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

February 25, 2002
Contract No. NRC-02-97-009
Account No. 20.01402.765

U.S. Nuclear Regulatory Commission
ATTN: Mrs. Deborah A. DeMarco
Two White Flint North
11545 Rockville Pike
Mail Stop T8 A23
Washington, DC 20555

Subject: Programmatic review of two papers for Conferences

Dear Mrs. DeMarco:

The enclosed papers, which will be submitted for publication in the proceedings volumes for two conferences, are being submitted for programmatic review. The titles 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. (Risk Analysis 2002, Sintra, Portugal, June 19–21, 2002.)

"Mean-based Sensitivity or Uncertainty Importance Measures for Identifying Influential Parameters" by S. Mohanty and Y-T. (Justin) Wu. The paper describes sensitivity measures developed under an Internal Research and Development project at the Southwest Research Institute and implementation in the Yucca Mountain problem. (PSAM 6, San Juan, Puerto Rico, June 23–28, 2002.)

Please advise me of the results of your programmatic review. Your cooperation in this matter is appreciated. Please contact Sitakanta Mohanty at (210) 522-5185 if you have any questions regarding this paper.

Sincerely yours,

Budhi Sagar
Technical Director

Enclosure

cc:

J.Linehan	T. Essig	R.K. Johnson	W. Patrick	R. Benke
E. Whitt	S. Wastler	T. McCartin	CNWRA Directors	O. Pensado
B. Meehan	J. Firth	C. McKenney	CNWRA Element Managers	S. Mayer
J. Greeves	R. Codell	J. Peckenpaugh	S. Mohanty	O. Povetko
W. Reamer	D. Esh	M. Rahimi	P. LaPlante	L. Howard
J. Schuller	C. Grossman	M. Thaggard	M. Smith	

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

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

Sitakanta Mohanty[1] and Richard Codell[2]
[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 jth realization is

$$y_j = f\left(x_{1,j}, x_{2,j}, \dots, x_{i,j}, \dots, x_{I,j}, d_k, a_m\right) \tag{1}$$

where I 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 $X_j = \{x_{1,j}, x_{2,j}, \dots, x_{i,j}, \dots, x_{I,j}\}$. 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 goodnessof-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

$$y = m_1 x_1 + m_2 x_2 + \ldots + m_n x_n + b \tag{2}$$

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_i, 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, □x_i, to estimate the first derivative of the performance: $\partial y / \partial x = [y(x_i + \Delta x_i) - y(x_i)] / \Delta x_i$. 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 $S_i = (\partial y / \overline{y}) / (\partial x_i / \overline{x_i})$, where $\overline{x_i}$ and $\overline{y}$ 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:

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

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, $k = 1, 2, ..., I, k \neq i$) 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

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

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 $0 \le s \le 2\pi$. 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 $0 \le x_i \le 1$, 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_i, 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_{l+} and N_{l-}. 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_{1+} ($L_{1+} \leq N_{1+}$) and L_{1-} ($L_{1-} \leq N_{1-}$). The difference between L_{1+} / N_{1+} and L_{1-} / N_{1-} is a measure of sensitivity of y to x_l. 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 in 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 (+-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)× 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 5^{th} and 95^{th} 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 4^{th} 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.

MEAN-BASED SENSITIVITY OR UNCERTAINTY IMPORTANCE MEASURES FOR IDENTIFYING INFLUENTIAL PARAMETERS

Sitakanta Mohanty[1], Y-T. (Justin) Wu[2]

1. CNWRA, Southwest Research Institute, San Antonio, Texas, USA
2. Applied Research Associates, Inc., Raleigh, North Carolina, USA

ABSTRACT

Two sensitivity (or uncertainty importance) measures particularly relevant to the disposal of HLW are presented. These measures are referred to as performance-mean-based sensitivity measures, $\partial\mu_Y / \partial\mu_{X_i}$ and $\partial\mu_Y / \partial\sigma_{X_i}$, where μ_Y is the mean of the model output Y and σ_{X_i} is the standard deviation of the input variable X_i. These two sensitivity measures are demonstrated using the U.S. Nuclear Regulatory Commission's total-system performance assessment model, for evaluating the proposed repository at Yucca Mountain. Based on $\partial\mu_Y / \partial\mu_{X_i}$, fifteen out of 330 variables are identified as significantly contributing to sensitivities at 95% acceptance limit. Similarly, based on the calculated $\partial\mu_Y / \partial\sigma_{X_i}$, twenty variables are identified as significantly contributing to sensitivities. Because of the large variability in the performance, approximately 700 samples are needed for the ranking of the variables to be stabilized.

