

**Application of a Variance-Based Sensitivity Analysis
to the Output of a Complex Computer Program**

Abbreviated Title:

App Variance Sensitivity

by

Randall D. Manteufel

Senior Research Engineer

Center for Nuclear Waste Regulatory Analyses

Southwest Research Institute

6220 Culebra Road

San Antonio, TX 78238-5166

210/522-5250

Abstract

The most important input parameters in a complex computer program are identified using a variance-based sensitivity method. The variance-based method differs from regression-based methods in that it does not require assumptions about the functional form relating input and output parameters. The variance-based method is used with the Latin Hypercube Sampling (LHS) scheme where the program is run using numerous realizations of the randomly matched matrix of input parameters. Importance measures are based on variance ratios such as the F-statistic. Based on the magnitude of the importance statistic, a set of important parameters is identified. The computer program is run repetitively to select additional parameters and reduce the variance of the output. Conditional complementary cumulative distribution functions derived by fixing important parameters are used to quantify the reduction in the output variability. Finally, a set of verification runs is performed to demonstrate the degree to which the important parameters truly control the mean and variance of the output. The variance-based method has the advantages of being tightly integrated with the LHS scheme, and not requiring assumptions about the functional relationship between input and output parameters. However, it has the drawbacks of requiring heuristic assessments of threshold variance ratios above which a parameter is considered important, and it also requires numerous executions of the computer program, which may be computationally expensive. For programs that require only a short time to run, the generation of conditional distributions are worthwhile and can enhance a user's understanding of the results from a complex computer program.

Key Words: Sensitivity/uncertainty analysis, importance analysis, probabilistic performance assessment.

Introduction

The problem of interest is illustrated in Figure 1. A computer program simulates a real complex system such as a nuclear power plant during accident scenarios⁽¹⁾ or a nuclear high-level waste repository over long time periods^(2,3). The response of the system, y , is dictated by the input parameters, x 's.

$$y = y(x_1, x_2, \dots, x_n) \quad (1)$$

In general, the input parameters have some uncertainty or range of plausible values. One typically has knowledge about the mean and the error bands associated with a parameter. The variability or uncertainty in each parameter is described using a probability distribution function (PDF) or a combination of them with a joint PDF. In general, a variety of PDFs describe the location and spread of each input parameter. The PDFs can be sampled to obtain specific values of each input. Having obtained discrete values for all of the input parameters, the output can then be calculated using the computer program. Typically, this process is performed in a probabilistic manner in which the x 's are selected from their PDFs and used to generate a set of y predictions.

$$y_k = y(x_{k,1}, x_{k,2}, \dots, x_{k,n}) \quad (2)$$

where $1 \leq k \leq$ number of runs. The statistics such as the mean, variance, and skewness, can be used to describe the set of y 's. If needed, a histogram or a PDF can be established for the output. The PDF for the output provides information about the location and range of estimated y 's. These descriptions are frequently sought by analysts.

For complex computer programs, it frequently is not obvious which parameters are

most important in dictating the magnitude and spread of the outputs. In many cases, a complex computer program may require hundreds of input parameters, yet only a relatively small subset of the inputs strongly influence the output. It is often the analyst's objective to identify which parameters are most important in the problem. There are many reasons for seeking the most important parameters, including comparing results between programs modeling the same problem, guiding the design of the system being modeled to reduce the vulnerability to particular phenomena, and guiding future work in specific areas to better quantify important yet poorly characterized phenomena.

Overall, there are two main types of approaches to identifying important parameters—regression- and variance-based methods. The regression-based methods have been used extensively throughout the literature^(4,5,6) while variance-based methods are relatively new⁽⁷⁾. In general, the regression methods are based on establishing functional relationships between the input and output parameters. Frequently, linear relationships are sought, yet nonlinear and coupled relationships can also be used. For linear relationships, the slope between each input and the output is used to gauge the importance of each input parameter. The slope can be multiplied by the range, or standard deviation, of the input and divided by the overall variability, or standard deviation, of the output. This ratio equals the standardized regression coefficient (SRC)⁽⁴⁾. An input parameter is important if it has a large SRC, hence, is responsible for contributing to a large variability of the output.

