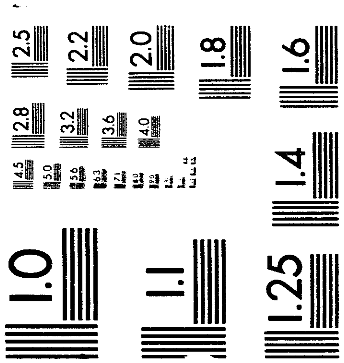
1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6

1 of 1

Conf-940312--35

LA-UR -93-3548

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

TITLE A PROCEDURE FOR ASSESSING UNCERTAINTY IN MODELS

AUTHOR(S) Michael D. McKay and Richard J. Beckman

SUBMITTED TO PSAM II International Conference, March 20-24, 1994
San Diego, CA

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

By acceptance of this article the publisher recognizes that the U S Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution or to allow others to do so, for U S Government purposes

The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U S Department of Energy

MASTER

Los Alamos Los Alamos National Laboratory
Los Alamos, New Mexico 87545

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

A PROCEDURE FOR ASSESSING UNCERTAINTY IN MODELS

Michael D. McKay and Richard J. Beckman

Statistics Group, MS F600
Los Alamos National Laboratory
Los Alamos, NM 87545

INTRODUCTION

This paper discusses uncertainty in the output calculation of a model due to uncertainty in inputs values. Uncertainty in input values, characterized by suitable probability distributions, propagates through the model to a probability distribution of an output. Our objective in studying uncertainty is to identify a subset of inputs as being important in the sense that fixing them greatly reduces the uncertainty, or variability, in the output.

The procedures we propose are demonstrated with an application of the model called MELCOR Accident Consequence Code System (MACCS), described in Helton *et al.* (1992). The purpose of MACCS is to simulate the impact of severe accidents at nuclear power plants on the surrounding environment. In any particular application of MACCS there are likely to be many possible inputs and outputs of interest. In this paper, attention focuses on a single output and 36 inputs. Our objective is to determine a subset of the 36 model inputs that can be said to be dominant, or important, in the sense that they are the principal contributors to uncertainty in the output.

OUTPUT, INPUTS AND PROBABILITY DISTRIBUTIONS

MACCS is used to calculate consequences of a reactor accident at a nuclear power station whose characteristics and those of the surrounding environment are defined by inputs. The output selected for examination is Early Fatalities, meaning the number of fatalities within one year of the accident. MACCS is composed of submodels for source term, plume rise, atmospheric transport, dry deposition, wet deposition, evacuation, food chain transport, and dosimetry and health effects. For each of the inputs, analysts determined plausible ranges of uncertainty for the inputs from the literature, experimental results and submodel considerations. Because of the preliminary nature of this particular analysis, uniform and loguniform probability distributions defined on input ranges were used. For subsets of inputs that could not be treated reasonably as stochastically independent, joint probability distributions or correlation coefficients were determined. For many more details on this part of the analysis process see Helton *et al.* (1992).

In a discussion, it is easy to gloss over the initial step of deciding upon probability distributions for inputs. Nevertheless, this step constitutes the basis for all inference regarding uncertainty. The probability distribution of the output and the identification of important input subsets depend directly on the probability distribution (form and range) of inputs. A very reasonable last stage of uncertainty analysis would be investigation of sensitivity of results to assumed distributions for inputs.

SOME DETAILS

The output of interest is Early Fatalities. In more precise terms, the quantity that MACCS computes is the complementary cumulative distribution function (CCDF) of Early Fatalities induced by treating weather conditions at the time of the accident as a random phenomenon. Tables of weather parameters (one year of hourly readings of wind speed, wind direction, atmospheric stability, and precipitation) are sampled repeatedly during the MACCS run to produce, in effect, a Monte Carlo estimate of the CCDF as a function of time. We denote this output CCDF by $Y(t)$. Then, we have

$$\begin{aligned}
X &= \text{ Input vector of length 36,} \\
Z &= \text{ Weather conditions,} \\
Y(t) &= \text{ Output CCDF of Early Fatalities } (EF), \\
&= Pr\{EF > t\} \text{ for } t \in T = \text{ a range for } EF, \\
&= M(X, Z), \text{ the model MACCS,} \\
f_x(x) &= \text{ probability distribution of } X, \\
f_y(y) &= \text{ probability distribution of } Y(t) \text{ at} \\
& \text{ a specific but arbitrary value of } t.
\end{aligned} \tag{1}$$

