

CNWRA *A center of excellence in earth sciences and engineering*

A Division of Southwest Research Institute™
6220 Culebra Road • San Antonio, Texas, U.S.A. 78228-5166
(210) 522-5160 • Fax (210) 522-5155

July 24, 2002
Contract No. NRC-02-97-009
Account No. 20.01402.765

U.S. Nuclear Regulatory Commission
ATTN: Mr. James Firth
Office of Nuclear Material Safety and Safeguards
Division of Waste Management
Environmental and Performance Assessment Branch
Mail Stop 7C-18
Washington, DC 20555

Subject: Completion of Two Peer-Reviewed Papers on Advances in Sensitivity, Uncertainty, and Importance Methods—Journal Articles IM 20.01402.765.220

Dear Mr. Firth:

The purpose of this letter is to transmit copies of two peer-reviewed papers that were published in fiscal year 2002 in fulfillment of IM 20.01402.765.220, *Advances in Sensitivity, Uncertainty, and Importance Methods—Journal Articles*. These two papers, one of which was presented at the PSAM 6 Conference in San Juan, Puerto Rico, and one which was presented at Risk Analysis 2002 in Sintra, Portugal, were previously submitted to NRC for programmatic review and were approved for publication. The titles of the attached papers are:

Sensitivity Analysis Methods for Identifying Influential Parameters in a Problem with a Large Number of Random Variables by S. Mohanty and R. Codell. The paper describes a comparison of sensitivity analysis methods that have been applied to the TPA Version 4.1 code as well as to other, complex probabilistic models used in conducting risk assessments. This paper was presented at Risk Analysis 2002.

A Partitioning Method for Identifying Important Model Parameters by O. Pensado, V. Troshanov, G. Wittmeyer, and B. Sagar. The paper presents a novel partitioning procedure based on a variety of metrics for identifying important parameters in a probabilistic model. The procedure is also applied to output results from the TPA Version 4.1 code and produces results that are consistent with those obtained by other methods. This paper was presented at PSAM 6.



Washington Office • Twinbrook Metro Plaza #210
12300 Twinbrook Parkway • Rockville, Maryland 20852-1606

James Firth
July 24, 2002
Page 2

If you have any technical questions about the content of these papers or would like to know how each paper was received when presented at the conferences, please contact the senior authors Dr. Sitakanta Mohanty at (210) 522-5185 and Dr. Osvaldo Pensado at (210) 522-6084. Please contact me at (210) 522-5082 if you have programmatic questions.

Sincerely yours,



Gordon W. Wittmeyer, Ph.D.
Manager, Performance Assessment

GW/cw

With Enclosure					Without Enclosure	
cc: J. Linehan	W. Reamer	R. Codell	M. Rahimi	W. Patrick	CNWRA Directors	
B. Meehan	T. Essig	D. Esh	C. McKenney	T. Nagy (SwRI Contracts)	CNWRA Element Managers	
D. DeMarco	J. Schlueter	C. Grossman	J. Danna	P. Maldonado		
D. Riffle	S. Wastler	R. Johnson	K. Stablein	O. Pensado		
J. Greeves	T. McCartin	J. Peckenpaugh	L. Campbell	S. Mohanty		

Sensitivity analysis methods for identifying influential parameters in a problem with a large number of random variables

Sitakanta Mohanty¹ and Richard Codell²

¹*Center for Nuclear Waste Regulatory Analyses, USA*

²*U.S. Nuclear Regulatory Commission, Washington, D.C., USA*

Abstract

Risk analysis can benefit from applications of sensitivity techniques to identify the important parameters. This paper compares the ranking of the ten most influential variables among a possible 330 variables for a model describing the performance of a repository for radioactive waste, using ten different statistical and non-statistical parametric sensitivity analysis methods. Because each method has its advantages and limitations, the selection of the final list of influential parameters is based on the number of times the parameter achieves a high ranking by different methods. The scoring method appears to successfully isolate the most influential parameters.

1 Introduction

Computer modeling provides an avenue to simulate the behavior of complex systems. Many of the input model parameters have large uncertainties. Sensitivity analysis can be used to investigate the model response to these uncertain input parameters. Such studies are particularly useful to identify the most influential parameters affecting model output and to determine the variation in model output that can be explained by these variables.

There are a large variety of sensitivity analysis methods, each with its own strengths and weaknesses, and no method clearly stands out as the best. In this paper, we have picked ten different methods and have applied these methods to a high-level waste repository model, which is characterized by a large number of variables (e.g., 330), to identify influential input variables.

2 Sensitivity Analysis Techniques

Most techniques used herein rely on the Monte Carlo (or its stratified equivalent, Latin Hypercube Sampling) method for probabilistically determining system performance. Many of the input parameters are not precisely known. The Monte Carlo technique makes a series of calculations (called realizations) of the possible states for the system, choosing values for the input parameters from their probability distributions.

A sensitive parameter is one that produces a relatively large change in model response for a unit change in an input parameter. The goal of the sensitivity analyses presented in this paper is to determine the parameters to which model response shows the most sensitivity. The goal of the uncertainty analyses is to determine the parameters that are driving uncertainty (i.e., variation) in response.

2.1 General Model

The response of the system is denoted as y , which is generally a function of random parameters, x_i ; deterministic parameters, d_k ; and model assumptions, a_m . The system response for the j th realization is

$$y_j = f(x_{1,j}, x_{2,j}, \dots, x_{i,j}, \dots, x_{l,j}, d_k, a_m) \quad (1)$$

where l is the total number of sampled parameters in the model, k is the number of deterministic parameters and m is the number of model assumptions. It is assumed that the behavior of the system is simulated by appropriately sampling the random parameters and then computing the system response for each realization of the parameter vector $X_j = \{x_{1,j}, x_{2,j}, \dots, x_{i,j}, \dots, x_{l,j}\}$. For the purposes of identifying influential random parameters and develop understanding of their relationship to the response, we do not consider the dependence of y on deterministic parameters and model assumptions.

