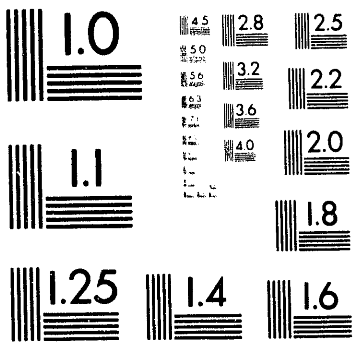
1.0
4.5
2.8
2.5
5.0
3.2
2.2
5.6
6.3
3.6
2.0
4.0
1.1
1.8
1.25
1.4
1.6

1 of 1

Conf-9309215--1

LA-UR -93-3160

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 METHODOLOGY FOR QUANTIFYING UNCERTAINTY IN MODELS

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

SUBMITTED TO Symposium on Surety Technology, 9/28-10/2,1993

DISCLAIMER

MASTER

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

FORM NO 836 R4
ST NO 2629 5/81

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

A METHODOLOGY FOR QUANTIFYING UNCERTAINTY IN MODELS

Michael D. McKay and Richard J. Beckman
Los Alamos National Laboratory, Los Alamos, New Mexico 87545

ABSTRACT

This paper, condensed from McKay et al. (1992) outlines an analysis of uncertainty in the output of computer models arising from uncertainty in inputs (parameters). Uncertainty of this type most often arises when proper input values are imprecisely known. Uncertainty in the output is quantified in its probability distribution, which results from treating the inputs as random variables. The assessment of which inputs are important (sensitivity analysis) with respect to uncertainty is done relative to the probability distribution of the output.

1 INTRODUCTION

The evaluation of models in the form of computer codes (computer programs) becomes more important when the models are used in making decisions that have far reaching effects. For example, the complex models used to study global warming, nuclear reactor safety, and environmental safety and restoration provide vital input to regulatory agencies, whose decisions have large impact on our lives. Although models like those used for policy decisions in government vary widely in their mathematical form, they share some important characteristics. Namely, they often "predict" or calculate things one hopes never to observe, for example, serious accidents at nuclear reactors. Secondly, they are functions of many inputs for which costly data collection may be required to determine appropriate values, ranges and so forth. Finally, the relationship between inputs and output is complex.

There are many aspects to the evaluation of the quality of output of a model. The subject addressed in this paper concerns uncertainty in the output attributable to uncertainty in model inputs (or parameters). Within this area, discussion will focus on the sensitivity or importance of the inputs.

2 UNCERTAINTY AND SENSITIVITY

The more traditional, historical approach to sensitivity is founded in the derivative of the output with respect to each input. Another approach is to consider the output as a random variable and find a meaningful decomposition of variance based on the inputs. In a more general approach, this paper views importance of inputs with respect to uncertainty in the output. We are interested in the type of uncertainty that can be characterized as being due to the values used for the inputs. The quantity of interest for uncertainty is the probability distribution of the model output, which is determined by that of the inputs and the transformation of inputs to output via the model. The sensitivity and importance of inputs is relative to the probability distribution of the model output.

3 MATHEMATICAL FRAMEWORK

Models often have multiple outputs that may be functions of time or location. So as not to needlessly complicate the issue, we consider the case of a single scalar output. Let Y denote the calculated output, which depends on the input vector, X, of length p through the computer model, h(•). Because proper values of the components of X may be unknown or imprecisely known, or because, in some cases, they can only be described stochastically, it is reasonable to treat X as a random variable and to describe uncertainty about X with a probability distribution. Uncertainty in the calculation Y is captured by its own probability distribution. In summary, then,

$$\begin{aligned} Y &= h(X) \\ X &\sim f_x(x)\ ,\ x \epsilon R^p \\ Y &\sim f_y(y)\ . \end{aligned} \tag{1}$$

For now, we treat f_x as known, although in practice, knowledge about it is at best incomplete.