Strictly speaking, for each set $\{t_1, t_2, \cdots, t_m\}$, the set $\{Y(t_1), Y(t_2), \cdots, Y(t_m)\}$ has a joint distribution. However, it is sufficient for our purposes to examine the distributions of $Y(t)$ for each t separately. In the notation we suppress the dependence of f_y on t.

We follow an *ad hoc* screening procedure that resembles what one might do when selecting a subset of independent variables in a step-up regression analysis. Namely, we iteratively run MACCS at a sample of input values and examine input and output values to select one or more "important" inputs. Those inputs designated as important are fixed at nominal values and the steps are repeated until variabilitv in the output falls below a threshold. Finally, we examine the cumulative effect of the set of important inputs over the entire range of input values, not just at nominal values. We now discuss the two elements of the sequential procedure, the input sampling procedure and the indication of importance.

INPUT SAMPLING

The input sampling method is a variation of Latin hypercube sampling (McKay, Conover and Beckman, 1979) with independent replications. In an LHS of size k, the range of each input is divided into k intervals of equal probability content. One value is selected from each interval according to the input's conditional probability distribution. Values across inputs are matched at random and without replacement to form k input vectors. In the variation, the probability midpoint of each interval, rather than a sampled value, is used. As a result, an independently replicated sample uses the same values for each input but matches them with different values across inputs. If we denote the number of replicated samples by r, then the total sample size is $n = r \times k$. The reason this type of design was chosen is to accommodate estimation of the importance indicator described in the next section.

For the MACCS study we chose arbitrarily $r = 10$ and $k = 50$ giving $n = 500$ runs. It is very likely that we would have obtained similar results with $r = 5$, and, even, $k = 25$. Sample size requirements were not investigated. The first column of Fig. 1 shows $Y(t)$ for the 50 runs constituting the first LHS replicate within the 500 runs. The variability (uncertainty) in the output as the 36 inputs vary across the input space is apparent. The 81 values of t in this study have been arbitrarily labeled 1, 2, and so forth.

IMPORTANCE

Interest lies in importance of inputs relative to uncertainty in the output. Assuming a suitable definition of importance is found, one could try to determine the importance of each input separately, or try to identify most important subsets of inputs of size 2, 3 and so forth. For this study, we look at the simple case of trying to find a subset of the 36 inputs which accounts for essentially all of the variability in Y.

Suppose we have a partition of inputs into two sets, $X = S_x \cup S_x^c$. Upper case S_x means a subset of p inputs and lower case s_x means the numerical values of the p inputs.

$$\begin{aligned} S_x &= \{X_{i_1}, X_{i_2} \cdots, X_{i_p}\} \\ &= \text{a subset of } p \text{ inputs.} \\ s_x &= \{x_{i_1}, x_{i_2}, \cdots, x_{i_p},\} \\ &= \text{numerical values of the } p \text{ inputs.} \end{aligned} \qquad (2)$$

If S_x effectively identifies important inputs, then the family of conditional distributions $f_{y|s_x}$ indexed on the numerical values of the inputs in the subset S_x will be widely dissimilar. (See, for example, the right side of Fig. 2.) On the other hand, if S_x contains only unimportant inputs, the family of conditional distributions $f_{y|s_x}$ will be similar among themselves and to the marginal distribution of the output, f_y. (See, for example, the left side of Fig. 2.) One motivation for this approach comes from the following consideration. If Y depends only on S_x, then it is stochastically independent of all the other inputs, S_x^c. In that case, S_x^c would be clearly a set of "unimportant" inputs and $f_y = f_{y|s_x^c}$ for all numerical values of the inputs in S_x^c. Realistically, though, importance is not an either-or attribute.