2.2 Regression Analyses Methods

Single Linear Regression on One Variable

Single linear regression (i.e., regression with only the first power of a single independent variable), is useful to understand the nature and strength of relationships between input and response variables of a model. The coefficient of determination, R^2 , gives a quantitative measure of the correlation. Even when the response variable is linearly dependent on the input variable being studied, univariate linear regression of Monte Carlo results may fail to show unambiguous correlation because other sampled parameters that affect the response are varying at the same time. When R^2 is small, it is not necessarily a good indicator of the importance of the variable. A better indication of influence is to determine by means of a T-test whether the probability that the slope of the linear regression line is significantly different from zero [1].

The correlation between input and response variables can be enhanced by transforming the variables. In general, variables are transformed by (i) eliminating dimensionality, (ii) reducing the role of the tails of the distributions, (iii) properly scaling the resulting sensitivities to the variability of the input variables, and (iv) using input variable ranks. While transformations generally increase the goodness-

of-fit, they may distort the meaning of the results. For example, transformations such as rank, logarithmic, and power law applied to the response variable, frequently give unfair weight to small response values, which do not affect the mean results as much as the large response values. If the mean response is a desirable quantity, regression results based on transformed variables should be used cautiously.

2.3 Stepwise Multiple Linear Regression

Stepwise multiple linear regression (stepwise regression) determines the most influential input parameters according to how much each input parameter reduces the residual sum of squares (RSS) [2]. The form of the regression equation is

$$y = m_1x_1 + m_2x_2 + \dots + m_nx_n + b \quad (2)$$

where y is the dependent variable, x_i are independent variables (could be raw, transformed, or rank variables), m_i are regression coefficients, and b is the intercept. The regression coefficient, which is the partial derivative of the dependent variable with respect to each of the independent variables, is a measure of linear sensitivity of y to input x_i [3]. The stepwise algorithm calculates the reduction in RSS for the independent variables in the order that gives the greatest reduction first. In the implementation of the procedure, a multiple linear regression model is fitted to the data in an iterative fashion. The procedure starts with the variable, x_n , that explains most of the variation in the model response, y . Then it adds additional variables (one at a time) to maximize the improvement in fit of the model according to the R^2 value, which is an indicator of the fraction of variability in the dependent variable that is explained by the variability of x_i . The sequence in which the inputs are selected and the magnitude of the increment in R^2 provides the measure of uncertainty importance.

2.4 The Kolmogorov-Smirnov (K-S) Test

The K-S test is nonparametric, i.e., a statistical test that does not require specific assumptions about the probability distributions of the data [4]. Probability distribution of a subset (e.g., top 10 percent) of the observations of the input variables is compared to the theoretical (i.e., true) distribution of that variable. If the two distributions are equivalent, then response is not sensitive to the variable in question. Conversely, if the distributions are different, then the variable in question does have an effect on response. For the present study, there are 4,000 vectors in the entire set, and the subset consists of the 400 vectors with the highest responses. The significance of the K-S test was determined at the 95-percent confidence level.

2.5 The Sign Test

The Sign test is also nonparametric. In the Sign test, each observation of the input variable is represented by either a plus sign (+) or a minus sign (-) depending on if it is greater than or less than the median value of the theoretical distribution. A subset of the input parameter values (e.g., 10 percent) corresponding to calculated responses is compared to the theoretical distribution of that input variable. For the present study, there are 4,000 vectors in the entire set, and the subset consists of the 400 vectors with the highest responses. The significance of the Sign test was determined at the 90-percent confidence level.

2.6 Differential Analysis Technique

In the differential analysis technique for determining the most influential input parameters, multiple deterministic runs are made in which an input parameter, x_i , is changed (one at a time) by a known amount, Δx_i , to estimate the first derivative of the performance: $\partial y / \partial x_i = [y(x_i + \Delta x_i) - y(x_i)] / \Delta x_i$. Usually Δx_i in this derivative is relatively small (e.g., 1 percent of the parameter value). Consequently, differential analysis determines sensitivity of parameters only at local points in parameter space and does not consider the wide range of parameter variations as does the Monte Carlo method. This concern is alleviated by evaluating derivatives at several randomly selected points in the sample space and averaging the corresponding sensitivities that are derived from these derivatives. In the analyses presented herein, the derivative is transformed in one of two ways to allow for comparison of sensitivity coefficients between parameters whose units may differ. The first transformation is described by $S_i = (\partial y / \bar{y}) / (\partial x_i / \bar{x}_i)$, where \bar{x}_i and \bar{y} are the mean values of x_i and y , respectively and S_i is the dimensionless normalized sensitivity coefficient. These normalized sensitivity coefficients presented in the above equation are equivalent to the coefficients of the regression equation using the logs of the normalized response and independent variables. Because S_i does not account for the range of the input parameter, a second transformation of the derivative is also performed where the derivative is multiplied by the standard deviation of the input parameter distribution. This transformation is described by $S_\sigma = (\partial y / \partial x_i) \sigma_{x_i}$.

Differential analysis determines sensitivity unambiguously because it deals with changes in only one independent variable at a time. In contrast, regression analysis on the Monte Carlo results can only determine the most influential parameters when those parameters also have large-enough correlation coefficients that they are distinguishable from the confounding effects of the simultaneous sampling of all other independent variables.

2.7 Morris Method Technique

In the Morris method [5], the random variable, $\partial y / \partial x_i$, is evaluated using the current and the previous values of y :

$$\frac{\Delta y}{\Delta x_i} = \frac{y(x_1 + \Delta x_1, \dots, x_i + \Delta x_i, \dots, x_j) - y(x_1 + \Delta x_1, \dots, x_i, \dots, x_j)}{\Delta x_i} \quad (3)$$

To compute $\partial y / \partial x_i$, a design matrix is constructed by (i) subdividing the range of each input variable x_i into $(p-1)$ intervals using $(p-1)$ equally spaced points, (ii) randomly sampling x_i (normalized) from these p intervals of size $\Delta_i = p/2(p-1)$.