We look to the probability distribution, f_y, for answers to the question "What is the uncertainty in Y?" That is to say, we can use the quantiles of the distribution of Y to construct probability intervals. Alternatively, one might use the variance of Y to quantify uncertainty. In either case, under the assumption that f_y can be adequately estimated, questions answerable with quantiles or moments are covered. However, as has already been mentioned, the issue of how well f_x is known will surely have to be addressed in practice.

Questions of importance of inputs are relative to the probability distribution of Y. That is, they are questions like "Which variables really contribute to (or affect) the probability distribution of the output?" The meaning of importance is given in somewhat of a backwards way as being the complement of unimportant. We say that a subset of inputs is unimportant if the conditional distribution of the output given the subset is essentially independent of the values of the inputs in the subset. These ideas are now examined in more detail.

Suppose that the vector X of inputs is partitioned into X_1, to be the important components, and X_2, to be the unimportant ones. Corresponding to the partition, we write

$$\begin{aligned} Y &= h(X) \\ &= h(X_1, X_2)\ . \end{aligned} \tag{2}$$

We address the question of the unimportance of X_2 by looking at the conditional distributions

$$f_{y|x_2} = \text{distribution } Y \text{ given } X_2 = x_2 \tag{3}$$

as compared to f_y. We say that X_2 is unimportant if f_y and $f_{y|x_2}$ are not substantially different for all values of X_2. Similarly, by looking at the conditional distributions

$$f_{y|x_1} = \text{distribution } Y \text{ given } X_1 = x_1 \tag{4}$$

we say that X_1 contains important inputs if the conditional distributions $f_{y|x_1}$ show large differences for different values of x_1. Of course, a technical way to compare and assess the conditional distributions must be determined. We are currently investigating the use of entropy.

4 APPLICATION

These ideas are applied to the analysis of a compartmental model used to describe the flow of material in an ecosystem. The model calculates concentrations in 15 subsystems, or compartments, as functions of time. For presentation, we have chosen to study the concentration, Y, in one of the compartments at time corresponding to system equilibrium. The flow among compartments, diagrammed in Figure 1, is modeled by a system of linear differential equations. The "transfer coefficients" of the model are functions of the 82 input variables X.

After identifying the model output and inputs, independent beta probability distributions, f_x, were assigned to the inputs. The beta family of distributions was used because