KEYWORDS

Sensitivity Analysis, Uncertainty Analysis, Risk Assessment, System Modeling, Nuclear Fuel Cycle, and Waste Management

INTRODUCTION

Physics-based probabilistic analysis of engineered and natural systems is emerging as an important tool for studying reliability in addition to field and laboratory tests. However, new challenges exist because highly complicated physics-based models are computationally intensive and involve a large number of parameters. The performance assessment of a high-level radioactive waste (HLW) disposal is an example. The performance assessment model has a large number of input parameters that are described by probability distribution functions representing uncertainty and variability. Sensitivity analysis of the performance assessment model is conducted to explain the variability in the output due to uncertainties in the model (not considered in the paper) and input parameters and to determine the most influential input parameters that control the behavior of the output. Knowledge of the most influential input parameters is important because (among other reasons) it can provide an insight on where more efforts should be devoted to reduce the uncertainties in the output and to significantly improve the understanding of the system.

A variety of sensitivity measures have been used in the literature to identify influential parameters emphasizing different aspects of the input-output relationships. In a recently published article by Mohanty and Wu [1], two sampling-based sensitivity measures in the context of the CDF-sensitivity analysis function were presented. However, the HLW problem requires sensitivity measures that are consistent with the regulatory criteria, such as the peak expected dose for compliance [2]. Two performance-mean-based sensitivity measures, $\partial\mu_Y / \partial\mu_{X_i}$ and $\partial\mu_Y / \partial\sigma_{X_i}$, have been proposed in the past in [3] for importance analysis for HLW applications in which components of the repository are artificially neutralized to identify important components. However, applicability of these measures has not been established in the context of sensitivity analysis.

This paper summarizes the development and application of these two mean-based sensitivity measures. Details of the development of these measures in conjunction with the cumulative distribution function (CDF)-based sensitivity analysis method and their comparison with the previously developed [4] and implemented [1] sensitivity measures is a subject of a future paper. In the following sections, we present a very brief description of the processes involved in the performance assessment model, a brief description of the mean-based sensitivity measures, and the results from the application of these measures to the NRC performance assessment model.

THE PERFORMANCE ASSESSMENT COMPUTER MODEL

Performance assessment models often use a probabilistic approach to propagate uncertainties (sometimes variability) in model parameters, conceptual models, and future system states (i.e., scenario classes). A probabilistic model, as implemented in the NRC TPA code [5], simulates (at the process level) thermal, hydrological, mechanical, and chemical processes of the repository system. This paper uses only the portion of the TPA code that models the most likely scenario. This scenario involves the degradation of waste package (WP) in which high-level waste is disposed in the engineered barrier system (EBS), the release of radionuclides when the water infiltrating the ground surface contacts exposed spent nuclear fuel, and transports the radionuclides through the partially water-saturated geologic medium beneath the repository and subsequently in the saturated zone to a reasonably maximally exposed individual assumed to be located at 20 km down-gradient of the repository [5]. The TPA code estimates dose from released radionuclides during specified time periods (e.g. regulatory compliance period). Input parameters are sampled from assigned probability distributions using Latin Hypercube Sampling (LHS). The code contains 961 input parameters out of which 330 are sampled from specified distribution functions. Several sampled input parameters are specified to have correlation with other parameters.

SENSITIVITY MEASURES

Based on a reliability sensitivity concept [4], the response CDF is defined as the integral of the joint probability-density-function of the parameters, with a domain of integration that corresponds to the domain of the identified samples. The response CDF sensitivities are then calculated from the derivatives of the probability integral. The derivatives are statistically estimated from the samples and used to identify and rank the importance of the random variables.

The CDF of a performance $Y = Y(X)$ can be represented as:

$$p = F_Y(y_0) = P(Y < y_0) = \int_\Omega \dots \int f_X(x)dx \tag{1}$$

where Ω is the region of X for $Y(X) < y_0$. From Eq. 1, the sensitivity of p with respect to a distribution parameter θ (e.g., mean or standard deviation) can be formulated as:

$$\frac{\partial p / p}{\partial \theta / \theta} = \int_\Omega \dots \int \frac{\theta \partial f_X}{f_X \partial \theta} (\frac{f_X}{p}) dx \tag{2}$$

in which (f_X / p) is the sampling density function that corresponds to the sampling region Ω. By applying Eq. 2 for a number of different percentiles, the sensitivities for the entire CDF of Y can be estimated from random samples. Two CDF sensitivities, the standard-deviation sensitivity, $S_{\sigma_i} = (\partial p / p)/(\partial \sigma_i / \sigma_i)$, and the mean sensitivity, $S_{\mu_i} = (\partial p / p)/(\partial \mu_i / \sigma_i)$, were developed in [4] and implemented in [1]. Parameters μ_i and σ_i are the mean and the standard deviation, respectively, of the random variable X_i.