The Morris method considers $\partial y / \partial x_i$ as a random variable and uses its mean and standard deviation of the random variable to determine the sensitivity of y to x_i . A large value of mean $\partial y / \partial x_i$ implies that x_i has a large overall influence on y . A large value of standard deviation implies that either x_i has significant interactions

with other input parameters (i.e., x_k , $k = 1, 2, \dots, I$, $k \neq i$) or its influence is highly nonlinear.

2.8 The Fourier Amplitude Sensitivity Test (FAST) Method

Both the differential analysis and the Morris method handle one input parameter at a time. For a nonlinear computational model, in which input parameters are likely to have strong interactions, it would be desirable to have a sensitivity analysis method that would investigate the influence of all input parameters at the same time. The FAST method [6] does this. It first applies the trigonometric transformation $x_i = g_i(\sin \omega_i s)$ to the input parameters. Transformations for various input distribution functions can be found in Lu and Mohanty [7]. The output variable can then be expanded into a Fourier series

$$y(s) = \frac{A_0}{2} + \sum_{i=1}^I A_i \sin(\omega_i s) = y(s + 2\pi) \quad (4)$$

where A_i 's are the Fourier amplitudes of the output variables corresponding to frequencies ω_i .

The trigonometric transforms relate each input variable, x_i , to a unique integer frequency, ω_i . All transforms have a common parameter s , where $0 \leq s \leq 2\pi$. As s varies from 0 to 2π , all the input parameters vary through their ranges simultaneously at different rates controlled by the integer frequencies assigned to them through $x_i = g_i(\sin \omega_i s)$. Equally spaced values of s between 0 and 2π are chosen to generate values of x_i . Because trigonometric transforms and integer frequencies are used, the response, y , becomes periodic in s , and the discrete Fourier analysis can be used to obtain the Fourier coefficients of y with respect to each integer frequency. The sensitivity of y to x_i is measured by the magnitudes of the Fourier coefficients with respect to ω_i , and y is considered sensitive to the input parameters with larger magnitudes of Fourier coefficients.

The use of integer frequencies causes some errors due to "aliasing" (see [7] for an explanation) among Fourier coefficients. The integer frequencies in $x_i = g_i(\sin \omega_i s)$ were chosen to minimize interactions among Fourier coefficients to ensure, as much as possible, that the particular coefficient, A_i , through the particular integer frequency, ω_i , represents only the influence of the corresponding input parameter, x_i . Assuming $0 \leq x_i \leq 1$, the trigonometric transformation functions used here is $x_i = 1/2 + 1/\pi \arcsin[\sin(\omega_i s + r_i)]$, where r_i 's are random numbers.

Because implementing the FAST method is computationally intensive, the number of input variables was limited to 50. According to Cukier et al. [8], as many as 43,606 realizations are needed to perform a satisfactory analysis on 50 input parameters to avoid aliasing among any four Fourier amplitudes.

2.9 Parameter Tree Method

The parameter tree method evaluates relative sensitivity and correlations of the output variable to one or a subgroup of input parameters. In this technique, the Monte Carlo method is used to produce a pool of realizations, which is then

partitioned into bins according to several rules; e.g., all sampled input parameters above their median value. The bins are then examined to determine which input variables appear to have significant effects on the output variable [9].

A tree structure develops by partitioning input parameter space into bins, each forming a branch of the tree based on a partitioning criterion similar to an event tree. The simplest branching criterion is a classification based on parameter magnitude that treats sampled input values as either $a +$ or $a -$ depending on whether the sampled value is greater or less than the branching criterion value. First, a number of Monte Carlo realizations are generated for a given scenario class. Next, the realizations are partitioned into two subsets determined by whether the first influential parameter, x_j , is greater than or less than a specified level. Realizations with a high value are all treated as $a +$ and low as $a -$, regardless of their position within the subset. For example, realizations with all five influential input parameters in a subgroup of five influential parameters sampled above the median would be placed in the same bin. Similarly, all realizations where the first four influential parameters are $a +$ and the last one is $a -$ would be placed in another bin and so on.

Let the number of realizations associated with the two branches be N_{j+} and N_{j-} . Next, the response variable is examined for realization associated with each branch of the tree. The number of realizations with y greater than a partition criterion (e.g., mean) is counted for both the branches. Let these numbers be L_{j+} ($L_{j+} \leq N_{j+}$) and L_{j-} ($L_{j-} \leq N_{j-}$). The difference between L_{j+} / N_{j+} and L_{j-} / N_{j-} is a measure of sensitivity of y to x_j . The procedure is repeated in each of these two subsets with the next influential parameter to be considered and so on until each of the influential parameters is considered. Note that, in this approach, the selection of the second parameter is dependent on the first and so on.

While the parameter tree method is powerful method for dealing with a subgroup of parameters, it is limited to determining a relatively small number of significant variables because at each new branch of the tree, the number of realizations available for analysis decreases on average by half.

2.10 Fractional Factorial Method

Factorial methods are used in the design of experiments [10] and more recently, in testing of computer codes and models [11]. The basic approach is to sample each of the parameters at two or three levels (e.g., a median value divides the parameter range into two levels) and then to run the model to determine the response. A full-factorial design looks at all possible combinations of sampled input variables; e.g., for two levels, there would have to be 2^N samples, where N is the number of variables. Since the current problem has as many as 330 sampled variables, and each run requires several minutes of computer time, a full-factorial design is infeasible.

Fractional factorial designs require fewer than 2^N runs, but at the expense of ambiguous results. For example, a so-called "level 4" design for 330 variables requires 2048 runs. The results from such a level-4 experimental design can yield results for which the main effects of all variables are distinct from each other and two-way interactions of other variables, but can be confounded by some three-way or higher interactions of other variables. However it is possible to use other information generated in the runs to determine in many cases if the results of the

fractional factorial design are truly measuring the response to the variable or combinations of other variables.

