Proceedings of the 1986 Winter Simulation Conference
J. Wilson, J. Henriksen, S. Roberts (eds.)

MODELING INPUT PROCESSES

Ronald L. Iman
Safety and Environmental Studies
Division 6415
Sandia National Laboratories
Albuquerque, NM 87185

ABSTRACT

Computer models for various applications are closely scrutinized both from the standpoint of questioning the correctness of the underlying mathematical model with respect to the process it is attempting to model, and from the standpoint of verifying that the computer model correctly implements the underlying mathematical model. A process that receives less scrutiny, but is none the less of equal importance, concerns the individual and joint modeling of the inputs. This modeling effort clearly has a great impact on the credibility of results obtained from simulation studies. Model characteristics are reviewed that have a direct bearing on the model input process and reasons are given for using probabilistic based modeling with the inputs. Discussions are presented on how to model distributions for individual inputs and how to model multivariate input structures when dependence and other constraints may be present.

1. INTRODUCTION

The selection of appropriate models to represent the inputs utilized in simulation studies with complex computer models usually receives far less attention to detail than does the actual construction of the computer model itself. This is usually related to the fact that the modeler has spent a great deal of time and effort in developing the underlying mathematical model and in verifying the computer model that implements the mathematical model. Once developed and tested, it is then a simple matter to run the computer model. This latter viewpoint is reasonably correct if the analyst's objective is to perform a first cut screening analysis on the inputs to the computer model. However, if the analyst has objectives that are more statistical in nature, such as estimating quantiles, or the mean and variance of the computer model output variable, careful attention must be paid to modeling the multivariate structure of the input with respect to the appropriateness and validity of distribution assumptions and the degree of dependence that may be required among the inputs.

In this paper the characteristics of computer models are briefly reviewed with respect to their implications on the method of modeling the inputs. Reasons are presented for using probability distributions to model the inputs. This is followed by a discussion of how to model distributions for individual inputs. The paper concludes with a discussion of the importance of correctly modeling the multivariate structure of the inputs.

2. CHARACTERISTICS OF COMPUTER MODELS

Computer models are used in almost every setting imaginable. These settings range from quite simplistic to extremely complex, such as probabilistic risk assessments on nuclear power generating stations. There are several important computer model characteristics that have a direct bearing on the model input process:

1. The computer model may require many inputs (perhaps hundreds).
2. The computer model output may exhibit discontinuities with respect to some of its inputs.
3. The computer model inputs may not behave independently of one another.
4. The computer model predictions may be nonlinear, multivariate, time-dependent functions of the inputs.
5. The relative importance of the individual inputs may be a function of time.

The analyst needs to keep these characteristics clearly in mind when adopting a method for modeling the inputs. Monte Carlo based modeling techniques have a lot to offer the analyst with respect to these computer model characteristics. A discussion of the value of using Monte Carlo based modeling techniques is now given.

3. WHY USE PROBABILISTIC BASED MODELING OF INPUTS

It is not uncommon for an analyst to run a computer model with all values fixed except for one input whose value is varied at high and low values or some variation thereof. Such "one-at-a-time" approaches do not provide an efficient way to perform a screening analysis, provide only conditional information, and may be prohibitively expensive. It is highly unlikely that discontinuities in the output would be detected with a "one-at-a-time" approach. As long as pairs of inputs are physically reasonable (such as temperature and relative humidity not being set independent of one another) the "one- at-a-time" approach can be used with dependent inputs. The detection of nonlinear relationships between inputs and outputs can be difficult without utilizing the entire range of each input. Since the "one-at-a-time" approach yields a conditional analysis it would be difficult to argue that any relative importance ranking of the inputs is meaningful.

An alternative to the "one-at-a-time" approach is to model the inputs in a Monte Carlo fashion, that is, associating a probability distribution with each of the inputs. There are several reasons for preferring Monte Carlo modeling.

1. If properly done, Monte Carlo modeling can be designed to avoid the pitfalls mentioned above.
2. The Monte Carlo approach varies all inputs simultaneously, thoroughly explores the input space, and can be made very efficient.
3. If the probability distributions assigned to the inputs are meaningful then statistical estimates of output quantiles, means and variances can be made.