New Mean Response-Based Sampling Sensitivity Measures

Other sensitivity measures proposed for HLW applications include two performance mean-based measures $\partial \mu_Y / \partial \mu_{X_i}$ and $\partial \mu_Y / \partial \sigma_{X_i}$. The sampling-based methods for estimating these two sensitivities have been derived and a summary is given herein. More detailed derivations will be published in a future paper.

The variable transformation is used to transform X_i to Z_i. This transformation can be expressed as

$$\frac{Z_i - \mu_{Z_i}}{\sigma_{Z_i}} = \Phi^{-1}(F_{X_i}(x_i)) = u_i \tag{3}$$

where Z_i is a normal variable with mean value of $\mu_{Z_i} = 0$ and standard deviation of $\sigma_{Z_i} = 1$. Sensitivities with respect to the original variables can be expressed as:

$$\frac{\partial \mu_Y}{\partial \mu_{X_i}} = \frac{\partial \mu_Y}{\partial \mu_{Z_i}} \cdot \frac{\partial \mu_{Z_i}}{\partial \mu_{X_i}} \tag{4}$$

$$\frac{\partial \mu_Y}{\partial \sigma_{X_i}} = \frac{\partial \mu_Y}{\partial \sigma_{Z_i}} \cdot \frac{\partial \sigma_{Z_i}}{\partial \sigma_{X_i}} \tag{5}$$

In Eqs. 3-4, $\partial \mu_{Z_i} / \partial \mu_{X_i}$ and $\partial \sigma_{Z_i} / \partial \sigma_{X_i}$ are calculated numerically or analytically based on Eq. 3. The sensitivities $\partial \mu_Y / \partial \mu_{Z_i}$ and $\partial \mu_Y / \partial \sigma_{Z_i}$ are calculated from the random samples as described below.

$\partial \mu_Y / \partial \mu_{Z_i}$ *Sensitivity from Random Samples*

After the transformation using Eq. 3, the mean value of Y is:

$$\mu_Y = \int Y \phi_u (\boldsymbol{u}, \mu_Z, \sigma_Z) d\boldsymbol{u} \tag{6}$$

in which ϕ_u is the joint standard normal pdf. The mean-based sensitivity is (several intermediate steps are not presented):

$$S_{Y_\mu} = \frac{\partial \mu_Y}{\partial \mu_{Z_i}} = \int_{All\ u} Y(\boldsymbol{u}) \frac{\partial \phi(,\mu_Z, \sigma_Z)}{\partial \mu_{Z_i}} d\boldsymbol{u} = E[u_i Y(\boldsymbol{u})] \tag{7}$$

To distinguish if the sensitivity is statistically significant or not, we can test the hypothesis that $S_{Y_\mu} = 0$ and develop the acceptance limits. The test statistics is

$$Z_o = \frac{\overline{S}_{Y_\mu} - S_{Y_\mu}(=0)}{\sigma_{\overline{S}_{Y_\mu}}} \tag{8}$$

in which the sampling estimate is

$$\overline{S}_{Y_\mu} = \frac{1}{k}\sum_{j=1}^{k}[u_i Y]_j \tag{9}$$

Using normal distribution approximation, justified for sufficiently large k based on the central limit theorem, the following probability statement can be made:

$$P\left[-Z_{\alpha/2} \le \frac{\overline{S}_{Y_\mu} - S_{Y_\mu}}{\sqrt{E\left[Y^2\right]/k}} \le Z_{\alpha/2}\right] \le 1-\alpha \tag{10}$$

where $E[Y^2]$ can be estimated using the Monte Carlo or LHS samples. α is the significant probability level or the risk of making a wrong conclusion about the null hypothesis that u is unrelated to the performance Y and has zero sensitivity.