In general, the fractional factorial analysis was conducted in the following steps; (1) Develop a fractional factorial design for all variables in the problem taking into account the largest number of runs that can reasonably be handled; (2) From the results of the preliminary screening, perform an analysis of variance (ANOVA) to determine those variables that appear to be significant to a specified statistical significance; (3) Further screen the list of statistically significant variables on the basis of information other than the ANOVA results; and (4) repeat the analyses with a refined set of variables and higher-resolution designs until results are acceptably unambiguous.

3 TEST PROBLEM

The test problem is the TPA Version 4.1 Code [12] for which the most influential input parameters are to be identified. The analyses have been conducted using the nominal case data set (i.e., includes the most likely scenario and excludes low probability and high consequence events), which does not include disruptive external events. The parameters sampled are the ones where a significant amount of uncertainty remains in their value or they have been shown potentially significant to estimating response (output variable) in the process-level sensitivity analyses. Out of 965 input parameters, 330 input parameters are sampled parameters, 635 are deterministic parameters, and there are numerous model assumptions. Only a few of the 330 sampled parameters contribute significantly to the uncertainty in response

4 RESULTS AND ANALYSES

This section presents the sensitivity and uncertainty analysis results generated using methods described in the previous section. Statistical results are treated separately from the non-statistical methods. The nonstatistical methods include differential analysis, Morris method, FAST method and the fractional factorial design method. Detailed description of the meaning of the parameters and their relevance to the performance assessment is outside the scope of this paper.

4.1 Sensitivity Results from Statistical Methods

This section presents the sensitivity analyses based on an initial screening by statistical analysis of a 4,000-vector Monte Carlo analysis of the nominal case. The statistical tests used in the screening were (1) the K-S test; (2) the Sign tests; (3) Single-variable regression including (a) t-test on the regression of the raw data and (b) t-test on the regression of the ranks of the data; (4) Stepwise regression of (a) raw data, (b) the ranks of the data, and (c) the logarithms of the data.

For each of the statistical tests, the resulting regression coefficients were sorted, giving the highest values receiving the best score. Sensitivities that ranked below the 5th percentile in terms of either a t-statistic or F-statistic, were eliminated from consideration (score = ∞). The overall score for a variable consisted of two parts; (1) the number of times that the variable appeared in the six tests with a finite rank (0 to 6), and (2) the sum of the reciprocal of the rank for the six tests. A variant of the second test replaced the rank with its square, but the results did not change the

conclusions. The top 10 ranks from the statistical screening that combines method 1 to 4 are presented in the second column of table 1.

4.2 Sensitivity from Nonstatistical Methods

Results from Differential Analyses: Seven baseline values were randomly sampled for each of the 330 parameters around which values were perturbed. Perturbations ($\pm 1\%$ of the baseline or local value) to the parameters in these random sets were selected so that the parameter values were maintained in their respectively defined ranges. The selection of random values yields calculations similar to one realization of a probabilistic TPA code run. Sensitivities calculated using arithmetic mean of the absolute values of S_i (at 7 points) weighted by the standard deviation of x_i . Then the x_i 's were sorted in the descending order of the sensitivities to identify the influential variables. The top 10 influential input variables are presented in column 3 of table 1.

Results from the Morris Method: In Morris method, seven samples are collected for each random variable $\partial y / \partial x_i$. A 2316×330 matrix was generated and used in sampling input parameters to the TPA code. The 2317 realizations $[(330 + 1) \times 7]$ produced seven samples for each $\partial y / \partial x_i$, which were used to calculate mean and standard deviation for each $\partial y / \partial x_i$. Seven samples were chosen to be consistent with the differential analysis method.

The greater the distance $\partial y / \partial x_i$ for parameter x_i is from zero the more influential the parameter x_i is. Physically, a point with large values of both mean and standard deviation suggests that the corresponding input parameter has not only a strong nonlinear effect itself, but also strong interactive effects with other parameters on the response. Results are presented in column 4 of table 1.

Results from the FAST Method: Conducting sensitivity analyses for all 330 sampled parameters in the TPA code using the FAST method is impractical because it would take more than 40,000 realizations for only 50 parameters. Such a large number of realizations is needed to avoid aliasing among Fourier coefficients [8]. Therefore, preliminary screening was necessary to reduce the number of parameters evaluated with the FAST method. In this paper, the FAST method is applied to the 20 parameters identified by the Morris method. For the 20 parameters, only 4,174 realizations are needed to avoid aliasing among any four Fourier amplitudes. To account for the range of an input parameter, each Fourier amplitude was multiplied by the standard deviation of the corresponding input parameter.

Results from the FAST methods are somewhat limited by the initial selection of 20 parameters from the Morris method.

Results from the Parameter Tree Method: In the parameter tree approach, median, mean, and 90th percentile values were used for parameter distribution for the identified influential input parameters and the response variable. Using a median value cutoff criterion for the input and output variables, 143 out of 4,000 realizations had all 5 of the influential parameters with values above the median. Of these 143 realizations, 128 had responses above the median value for

all 4,000 realizations. These 143 realizations accounted for 24 percent of the population mean of responses. This analysis reinforces the notion that these are indeed influential parameters because 3.5 percent of the realizations account for over 24 percent of the mean from all realizations.

The number of variables that can be captured by this method is limited by the number of realizations because each new branch of the tree cuts the number of samples by approximately half. In table 1 there may be reasonable assurance that only approximately the top 5 variables are significant, and the others are likely to be spurious.

Results from Fractional Factorial Method: The initial screening with the fractional factorial method used a level-4 design for 330 input variables that needed 2,048 runs. There were two levels for each of the input parameter models, chosen to be the 5th and 95th percentiles of the parameter distributions. The TPA code was then run for this experimental design to calculate the responses.