The importance of S_x is related to degree of dependence, in the probabilistic sense, between Y and the subset of inputs S_x. Although there is no universally accepted, unique measure of dependence, one can begin by comparing the distribution f_y to the family of conditional distributions $f_{y|s_x}$ through the popular variance formula (Parzen, 1962)

$$V[Y] = V[E(Y \mid S_x)] + E[V(Y \mid S_x)]\,. \qquad (3)$$

The term on the left side of Eq. 3 is a property of f_y. The inner expectations of the terms on the right side are properties of $f_{y|s_x}$. They represent the mean (E) and variance (V) of the conditional distribution $f_{y|s_x}$. Finally, the outer expectations (V and E) are with respect to the probability distribution f_{s_x} of S_x.

SCREENING

Eq. 3 can be used to form the basis of a screening process where inputs are identified, individually, as potentially important. McKay *et al.* (1992) used the quantity

$$\lambda_j = \frac{V[E(Y \mid X_j)]}{V[Y]} \tag{4}$$

as an importance indicator for input number j. An equivalent indicator

$$I_j = \sqrt{V[Y] - E[V(Y \mid X_j)]} \tag{5}$$

was used by Hora and Iman (1986) and Iman and Hora (1990).

Because of the sampling design used, λ_j can be estimated in a straight forward manner. Each input X_j is represented in the sample by k distinct values repeated r times. For each input X_j, we can label the n output values by y_{st} with $s = 1, 2, \cdots, k$ corresponding to the k distinct values of X_j, and $u = 1, 2, \cdots, r$ corresponding to the r times each value is repeated. Now, we write the common analysis of variance sum of squares partition

$$\sum_{s=1}^{k}\sum_{u=1}^{r}(y_{su} - \overline{y})^2 = r\sum_{s=1}^{k}(\overline{y}_s - \overline{y})^2 + \sum_{s=1}^{k}\sum_{u=1}^{r}(y_{su} - \overline{y}_s)^2 \tag{6}$$

$$\text{SST} \;=\; \text{SSB} \;+\; \text{SSW}$$

where

$$\overline{y}_s = \frac{1}{r}\sum_{u=1}^{r} y_{su} \text{ and } \overline{y} = \frac{1}{k}\sum_{s=1}^{k}\overline{y}_s \, .$$

In the usual manner, $V[Y]$ is estimated by the sum of squares total (SST) divided by $r \times k - 1$, and $V[E(Y \mid X_j)]$ is estimated by the sum of squares between (SSB) divided by $k - 1$. Thus, we see that the usual

$$R^2 = \text{SSB / SST} \, . \tag{8}$$

is proportional to a ratio of estimators of the two components of variance. That is, the importance indicator can be estimated by

$$\widehat{\lambda}_j = R^2\left(\frac{rk-1}{k-1}\right) . \tag{9}$$

Eq. 9 shows that one can use R^2 from sample data as an indicator of importance in place of $\widehat{\lambda}_j$. The probability distribution of R^2 depends on the true distribution of Y. However, if Y is approximately normally distributed and independent of X_j, then R^2 has, approximately, a beta distribution that depends on r and k. Thus, percentage points from the beta distribution might be used for guidance in evaluating the observed size of R^2. An interpretation of R^2 as a measure of degree of stochastic dependence needs to be researched. For now, we point out the obvious: if Y depends exclusively of X_j, then R^2 attains its maximum value of 1. Also, R^2 can be used as an indicator for importance for both individual inputs and subsets of inputs.

We calculate R^2 with both original and rank transformed output values. The rank transformation seems to have advantages for screening when some of the output values appear as outliers.

