can be seen that the choice of distribution has little effect on the results, except for the the parameter variance for the HGs. This data was unusual and was not particularly well fitted by any of the distributions considered. However, there is clearly some sensitivity to the choice of input distribution. In all other cases, the observed data could be reasonably fitted by a number of the distributions. Therefore, less sensitivity would be expected in such cases. The lack of sensitivity to choice of input model is heightened by two further points. Firstly, we were only able to conduct the exercise for service time distributions. The inter-arrival time distributions were taken as fixed, yet optimal selection of these distributions could contribute considerably to the variability in the context of a queueing analysis. Secondly, we have chosen here to calculate only mean measures. It would be sensible to consider the calculation of some probability measures, for instance the probability of queueing occurring on the approach road to the toll booths or the probability of a toll booth being idle. Once would expect such measures to be much more sensitive to input distribution and this will be the subject of further investigation.

The mean delay, $\bar{y}$, has been calculated here from the simulation runs at the fitted parameters only. The bootstrap results have been used only in the variance estimation. It is possible to incorporate the results for the mean from the bootstrap simulation runs with the main simulation result to produce an estimator with lowest overall variance. Details of the procedure may be found in Cheng and Holland (1995).

REFERENCES

Banks, J., Carson II, J.S. and Nelson, B.L. (1984). *Discrete-Event Simulation, (2nd Edn).* Upper Saddle River, NJ: Prentice Hall.

Barton, R.R. and Schruben, L.W. (1993). Uniform and Bootstrap Resampling of Empirical Distributions. In *Proceedings of the 1993 Winter Simulation Conference* (ed. G.W. Evans, M. Mollaghasemi, E.C. Russell and W.E. Biles), IEEE Piscataway, New Jersey, 503-508.

Cheng, R.C.H. and Holland, W. (1995). The Effect of Input Parameters on the Variability of Simulation Output. *Proceedings of the Second United Kingdom Simulation Society Conference (*Eds R.C.H. Cheng and R.J. Pooley), North Berwick, April 1995, Edinburgh University, pp 29-36.

Cheng, R.C.H. and Holland, W. (1996). Sensitivity of Computer Simulation Experiments to Errors in Input Data. To appear in *J. of Statistical Computation and Simulation.*

Cheng, R.C.H. and Iles, T.C. (1990). Embedded Models in Three-Parameter Distributions and their Estimation. *J. R. Statist. Soc.* B, **52**, 135-149.

D'Agostino, R.B. and Stephens, M.A. (1986). *Goodness of Fit Techniques.* New York: Marcel Dekker.

Efron B. (1979). Bootstrap Methods : Another Look at the Jackknife. *The Annals of Statistics*, **7**, pp 1-26.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Vol. **38** of CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM.

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

Griffiths, J.D. and Williams J.E. (1984). Traffic Studies on the Severn Bridge. *Traffic Engineering and Control*, **25**, pp 268-71, 274

Helton, J.C. (1993). Uncertainty and Sensitivity Analysis Techniques for Use in Performance Assessment for Radioactive Waste Disposal. *Reliability Engineering and System Safety,* **42**, 327-367.

Helton, J.C. (1994). Treatment of Uncertainty in Performance Assessments for Complex Systems. *Risk Analysis,* **14**, 483-511.

Kendall, M.G. and Stuart, A. (1979). *The Advanced Theory of Statistics. Vol. 2: 4th Edn.* London: Griffin.

Kleijnen J.P.C. (1995). Sensitivity Analysis and Related Analyses : a Survey of Statistical Techniques (submitted for publication).

Law, A.M. and Kelton, W.D. (1991). *Simulation Modeling and Analysis 2nd Edition.* New York: McGraw-Hill.

Shanker, A. and Kelton, W. D. (1994). Measuring Output Error due to Input Error in Simulation: Analysis of Fitted vs. Mixed Empirical Distributions for Queues. To appear.

Swain, J.J., Venkatraman, S. and Wilson, J.R. (1988). Distribution Selection and Validation. *J. of Statist. Comput. and Simul.*, **29**, 271-297.

AUTHOR BIOGRAPHIES

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

WAYNE HOLLAND is a lecturer in Operational Research in the Institute of Mathematics and Statistics at the University of Kent at Canterbury. He has a B.Sc. in Mathematics and its Applications from the University of Wales. He also obtained his Ph.D. from the University of Wales. He is a member of the Operational Research Society. His research interests include numerical approximation of transient queueing measures, analysis of queueing networks with particular application to transportation and computer communication systems.

NEIL HUGHES is a Research Assistant in the Institute of Mathematics and Statistics at the University of Kent at Canterbury. He has a B.Sc. in Mathematics and its Applications from the University of Wales, and is preparing his doctoral dissertation on the automatic generation of terrain databases for use in computer generated imagery.