Results from the set of 2,048 runs were then analyzed by ANOVA, using a probability cutoff of 0.05. The ANOVA yielded a set of 100 potentially influential variables. The results were refined to a list of only 37 variables by observations from other information generated by the code; for example, it was possible to eliminate all variables related to seismic failure of the waste packages by observing from other code outputs that there were no seismic failures in any of the runs.

Using the reduced set of variables from the initial screening, we then set up another fractional factorial design with higher discriminatory power. We set up a level 5 run for 37 variables that yielded the list presented in Table 1. With only 37 variables, it was also possible to look at some of the two-way and 3-way interactions that were combinations of the main effects, and to make conjectures about 4th and higher order interactions of those variables that might be explored by additional factorial designs. With less than 10 variables from the second screening, a full factorial design would require only 1024 additional runs. This experiment will be run in the near future.

5 CONCLUSIONS

This paper describes a suite of sensitivity analysis techniques to identify model variable whose uncertainty and variability strongly influence model response. These techniques help focus attention on what are likely to be the most important to response and also can be used to identify deficiencies in the models and data.

The sensitivity analyses employed in this work were conducted using the functional relations between the model input variables and the response variable embodied in the TPA code. Variety of statistical techniques (e.g., regression-based methods and parameter tree method) using a large set of Monte Carlo runs (4,000 vectors) and nonstatistical techniques (differential analysis, Morris method, FAST method, and fractional factorials) using 250–4,000 TPA realizations were used in this analysis. The parameter tree method allowed the determination of combinations of variables that led to the highest responses. The Morris method and the FAST method were used to determine what further insights could be gained from techniques specifically designed for nonlinear models.

Results from the regression analyses were based on normalized, log-transforms of the normalized inputs and ranks. The normalized results weight each result equally, whereas the log-normalized results tend to overemphasize smaller doses. However, the log-transformed results generally provide a better fit for the regression equations. Results of the regression analyses are standardized to account for the ranges of the input variables and allow a more accurate ranking of sensitivity coefficients.

6 ACKNOWLEDGMENTS

The paper was prepared to document work performed by the Center for Nuclear Waste Regulatory Analyses (CNWRA) for the Nuclear Regulatory Commission (NRC) under Contract No. NRC-02-97-009. The activities reported here were performed on behalf of the Office of Nuclear Material Safety and Safeguards (NMSS). The paper is an independent product of the CNWRA and does not necessarily reflect the views or regulatory position of the NRC.

7 REFERENCES

- [1] Benjamin, J.R., and C.A. Cornell. *Probability, Statistics, and Decision for Civil Engineers*. New York: McGraw-Hill. 1970.
- [2] Helton, J.C., J.W. Garner, R.D. McCurley, and D.K. Rudeen. *Sensitivity Analysis Techniques and Results for Performance Assessment at the Waste Isolation Pilot Plant*. SAND 90-7103. Albuquerque, NM: Sandia National Laboratories. 1991.
- [3] Draper, N.R., and H. Smith, Jr. *Applied Regression Analysis*. 2nd Edition. New York: John Wiley and Sons, Inc. 1981.
- [4] Bowen, W.M., and C.A. Bennett, eds. *Statistical Methods for Nuclear Material Management*. NUREG/CR-4604. Washington, DC: Nuclear Regulatory Commission. 1988.
- [5] Morris, M.D. Factorial sampling plans for preliminary computational experiments. *Technometrics* 33(2): 161-174. 1991.
- [6] Cukier, R.I., C.M. Fortuin, K.E. Schuler, A.G. Petschek, and J.H. Schaibly. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I: Theory. *Journal of Chemical Physics* 59(8): 3,873-3,878. 1973.
- [7] Lu, Y. and S. Mohanty. Sensitivity analysis of a complex, proposed geologic waste disposal system using the Fourier amplitude sensitivity test method. *Reliability Engineering and System Safety* 72(3), pp 275-291. 2001.
- [8] Cukier, R.I., J.H. Schaibly, and K.E. Schuler. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. III: Analysis of the approximation. *Journal of Chemical Physics*, 63(3), pp. 1,140-1,149. 1975.
- [9] Jazemba MS, Sagar B. A Parameter Tree Approach to Estimating System Sensitivities to Parameter Sets. *Reliability Engineering and System Safety*. 67, pp. 89-102. 2000.
- [10] Box, G.E.P. and J.S. Hunter, "The 2^{k-p} Fractional Factorial Designs Part 1, 1961, republished in *Technometrics*. 42 (1), pp. 28-47. 2000
- [11] Schmidt, S.R. and Launsby, R.G. *Understanding Industrial Designed Experiments*. Colorado Springs: Air Academy Press. 1991

- [12] Mohanty S, McCartin TJ (coordinators). NRC Sensitivity and Uncertainty Analyses for a Proposed HLW Repository at Yucca Mountain, Nevada Using TPA 3.1—Volume I: Conceptual Models and Data. NUREG-1668. Washington, DC: US Nuclear Regulatory Commission: 2001.

Table 1. Top 10 influential parameters from statistical and non-statistical analyses. Entries in the columns under each method represents numerical representation of the variable name.

Parameter Rank	Statistics/Regression	Differential Analysis	Morris Method	FAST Method	Parameter Tree Method	Fractional Factorial Design Method
1	12	60	1	259	60	12
2	61	63	2	237	63	60
3	60	12	12	235	61	5
4	63	61	60	62	301	61
5	62	304	292	69	52	62
6	5	70	239	61	2	63
7	1	1	225	77	1	1
8	4	69	223	63	3	239
9	237	78	63	2	287	237 (*)
10	239	239	61	60	15	131 (*)

* These parameters are included for reference, but were below the 5 percentile cutoff from ANOVA probability.