A SEQUENTIAL PROCEDURE

From many possibilities, we present here a simple, sequential procedure for building S_x which can be used to display the effect of inputs progressively. In a sequence of steps, the output is computed for a sample of input values where the number of inputs fixed becomes progressively larger. For the first step, no inputs are fixed, all 36 inputs vary, to produce representative output traces $Y(t)$ and $R^2(t)$ for each input in the first column of Fig. 1. For the second column of Fig. 1, 9 of the inputs are fixed at their nominal values and the remaining 27 assume the same sample values as in step 1. The spread in output values is noticeably reduced by fixing those 9 inputs. For the final column of Fig. 1, 16 of the 36 inputs are held fixed at their nominal values. At each step, additional inputs to fix are selected according to values of R^2.

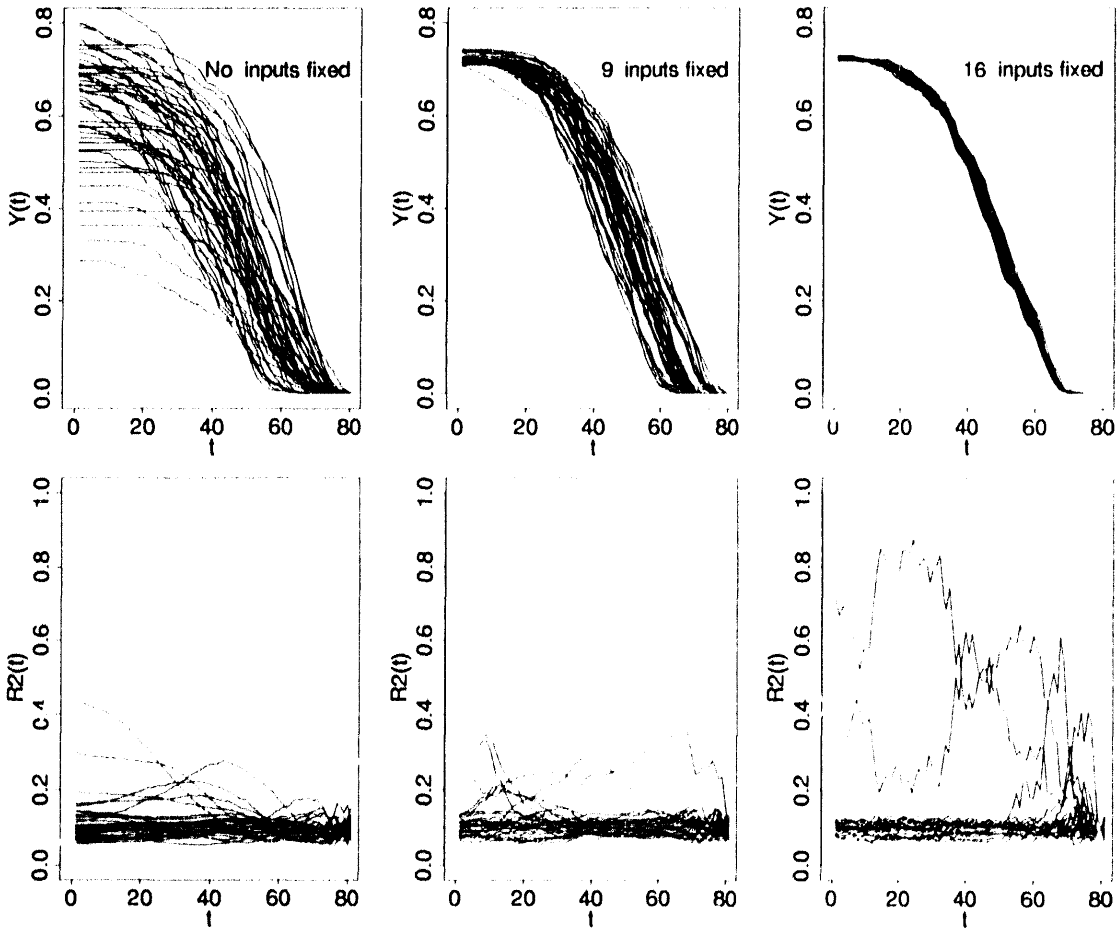

Figure 1. Representative $Y(t)$ and $R^2(t)$ for three steps

VALIDATION

Figure 1 indicates that fixing the 16 inputs in S_x at their nominal values greatly reduces variability in the output. To understand more fully the effects of the inputs in S_x and of the 20 unimportant inputs in S_x^c, we display two sets of probability distributions in Fig. 2 for $Y(t = 1)$. It is important to note the difference in vertical scales in the figure. The left side gives 50 conditional density functions $f_{y|s_x^c}$ where the unimportant inputs take on 50 representative, fixed values. The similarity among these densities indicates how little the unimportant subset affects the distribution of $Y(t = 1)$. On the right side of Fig. 2, the large dissimilarity among the conditional density functions $f_{y|s_x}$, which correspond to 50 representative values of the important inputs, indicates the effects of the important subset.