4. HOW DO YOU DETERMINE MODELING DISTRIBUTIONS

Information available on inputs varies from one computer modeling situation to the next. The following categories characterize many types of information with regard to modeling input:

1. The modeling information consists only of a range of values such as the interval [a,b] for a particular input without any associated probability distribution assigned to the interval.
2. The modeling information consists of a probability distribution over an interval where the probability distribution is either known or is possibly based on expert opinion.
3. The modeling information consists of empirical data without a fitted probability distribution.
4. The modeling information consists of empirical data to which a probability distribution is fit using maximum likelihood procedures and tested with goodness-of-fit techniques.
5. The modeling information consists of a prior probability distribution which is updated with empirical data utilizing Bayesian techniques. Each of these types of information will now be discussed in detail.

Range of the input is given. If only a range of values is available for a particular input, say the interval [a,b], then it may be convenient to use a uniform distribution to model the input. In many cases a simple uniform distribution will suffice; however, if the range covers several orders of magnitude, a loguniform distribution (uniform on the interval [log a, log b]) can be quite useful. Uniform distributions produce equally likely sampling of the entire range of the input (either on a linear or log scale) and provide insight into the relationship between the input (X) and the output (Y) when graphed in a scatterplot of X versus Y. Their use also provides a convenient method for generating random test problems for a computer model. Clearly it would be inappropriate to subject the computer model output to probabilistic interpretation when modeling the input with uniform distributions merely as a tool to facilitate the sampling. Helton and Iman (1982) provide an application of using uniform and loguniform distributions to model input.

A probability distribution is given for the input. In many computer modeling situations the input is modeled by probability distributions (which could include uniform and loguniform). These distributions may be attributable to some mathematical derivation, may be based on historical information, or may result from subjective opinion. When a probability distribution is given, it is only necessary to find the inverse of the distribution function of the input based on a uniform random number generated in some manner with respect to the interval [0,1].

The analyst should take care with respect to the source of the probability distribution when interpreting the results. That is, while modeling distributions derived from mathematical results or historical data can be subjected to probabilistic interpretation, results based on modeling distributions arising from use of expert opinion should be regarded as conditional on the specified distributions. Iman and Helton (1985) provide several examples of modeling input with probability distributions.

Use of empirical data. When empirical data are available the analyst has two choices. One is to construct an empirical distribution function (e.d.f.) from the data and use this as a model for the input. The e.d.f. looks like a stairstep function as it proceeds from left to right, having a value of 0 to the left of the smallest data observation and a value of 1 to the right of the largest data observation. In between, the cumulative step heights will start at $1/n$ at the smallest data observation and increase to $2/n$, $3/n$, and so on to $n/n = 1$ as additional sample observations are encountered from left to right along the horizontal axis. Samples are obtained from this empirical function in exactly the same manner as they are with a mathematically defined cumulative distribution function. That is, the inverse of the e.d.f. is found based on a uniform random number generated in some manner with respect to the interval [0,1]. This approach has the advantage of not introducing additional uncertainty into the analysis by assuming a probability model for the input that may provide a less than adequate representation of the data. Cranwell et al. (1982) provide an example with the input modeled empirically.

The second choice the analyst has is to use statistical techniques to fit a probability distribution to the data and then sample from the cumulative distribution function in the usual manner. That approach will now be discussed.

Fitting a distribution to the empirical data. The analyst is frequently confronted with fitting a probability distribution to the data from among many different distributions that may be tried for a given set of data. A good starting point is to graph the data as either an e.d.f or as a histogram in order to get some feel for the shape of the distribution with respect to symmetry, degree of skewness, or long tails. The standard procedure is to use maximum likelihood techniques to estimate the parameters of the desired distribution and then to use some goodness-of-fit technique to test the adequacy of the fit. For example, Iman (1982) provides graphical methods based on a plot of the e.d.f. for testing the adequacy of fits for normal, lognormal and exponential distributions. If a distribution function has been completely specified without making any parameter estimates from the data, the Kolmogorov goodness-of-fit test can be used to test the adequacy of fit for any distribution. Complete details of the Kolmogorov test are given in Conover (1980), as are references to other goodness-of-fit techniques. Once the distribution is fit the sampling takes place in the manner described previously.