A Partitioning Method For Identifying Important Model Parameters

Oswaldo Pensado¹, Velin Troshanov², Gordon Wittmeyer¹, and Budhi Sagar¹

¹Center for Nuclear Waste Regulatory Analyses
Southwest Research Institute, 6220 Culebra Road
San Antonio, Texas 78238-5166, USA

²College of Liberal Arts and Sciences,
University of Illinois at Urbana-Champaign, USA

ABSTRACT

A general method for identifying important parameters of complex stochastic models is presented. The method is applied to the analysis of a performance assessment model of a geologic repository. A small set of important parameters is derived and it is verified that this small set is sufficient to explain the nature of the model output.

KEYWORDS

Sensitivity analysis, Monte Carlo simulation, stochastic model, geologic repository, performance assessment

INTRODUCTION

For complex models incorporating multiple stochastic parameters, it is frequently helpful to determine the parameters that most influence their output, and values at which the parameters become influential. In this paper, we describe a novel method for accomplishing this task. This technique, referred to as the partitioning method, has greater power in identifying possible correlations among input and output variables than traditional methods such as linear regression, is computationally simple, and can be efficiently programmed.

The motivation for the partitioning method was to develop a technique to analyze results of a model to assess the performance of a proposed geologic repository at Yucca Mountain, Nevada. The model, implemented in the Total-system Performance Assessment (TPA) code (Mohanty and McCartin, 2000), has over 300 parameters (some of them correlated) that have assigned probability distributions. This code is executed in a Monte Carlo mode using the Latin Hypercube method to sample values of stochastic parameters. The main output of the code is a large number of realizations, each realization consisting of total effective dose equivalent (TEDE) to a reasonably maximally exposed individual as a function of time. The mean and confidence bounds for the TEDE as functions of time can be derived

from these multiple realizations. Each realization is associated with a particular set of values of input parameters. In the United States, disposal regulations applicable to Yucca Mountain require that the peak of the mean TEDE within 10,000 years be below a specified value. In this paper, the partitioning method is used to identify the set of most important stochastic parameters affecting different attributes of the TEDE (simply referred to as the annual dose from here on).

DESCRIPTION OF THE PARTITIONING METHOD

An outline of the partitioning method is provided as follows. Partition the output realizations into two bins, one bin containing those realizations contributing the most to the mean annual dose (contributing realizations) and a second bin containing all the remaining realizations (non-contributing realizations). We explored four different approaches for defining "contributing" and "non-contributing" realizations, discussed later. Let A be the parameter whose importance is to be evaluated. Plot a cumulative distribution function and a complementary cumulative distribution function for the set of values of A that are associated with the contributing and non-contributing realizations, respectively. Let (x_A, P_A) be the coordinates of the intersection of these two curves. The probability value P_A can be used to measure the importance of the parameter A . For example, the importance index for parameter A , z_A , can be defined as

$$z_A = 0.5 - P_A \quad (1)$$

High values of $|z_A|$ (i.e., $|z_A| > 0.1$) indicate an evident partitioning of parameter A into two subsets, related to the contributing and non-contributing realizations. The greater the value of $|z_A|$ the more important is variable A . Values of $|z_A| < 0.1$ suggest a lack of partitioning and a lack of importance of the parameter A . If $|z_A|$ is large (i.e., $|z_A| > 0.1$) and z_A is positive (negative), then there is a positive (negative) correlation between the parameter and the mean annual dose. Direct comparison of $|z_A|$ yields the ranking of the most important parameters in the stochastic model. The intersection value x_A also has an interesting interpretation. If $|z_A| > 0.1$ and z_A is positive, then there is a greater likelihood of $A > x_A$ for contributing realizations and $A < x_A$ for non-contributing realizations. In this sense, the value of the intersection, x_A , defines a partitioning value for the parameter A .

Four methods were explored to define contributing and non-contributing realizations. Method 1 was selected to detect the influence of any given realization on the peak of the mean annual dose, ignoring the time at which the peak may occur. Let p_{-i} be the maximum of the mean annual dose, computed without accounting for i th realization in the determination of the mean; i.e.,

$$p_{-i} = \max \left(\frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^n d_j(t) \right) \quad (2)$$

n is the total number of realizations and $d_j(t)$ is the annual dose as a function of time for the j th realization. The function max has the usual mathematical meaning. The discriminating index associated with the i th realization, a_i , is defined as

$$a_i = \frac{p_{-i} - p_T}{p_T} \quad (3)$$

where p_T represents the peak of the mean annual dose during the simulation time. For Method 1, the contributing realizations are defined as those satisfying $|a_j| > a_\mu$, where a_μ is the mean of the set of $|a_j|$ values ($j=1,2, \dots, n$). The non-contributing realizations are those for which $|a_j| \leq a_\mu$. It can be shown that $a_i \leq 1/n$. For most of the realizations considered in this paper, a_i is close $1/n$; thus, a_μ is also a number close to $1/n$.

Method 2 was selected to detect the influence of a realization on the peak of the mean dose, at the time at which the peak occurs. Let t_T be the time at which the peak dose, p_T , occurs. The discriminating index for the i th realization, b_i , is defined as

$$b_i = \frac{d_i(t_T) - p_T}{p_T} \quad (4)$$

$d_i(t_T)$ is the annual dose for the i th realization evaluated at time t_T . The contributing realizations are defined as those satisfying $b_i > 0$ and the non-contributing are all of the others. Since for the majority of the realizations considered in this paper the annual dose is negligible compared to the peak of the mean annual dose, b_i is in general close to -1.

Method 3 was designed to highlight influences on the mean dose over the complete simulation period. The norm in the space of continuous functions in the interval $[0, t_{\max}]$ is defined as

$$\|f\| = \sqrt{\int_0^{t_{\max}} \{f(t)\}^2 dt} \quad (5)$$

t_{\max} is the maximum time of the simulation period and f is a continuous function. Let $d_\mu(t)$ represent the mean annual dose as function of time. The discriminating index for the i th realization, c_i , is defined as

$$c_i = \frac{\|d_i - d_\mu\|^2}{\|d_\mu\|^2} \quad (6)$$

Chun et al. (2000) used an expression similar to Eqn. 6 to measure changes in output cumulative distribution functions. Let c_μ be the mean value of the set of c_j values ($j=1,2,\dots,n$). The contributing realizations are selected as those for which $c_i > c_\mu$ and the non-contributing realizations are all of the others. Since for the majority of the realizations considered in this paper the annual dose is negligible compared to the mean annual dose, c_i is in general close to one.