$\partial\mu_Y / \partial\sigma_{Z_i}$ *Sensitivity from Random Samples*

The mean-based sensitivity is:

$$S_{Y_\sigma} = \frac{\partial\mu_Y}{\partial\sigma_{Z_i}} = \int Y(\boldsymbol{u})\frac{\partial\phi(\boldsymbol{u},\mu_Z,\sigma_Z)}{\partial\sigma_{Z_i}}d\boldsymbol{u} = E\left[(u_i^2-1)Y(\boldsymbol{u})\right] \tag{11}$$

To test the hypothesis that $S_{Y_\sigma} = 0$, the test statistics is

$$Z_o = \frac{\overline{S}_{Y_\sigma} - S_{Y_\sigma}(=0)}{\sigma_{\overline{S}_{Y_\sigma}}} \tag{12}$$

in which the sampling estimate is

$$\overline{S}_{Y_\sigma} = \frac{1}{k}\sum_{j=1}^{k}\left[(u_i^2-1)Y\right]_j \tag{13}$$

Using the normal distribution approximation, the following probability statement can be made:

$$P\left[-Z_{\alpha/2} \le \frac{\overline{S}_{Y_\mu} - S_{Y_\mu}}{\sqrt{2\cdot E\left[Y^2\right]/k}} \le Z_{\alpha/2}\right] \le 1-\alpha \tag{14}$$

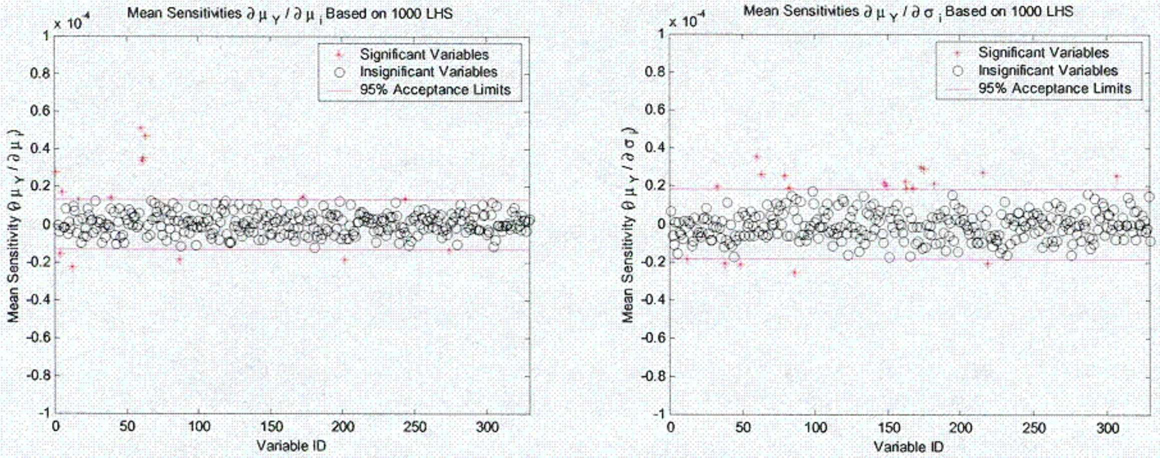
where $E[Y^2]$ can be estimated using the samples.

Acceptance Limits and Adaptive Sampling

If the calculated sensitivities are outside of the acceptance limits defined by Eqs. 10 or 14, we will accept the alternative hypotheses that the sensitivities are greater than zero at the corresponding confidence level. If the calculated point lies well outside of the limits, then the variable is likely to be important. In such cases, the magnitudes of the sensitivities may be used to rank the important variables. The number of samples can be adaptively increased to reduce the sampling error and to identify the important variables and their ranking with confidence.

RESULTS

Figure 1 shows the calculated sensitivities from 1000 LHS samples and the nominal case 10,000-yr compliance period response (peak dose) calculations using the TPA code. Based on $\partial\mu_Y / \partial\mu_{Z_i}$, 15 variables (corresponding to the data that are outside the acceptance limits) are identified as having significant sensitivities at $\alpha = 5\%$. Similarly, based on $\partial\mu_Y / \partial\sigma_{Z_i}$, 20 variables are identified as significant at $\alpha = 5\%$. The identified important variables are listed in table 1. The results show that the two sensitivity measures produce substantially different set of influential variables. But, when these two measures are applied to the previous version of the TPA code, the difference between the two sets of influential variables is small. Therefore, we believe that the difference between these two measures when applied to the latest version of the TPA code is a result of the new process models and the associated parameter ranges. A formal validation study is currently underway to ensure that the differences are logical and justified.