Bayesian updating. If a particular input has been previously modeled by some probability distribution (that is, a prior distribution) and some new data become available, the new data can be used to produce a posterior distribution for modeling the input by incorporating Bayesian

techniques to update the prior distribution. For example, suppose a particular failure rate has previously been modeled by a lognormal distribution and some new data are made available consisting of x failures recorded in n demands. The new data could be modeled by a binomial distribution and used to update the prior distribution. The input values for the computer model are obtained from the cumulative distribution function of the posterior distribution. The posterior distribution may or may not be some well recognized distribution and may require numerical integration and iteration to produce the needed inverse values for the inputs. Hora and Iman (1986) present a tutorial on Bayesian updating with application to system unavailability.

5. MODELING DEPENDENCE AND OTHER CONSTRAINTS AMONG THE INPUTS

Many times the inputs are not independent of one another, and yet are not functionally related. For example, temperature and relative humidity could represent two inputs to a computer model. It is known that the two quantities do not behave independent of one another, but they do not necessarily behave as a functional relationship. It would be a mistake to treat two such inputs as independent of one another, as any output generated under such an assumption would be meaningless. Generally, dependence is quantified in terms of a correlation coefficient. Correlation works fine if the affected inputs both have been modeled with normal distributions. Otherwise, the correlation coefficient quickly loses meaning. A quantification technique that can be used to model all types of distributions, including normal distributions, is the rank correlation coefficient. The rank correlation coefficient is merely the simple correlation coefficient computed on the ranks of the data (see Iman and Conover, 1983).

Iman and Conover (1982) have provided a distribution free technique for pairing observations in a multivariate structure based on rank correlations. This technique is easy to use, preserves the integrity of the sampling scheme (perhaps simple Monte Carlo, or stratified Monte Carlo such as Latin Hypercube sampling as given in McKay, Conover and Beckman, 1979), and preserves the marginal modeling distributions of the individual inputs. This procedure has been incorporated into a computer program for producing Monte Carlo samples by Iman and Shortencarier (1985). This program will produce simple random samples, simple Latin hypercube samples, random samples with restricted pairing, and Latin hypercube samples with restricted pairing.

The restricted pairing can be used either to *induce desired rank correlations* among the inputs or to *reduce spurious rank correlations* among the inputs that occur in simple Monte Carlo with random pairing. To demonstrate the impact of the restricted pairing technique with independent inputs, consider the following illustration. A sample of size 20 was generated for 10 independent inputs. The magnitude of the correlation coefficient was observed for each of the 45 possible pairs of inputs. This process was repeated with restricted pairing. The results are summarized in Table 1.

Table 1.Summary of the magnitudes of the correlations among the 45 possible pairs of correlations for random pairing and restricted pairing.

Absolute Value of correlation	Random pairing Number of pairs	Restricted pairing Number of pairs
.00 to .05	11	37
.05 to .10	10	8
.10 to .15	9	
.15 to .20	1	
.20 to .25	9	
.25 to .30	2	
.30 to .35	1	
.35 to .40	1	
.40 to .45	1	
	45	45

Table 1 makes it clear that spurious correlations can be eliminated from the multivariate sample structure by using restricted pairing whenever desired. On the other hand if the analyst desires to *induce* correlation among one or more pairs of variables, the restricted pairing technique can be used to accomplish this objective also. Iman and Davenport (1982) provide scatterplots of inputs arising from various distributions that have been subjected to varying degrees of rank correlation. Helton et al. (1986) provide an application of input modeling with correlated input.