Method 4 was also designed to highlight the influences on the mean annual dose over the complete simulation period. The discriminating index for the i th realization, \bar{c}_i , is defined as

$$\bar{c}_i = \frac{\|d_{-i} - d_\mu\|^2}{\|d_\mu\|^2} \quad (7)$$

d_{-i} is computed as

$$d_{-i} = \frac{1}{n-1} \int_0^{t_{\max}} \sum_{\substack{j=1 \\ j \neq i}}^n d_j(t) \quad (8)$$

The contributing realizations are defined as those having values of \bar{c}_i greater than the mean of the set of \bar{c}_j values ($j=1,2,\dots,n$). It can be shown that $\bar{c}_i \approx c_i/n^2$, provided that the number of realizations, n , is large enough. Thus, Method 4 is equivalent to Method 3; the yield identical results if the number of realizations is large, as is the case in this paper.

RESULTS

Data generated with 4000 realizations of the TPA Code Version 4.1j were used for the analyses. In this version of the TPA code, 330 stochastic input parameters are considered in the non-disruptive base case. In general, the discriminating indices (i.e., a_i , b_i , and c_i) defined above tend to be close to a constant value (i.e., 1/4000, -1, 1, respectively). Thus, linear regression between the discriminating index and parameter values yields a slope that is not clearly different from zero. In other words, linear regression cannot be used to identify a correlation between parameter values and the discriminating index. On the other hand, the partitioning is capable of detecting correlations, if they exist.

The importance indices, z_A , were computed for all of the stochastic parameters using the three methods defined above (methods 3 and 4 are equivalent). The parameters were sorted according to decreasing values of $|z_A|$. The most important parameters are those with highest values of $|z_A|$. The list of the most important parameters for 10,000 year and 100,000 year realizations are included in Table 1.

TABLE 1
LIST OF MOST IMPORTANT PARAMETERS

Parameter	100,000 yr	10,000 yr	Correlated	Meaning
<i>Preexponential_SFDissolutionModel2</i>	×	×		Factor modulating the spent fuel dissolution rate
AlluviumMatrixRD_SAV_Np	×	×	C ₁	Retardation coefficient for Np in the alluvium
<i>SubAreaWetFraction</i>	×	×	B ₁	Related to the amount of water at the drift
AA_1_1[C/m2/yr]	×			Corrosion rate of Alloy 22
<i>ArealAverageMeanAnnualInfiltrationAtStart [mm/yr]</i>	×	×	B ₂	Mean annual infiltration for current climate
AlluviumMatrixRD_SAV_Pu	×	×	C ₂	Retardation coefficient for Pu in the alluvium
AlluviumMatrixRD_SAV_Am	×	×	C ₂	Retardation coefficient for Am in the alluvium
AlluviumMatrixRD_SAV_U	×	×	C ₂	Retardation coefficient for U in the alluvium
DistanceToTuffAlluviumInterface[km]	×	×		Related to location of tuff/alluvium interface
<i>WastePackageFlowMultiplicationFactor</i>	×	×		Related to the amount of water for release
MatrixPermeability_TSw_[m2]	×	×	B ₂	Matrix permeability for Topopah Spring tuff-welded
WellPumpingRateAtReceptorGroup20km [gal/day]*	×	×		Well pumping rate for farming receptor group located at a distance greater than 20 km
AlluviumMatrixRD_SAV_Th	×		C ₂	Retardation coefficient for Th in the alluvium
FractionOfCondensateTowardRepository[1/yr]	×	×		Fraction of condensed water moving towards the repository
ImmobilePorosityPenetrationFraction_ST FF	×			Effective fraction of saturated rock matrix accessible to matrix diffusion
MatrixKD_UCF_Am[m3/kg]	×			Matrix sorption coefficient (Upper Crater Flat) for Am

SolubilityNp[kg/m3]	×			Solubility of Np
InterceptionFraction/Irrigate	×	×		Fraction of irrigation interception
<i>DefectiveFractionOfWPs/cell</i>		×		Related to the number of waste packages assumed initially failed
DripShieldFailureTime[yr]		×		Time of failure of the drip shield
FractionOfCondensateRemoved[1/yr]		×		Fraction of condensed water not intersecting the drifts
RntoDetermineFaultOrientation		×		Random number to determine fault orientation
* In the TPA Code Version 4.1j, non-disruptive base case, the pumping rate is sampled from a probability distribution function. United States regulations for the proposed Yucca Mountain Site require that the pumping rate of well water considered for performance assessment analysis be a particular fixed value. Future versions of the TPA code will be made consistent with this recent regulatory requirement.				

The parameter names in Table 1 are the same as those used in the TPA Code. The three methods coincide in that pre-exponential factor modulating the rate of spent fuel dissolution is the most important parameter (listed as the first entry in Table 1). If a parameter ranked within the first 20, for at least two of the three methods, such a parameter was included in Table 1. The three methods coincide in the top nine (five) parameters —indicated in bold (italic) font in Table 1— for 100,000 (10,000) year simulations, although the ranking is slightly different from method to method. The 18 (17) most important parameters for 100,000 (10,000) year simulations are indicated by the label × under the 100,000 (10,000) year column in Table 1. In the non-disruptive base case, the parameters labeled with B₂ and C₂, under the Correlated column, are correlated to the parameters labeled with B₁ and C₁, respectively. Parameters labeled with B₂ and C₂ appear important because they are correlated to the parameters labeled with B₁ and C₁, as is shown later. Several runs of the TPA code were completed to verify that the parameters in Table 1 are sufficient to reproduce the variance of the annual dose and the magnitude of the mean annual dose. The results are reported in Figure 1.