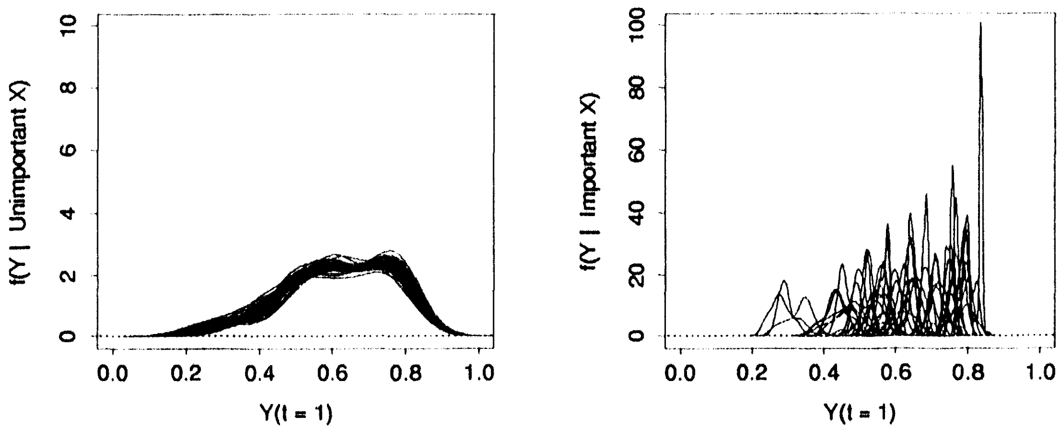

Figure 2. 50 representative conditional density functions when unimportant inputs are fixed, on the left, and when important inputs are fixed, on the right.

SUMMARY

A subset of 16 of the 36 inputs that accounts for essentially all of the variability in the output was identified via a sequential procedure that sampled with replicated, modified LHS and screened with the scaled variance-ratio importance indicator R^2. Conditional distributions were examined to validate the important inputs selected.

ACKNOWLEDGMENTS

This work was supported by the United States Nuclear Regulatory Commission, Office of Nuclear Regulatory Research.

REFERENCES

Helton, J. C., Rollstin, J. A., Sprung, J. L., and Johnson, J. D. (1992). An exploratory sensitivity study with the MACCS reactor accident consequence model. *Reliability Engineering and System Safety*, 36:137-164.

Hora, S. C. and Iman, R. L. (1986). A comparison of Maximum/Bounding and Bayes/Monte Carlo for fault tree uncertainty analysis. Technical Report SAND85-2839, Sandia National Laboratories, Albuquerque, NM.

Iman, R. L. and Hora, S. C. (1990). A robust measure of uncertainty importance for use in fault tree system analysis. *Risk Analysis*, 10(3):401-406.

McKay, M. D., Beckman, R. J., Moore, L. M., and Picard, R. R. (1992). An alternative view of sensitivity in the analysis of computer codes. In *Proceedings of the American Statistical Association Section on Physical and Engineering Sciences*, Boston, Massachusetts, August 9-13.

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*, 22(2):239-245.

Parzen, E. (1962). *Stochastic Processes*. Holden Day, San Francisco, p. 55.

DATE

FILMED

2 / 7 / 94

END