One drawback of regression-based methods is that an algebraic function relating inputs and output must be developed. This approach is very successful when couplings and interactions between inputs are minimal, and the inputs are independently and linearly related

to the output. As interactions become stronger, it becomes more critical to specify an appropriate functional form relating inputs and output. In an attempt to relax the need to specify the form of the relationship, McKay⁽⁷⁾ has developed a variance-based sensitivity method. The method has the benefits of not requiring assumptions about the type of relationship between input and output. In addition, it is integrated with the LHS scheme^(8,9) and readily quantifies the effects of inputs on the output. A potential drawback of the variance-based method, however, is that it requires numerous executions of the computer program. In contrast, regression-based methods can be performed with fewer program computations yet require more elaborate auxiliary computations to establish the regression.

A stepwise multilinear regression analysis⁽¹⁰⁾ was used in a recently completed Iterative Performance Assessment (IPA) exercise for the proposed high-level waste (HLW) repository at Yucca Mountain, Nevada⁽³⁾. The computer program used in this paper and in the earlier IPA exercise is the Total-System Performance Assessment (TPA) code⁽¹¹⁾. The TPA code uses the LHS scheme with a total of 195 input parameters. The program predicts the release and transport of radionuclides from the proposed geologic HLW repository at Yucca Mountain, Nevada. The output predicted is the summed normalized release of radionuclides over the next 10,000 yr. The code tracks 20 radionuclides and predicts the time and quantity of release. In 40 CFR 191 (which is expected to be revised in the near future), release limits are given for each radionuclide which are used to normalize the predicted releases. The sum of individual normalized releases is tabulated and compared with regulatory limits. The details of the code are not of primary interest in this paper and are documented elsewhere⁽³⁾. The primary objective of this paper was to apply the variance-

based method and compare the results with regression-based results. For comparison purposes, both variance- and regression-based calculations were performed using the TPA computer program. Because the TPA code was slightly different from the one used in the IPA exercise, both variance- and regression-based calculations are performed using the same computer generated output.

Theory

The variance sensitivity method has the objective of identifying a subset of input parameters which most strongly drive the output. The uncertainty in each of the inputs contributes to the spread of the outputs. At the top of Figure 2, the spread of the output is characterized using either the PDF, Cumulative Distribution Function (CDF), or Complementary Cumulative Distribution Function (CCDF). The CCDF is used frequently in this paper. The spread is characterized by the slopes of the CDF and CCDF. The idea behind variance-based sensitivity is to identify the parameters that, when held constant, significantly reduce the spread of the output. In the bottom of Figure 2, a hypothetical distribution of outputs is shown with smaller variance, hence narrower PDF and steeper CDF and CCDF. The problem is then to identify the inputs such that if their variability is significantly reduced, then the variability of the output will be significantly reduced. If a parameter is identified as being important based on output variability, then it follows that significant changes in its mean value will lead to significant changes in the mean of the output. For analysis purposes, it is convenient to focus on output variability, noting that important parameters strongly affect both the mean and variance of the output.

Alternatively, one can state the problem as identifying inputs that have minimal

influence on output variability, hence, are not important. It appears more readily grasped to search for a positive declaration about a parameters importance than a negative.

Scatter Plots

Scatter plots are a simple and informative tool to visually investigate the relationship between any one input parameter and the output. Typically, the analysis will have generated a large set of valid inputs which are used to compute an equal number of outputs. A plot is constructed of the output y versus a single input parameter x . Each model evaluation or computer code run is represented as a single point. Depending on the distribution of points, one may identify a relationship between the input parameter and the output. If no pattern exists, then the scatter plot will appear as a cloud of uncorrelated points.