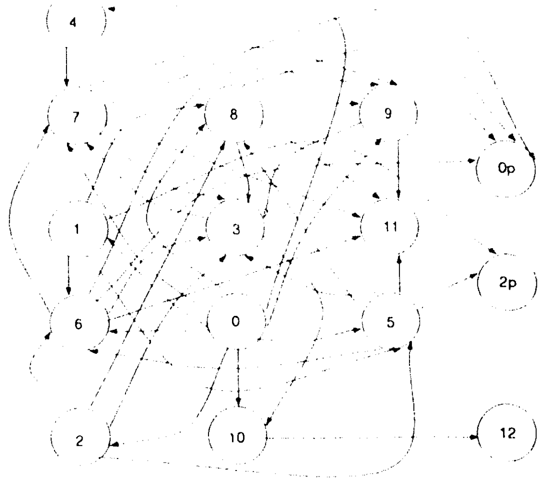

Figure 1. Compartment model

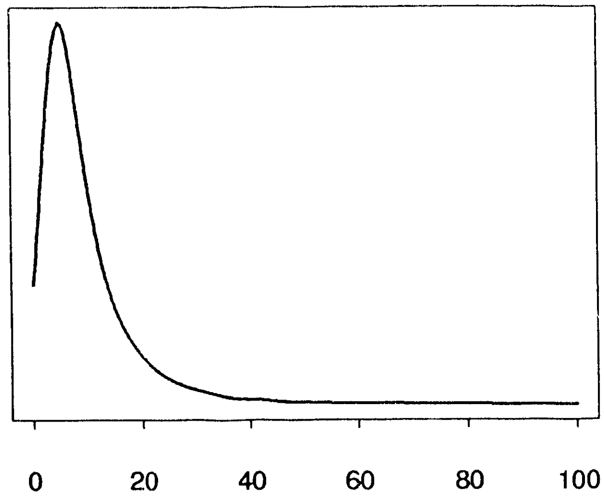

Figure 2. Density function f_y

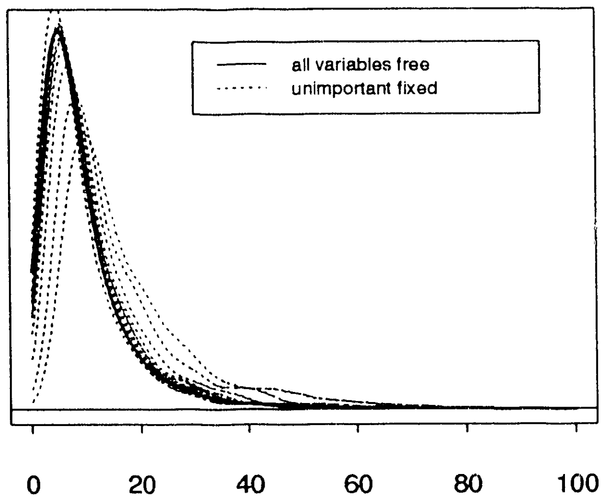

Figure 3. Density functions $f_{y|x_2}$ for 10 values of unimportant inputs X_2

of the wide range in shapes it accommodates. We used only unimodal shapes (none of the U-shaped forms) which included both symmetric and very skewed forms. Parameters of the distributions were inferred from range, best estimate and quantile values obtained from subject-area scientists.

Figure 2 shows the probability distribution of Y, f_y, created when all 82 inputs are free to vary was estimated in Monte Carlo fashion. Latin hypercube sampling (McKay, Conover and Beckman, 1979) as originally described and in a replicated form (McKay et al., 1992) was use throughout the study. In an iterative manner similar to what one might do for variable selection in regression, we selected 7 of the 82 inputs as being possibly important.

To see how the selection procedure performed, we look at 2 sets of density functions. First of all, we investigate whether any important inputs have been missed by looking at $f_{y|x_2}$, which describes Y as a function of X_1 (the "important" inputs) for fixed values of X_2 (the "unimportant" inputs). Figure 3 makes the comparison for 10 values of X_2. The figure indicates acceptable agreement for 8 of 10 values. For 2 of the values of X_2, the agreement between f_y and $f_{y|x_2}$ is not as close, and further analysis may be prudent. In general, however, the figure indicates reasonable agreement between the different $f_{y|x_2}$ and f_y, implying that X_2 identifies only unimportant inputs.

To examine the importance of the set X_1, we look at $f_{y|x_1}$ for 10 values of X_1. These densities are presented in Figure 4. Fixing X_1 produces densities quite different than the marginal density of Y. We do not know at this point, however, whether the set X_1 contains extraneous, unimportant inputs.

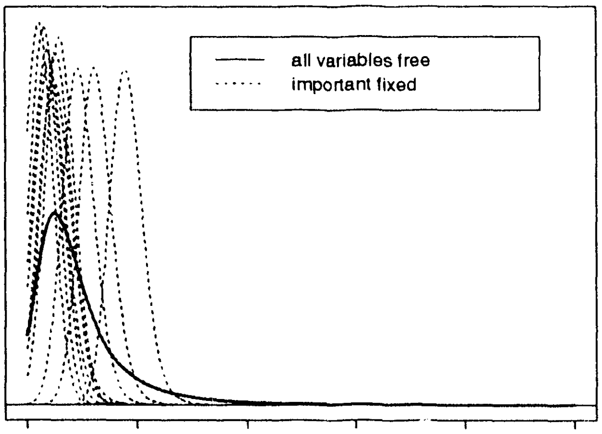

Figure 4. Density functions $f_{y|x_1}$ for 10 values of important inputs X_1

ACKNOWLEDGMENTS

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

REFERENCES

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 selection values of input variables in the analysis of output from a computer code. *Technometrics*, 22(2):239–245.

DATE
FILMED
11/5/93
END