(a) **(b)**

Figure 1. Influential variables identified by (a) S_{Y_μ} and (b) S_{Y_σ} sensitivities (see table 1 for top 10)

Table 1. Top ten random variables identified by S_{Y_μ} and S_{Y_σ} sensitivity

S_{Y_μ} or ($\partial\mu_Y / \partial\mu_{Z_i}$) sensitivity		S_{Y_σ} or ($\partial\mu_Y / \partial\sigma_{Z_i}$) sensitivity	
Rank	Variable Name	Rank	Variable Name
1	WastePackageFlowMultiplicationFactor	1	WastePackageFlowMultiplicationFactor
2	Preexponential_SFDissolutionModel2	2	MatrixKD_UFZ_Ra[m3/kg]
3	DefectiveFractionOfWPs/cell	3	MatrixKD_CHnvPb[m3/kg]
4	SubAreaWetFraction	4	FracturePorosity_TSw_
5	ArealAvgMeanAnnualInfiltrationAtStart[mm/yr]	5	Preexponential_SFDissolutionModel2
6	DripShieldFailureTime[yr]	6	KD_Soil_Se[cm3/g]
7	SFWettedFraction_SEISMO1_7	7	SFWettedFraction_FAULTO
8	MatrixPermeability_TSw_[m2]	8	SFWettedFraction_SEISMO1_6
9	FractionOfCondensateTowardRepository[1/yr]	9	MatrixKD_CHnzTh[m3/kg]
10	FractionOfCondensateRemoved[1/yr]	10	MatrixKD_CHnzU[m3/kg]

The mean sensitivity is expected to stabilize as the number of samples is increased. Figure 2 shows that the ranking convergence seems to become stabilized as the number of samples exceeds about 700.

C01

More samples will be generated to confirm the convergence. Several parameters that are known to have very little significance show up in table 1 (Rank 7 for S_{Y_μ} sensitivity), but this variable drops out as the number of samples is increased from 1000 to 2000. Investigation continues to address this issue.

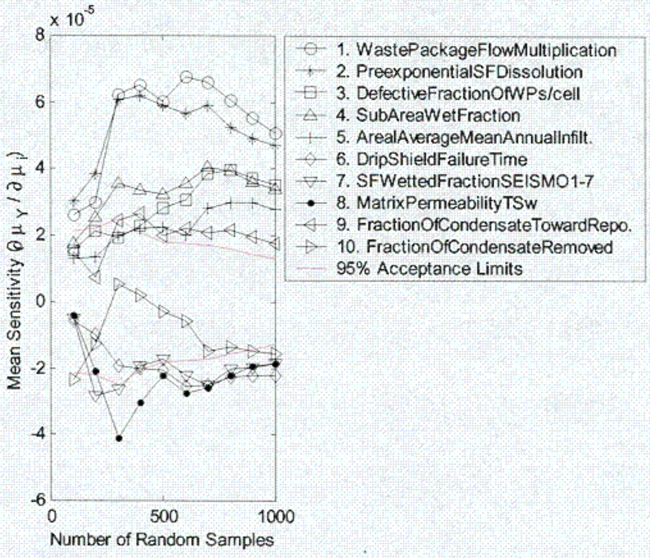


Figure 2. Mean sensitivity of performance to top 10 variables as a function of sample size

CONCLUSIONS

The development and successful implementation of two performance-mean-based sensitivity (or uncertainty importance) measures, $\partial\mu_Y / \partial\mu_{X_i}$ and $\partial\mu_Y / \partial\sigma_{X_i}$, that are particularly relevant to the disposal of HLW regulatory criteria are summarized. Based on $\partial\mu_Y / \partial\mu_{X_i}$ and $\partial\mu_Y / \partial\sigma_{X_i}$ sensitivities, fifteen and twenty out of 330 variables are identified as having significant sensitivities at 95% acceptance limit. Further studies are underway to determine the reason for significant differences in the list of influential variables identified through these two mean-based measures. It appears that 700 samples are sufficient for obtaining stable results at 95% confidence limit for the S_{Y_μ} sensitivity.

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.

REFERENCES

[1] Mohanty, S. and Y-T. (Justin) Wu. 2001. CDF sensitivity analysis technique for ranking influential parameters in the performance assessment of the proposed high-level waste repository at Yucca Mountain, Nevada, USA, *Reliability Engineering and System Safety*, Vol 73/2, pp 167-176.
[2] U.S. Nuclear Regulatory Commission, 10 CFR Part 19, Disposal of High-Level Radioactive Wastes in a Proposed Geological Repository at Yucca Mountain, Nevada, Part 63, Washington, DC, U.S. Nuclear Regulatory Commission, 2001.
[3] Eisenberg NA, Sagar B. Importance Measures for Nuclear Waste Repositories. Reliability Eng. Safety. 70, 217-239: 2000.

C02

[4] Wu Y-T. Computational methods for efficient structural reliability and reliability sensitivity analysis. AIAA J. 32(8), 1717–1723: 1994.
[5] 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.