There may also occur situations in which the analyst must impose restrictions among a pair of variables which cannot be accomplished through inducing a correlation. For example, suppose it is required for X_i to be less than or equal to X_j and that X_i is uniform on the interval [a,b] and X_j is uniform on the interval [c,d] with the requirement that $a \geq c$ and $b \geq d$. In the rectangular region defined from a to b on the horizontal axis and from c to d on the vertical axis, the only pairs of points allowed are those that are either on or above the line $X_i = X_j$. If both X_i and X_j are generated in the usual manner and all pairs (X_i, X_j) are transformed to pairs (X_i, X_j^*) where

$$X_j^* = (X_j - c)(d - X_i)/(d - c) + X_i$$

then X_i will be less than or equal to X_j^*. Under this transformation X_i will remain uniform on the interval [a,b] [a,b] but X_j^* will be uniform on the interval $[X_i,d]$, that is, the distribution of X_j^* becomes conditional on the value of X_i. Moreover, the conditional distribution creates a correlation between X_i and X_j^*. Examples of this type of modeling can be found in Iman and Helton (1985).

6. CONCLUSIONS

The correct modeling of inputs to computer models is a task that deserves careful attention, but frequently receives only superficial treatment. This paper has indicated how different sources of information can be incorporated into the modeling of the individual inputs. The modeling of the multivariate structure usually receives even less attention than the modeling of the individual inputs. This paper has presented a brief discussion of how to model multivariate input and provided references for the interested analyst.

REFERENCES

Conover, W. J. (1980). *Practical Nonparametric Statistics*, Second Edition. John Wiley, New York.

Cranwell, R. M., Campbell, J. E., Helton, J. C., Iman, R. L., Longsine, D. E., Ortiz, N. R., Runkle, G. E. and Shortencarier, M. J. (1982). Risk Methodology for Geologic Disposal of Radioactive Waste: Final Report. Technical Report, SAND81-2573, Sandia National Laboratories, Albuquerque, NM 87185.

Helton, J. C. and Iman, R. L. (1982). Sensitivity Analysis of a Model for the Environmental Movement of Radionuclides. *Health Physics* **42**(5), 565-584.

Helton, J. C., Iman, R. L., Johnson, J. D., and Leigh, C. D. (1986). Uncertainty and Sensitivity Analysis of a Model for Multicomponent Aerosol Dynamics. *Nuclear Technology* **73**(2), 320-342.

Hora, S. C. and Iman, R. L. (1986). A Comparison of Maximus/Bounding and Bayes/Monte Carlo for Fault Tree Uncertainty Analysis. Technical Report, SAND85-2839, Sandia National Laboratories, Albuquerque, NM 87185.

Iman, R. L. (1982). Graphs for Use with the Lilliefors Test for Normal and Exponential Distributions. *The American Statistician* **36**(2), 108-112.

Iman, R. L. and Conover, W. J. (1982). A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables. *Communications in Statistics* **B11**(3), 311-334.

Iman, R. L. and Conover, W. J. (1983). *Modern Business Statistics.* John Wiley and Sons, Inc., New York.

Iman, R. L. and Davenport, J. M. (1982). Rank Correlation Plots for Use with Correlated Input Variables. *Communications in Statistics* **B11**(3), 335-360.

Iman, R. L. and Helton, J. C. (1985). A Comparison of Uncertainty and Sensitivity Analysis Techniques for Computer Models. Technical Report, SAND84-1461, Sandia National Laboratories, Albuquerque, NM 87185.

Iman, R. L. and Shortencarier, M. J. (1985). LHS: A Program to Generate Input Samples for Multivariate Simulations. Computer program announcement in *The American Statistician* **39**(3), 212.

McKay, M. D., Conover, W. J. and Beckman, R. J. (1979). A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics* **21**, 239- 245.

AUTHOR'S BIOGRAPHY

Ronald L. Iman is a member of the technical staff at Sandia National Laboratories. He received a B.S. in mathematics education from Kansas State University in 1962, an M.A. in mathematics from Emporia State University in 1965, and M.S. and Ph.D. degrees in statistics from Kansas State University in 1970 and 1973 respectively. At Sandia he has been responsible for the development of uncertainty and sensitivity analysis techniques for computer models used in probabilistic risk assessments on nuclear power generating stations since 1975. His research interests include uncertainty and sensitivity analysis techniques, and statistical analysis based on rank transformations. He is the co-author of three statistics textbooks published by Wiley. He is a life member and Fellow of ASA.

Ronald L. Iman
Division 6415
Sandia National Laboratories
Albuquerque, NM 87185
(505) 844-8834