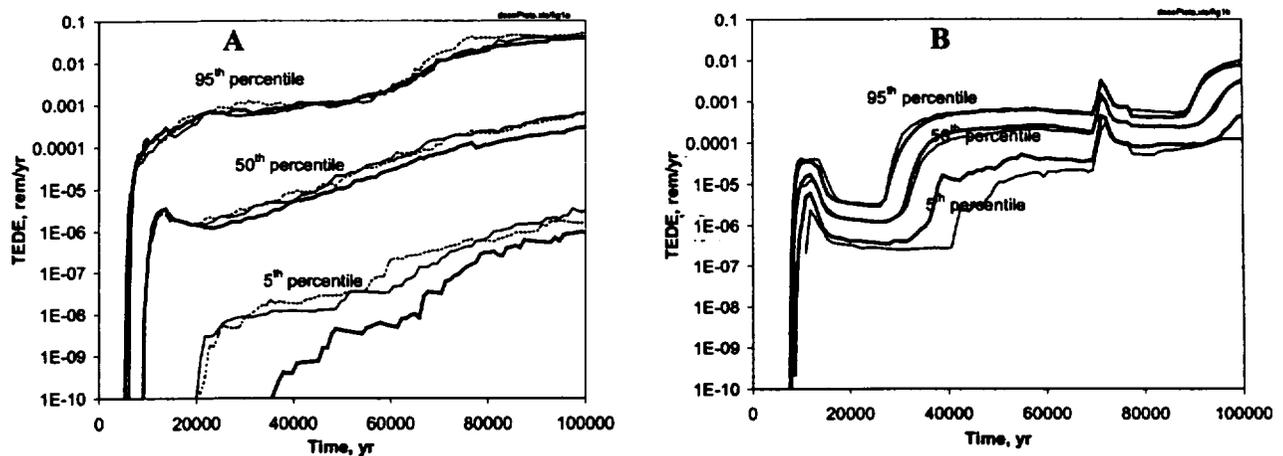


Figure 1: Plots of the 5th, 50th, and 95th percentile of the annual dose versus time.

(A) Base case, and cases 1 and 2.

(B) Cases 3 and 4. See main text for the definition of cases 1 to 4.

Figure 1 includes 5th, 50th, and 95th percentile curves for the annual dose versus time. Figure 1-A presents results for the base case and cases 1 and 2. The thick lines are associated to results of the non-disruptive base case (500 realizations). For the case 1 (thin lines), the important parameters in Table 1 were sampled stochastically (300 realizations) in the range defined in the base case, with the exception of those parameters labeled with B₂ and C₂. All of the other parameters, including those labeled with B₂ and C₂, were fixed at their mean values. A total of 16 parameters were sampled. For the case 2 (dotted lines in Figure 1-A), the parameters in Table 1 in bold or italic font, except those labeled with B₂ and C₂, and the parameter DripShieldFailureTime[yr] were sampled stochastically

(300 realizations). All of the other parameters, including the B_2 and C_2 parameters, were fixed at their mean values. Thus, a total of 8 parameters were sampled for the case 2. The confidence intervals for the base case and cases 1 and 2 are very similar, with relevant variations only in the 5th percentile curves. It is concluded that 8 parameters suffice to account for the gross variance of the annual dose. The 50th and 95th percentile curves compare within less than an order of magnitude for the three cases. Furthermore, the mean annual dose curves (not included in Figure 1) are almost the same for the three cases (they differ by much less than an order of magnitude at all times for the three cases).

Figure 1-B, includes results for cases 3 and 4. For case 3 (thick lines), all of the important parameters in Table 1 were fixed at their mean values (a total of 22 parameters) and all of the others were sampled. Case 4 (thin lines) is the reverse to case 2; the 8 parameters of the case 2 were fixed at their mean values, and the remaining 322 input parameters were sampled. Figure 1-B summarizes 300-realization runs.

In Figure 1 it is noted that the variance in the annual dose deriving from the variance of 322 parameters is small compared to that resulting from the variance of the 8 most important parameters identified in cases 2 and 4. Some of the parameters (those with labels B_2 and C_2 in Table 1) are ranked high by the partitioning method because they are correlated to important parameters. Method 3 was capable of ranking the parameter associated with the failure time of the drip shield within the highest six parameters, because it was designed to identify parameters affecting the annual dose in the complete simulation period, as opposed to methods 1 and 2, which focus on the peak of the mean annual dose. The partitioning method succeeded in identifying a small set of parameters controlling the variance of the annual dose.

CONCLUSIONS

The partitioning method for identifying important parameters was discussed. Although there was a direct motivation to analyze the performance assessment model of a geologic repository, the method is quite general and can be applied to the analysis of any data in which the output depends upon stochastic input parameters. In the particular example of the geologic repository, the partitioning method indicates that the mean annual dose rate is influenced most by parameters controlling the rate of release of radionuclides, corrosion rates of container materials, the amount of water available for radionuclide transport, and retardation coefficients for neptunium.

ACKNOWLEDGEMENTS

This paper was prepared to document work performed on behalf of the U.S. Nuclear Regulatory Commission (NRC), Office of Nuclear Material Safety and Safeguards, under Contract No. 02-97-009. The views expressed in this paper are those of the authors and do not necessarily reflect the views or regulatory position of the NRC.

REFERENCES

- Chun M.-H., Han S.-J., and Tak N.-I. (2000). An uncertainty importance measure using a distance metric for the change in cumulative distribution function. *Reliability Engineering and System Safety*. **70**, 313-321
- Mohanty S. and McCartin T. J. Coordinators. (2000). Total-system Performance Assessment (TPA) Version 4.0 Code: Module Description and User's Guide (Draft), Center for Nuclear Waste Regulatory Analyses, San Antonio, Texas