In Figure 3, two scatter plots are shown from this work. The two parameters, $infil$ and $akr2$, are plotted against the output, normalized release. Each scatter plot contains 2,000 points representing 40 distinct LHS-50 runs. The term LHS-50 describes how each input parameter is discretized into 50 equal probability bins, and the mean value in each bin is used once and only once in a LHS run. In total, 50 executions of the computer program (runs) are performed for an LHS-50, regardless of the number of input parameters. Each run is based on a vector of inputs (equal in length to the number of parameters) determined by randomly matching binned input parameters. In this work, 40 different matchings of input parameters were used, hence, 40 distinct LHS-50 runs. As a result, the scatter plots consist of 40 points along 50 vertical lines. The first parameter, $infil$, describes the deep percolation of infiltrating meteoric water at the repository site. The second parameter, $akr2$, describes the gaseous fracture permeability of the Paintbrush hydrostratigraphic unit at the site. The

phenomena being described by each parameter is not important for this paper, only that infil has a visual correlation with normalized release while $akr2$ does not. One also notices that the normalized release is plotted on a log scale because it ranges over 5 orders-of-magnitude. A few of the outputs had zero value, and are set equal to 2×10^{-4} in order to be plotted.

Because the output values ranged over 5 orders-of-magnitude and contained a few zero values, the output was transformed in order to do either a variance- or regression-based analysis⁽¹²⁾. A linear rank transformation was used in which the outputs from all of the runs are ordered and replaced by a set of steadily increasing values. In this problem, the 40 LHS-50 generated 2,000 outputs, and the lowest output was transformed to $1/2,000$, the second lowest to $2/2,000$, and so forth up to the highest output being transformed to 1. This generated a uniform distribution in the output values of y .

In Figure 4(a), the scatter plot for the rank transformed output and infil are shown. The rank transformed data continue to visually show the trend in the data. The mean of the 40 values for each of the 50 bins is shown in Figure 4(b), as well as the standard deviations about the means in (c). The bin means show much more clearly the correlation between input and output. Figure 4 is also used to introduce the variance-based method.

Variance-Based Importance

The main goal behind the variance-based method is to determine what portion of the total output variance is explained by a trend through the means or is unexplained due to residual uncertainty attributed to other parameters. This appears to be an idea consistent with correlation and regression. A well-known variance identity used in analysis of variance relates the total variability of y to the variability between the bin means (explained) and

variability within the bins (unexplained)^(7,10,14). For this work, the identity is expressed as:

$$\sum_{j=1}^{Nrep} \sum_{i=1}^{Nvec} (y_{ij} - \bar{y}_{..})^2 = Nrep \sum_{i=1}^{Nvec} (\bar{y}_i - \bar{y}_{..})^2 + \sum_{i=1}^{Nvec} \sum_{j=1}^{Nrep} (y_{ij} - \bar{y}_i)^2 \quad (3)$$

where $Nvec$ =number of LHS vectors or bins (=50 in this work), and $Nrep$ =number of repetitions (=40 in this work). In words, Equation 3 states that the total variability equals the variability of bin means plus the mean of bin variabilities.

The average over all output y 's is used in Equation 3 and defined as:

$$\bar{y}_{..} = \frac{\sum_{j=1}^{Nrep} \sum_{i=1}^{Nvec} y_{ij}}{Nrep \cdot Nvec} \quad (4)$$

The average over all y 's in one LHS bin is used in Equation 3 and defined as:

$$\bar{y}_i = \frac{\sum_{j=1}^{Nrep} y_{ij}}{Nrep} \quad (5)$$

This variance identity is used to develop importance indices.

Importance Indices: R^2 , R_a^2 , F

If the input parameter has a negligible influence on the output, then the variability of bin means will be relatively small and the mean bin variability will be relatively large.

Conversely, if the input has a strong influence on the output, then the variability of bin means will be relatively large. This suggests an importance index which is the ratio of variances

This ratio is analogous to the coefficient of determination which is commonly denoted

$$R^2 = \frac{N_{\text{rep}} \sum_{i=1}^{N_{\text{vec}}} (\bar{y}_i - \bar{y}_{..})^2}{\sum_{j=1}^{N_{\text{rep}}} \sum_{i=1}^{N_{\text{vec}}} (y_{ij} - \bar{y}_{..})^2} \quad (6)$$

as R^2 , except Equation 6 is derived from an analysis of variance instead of a continuous regression-based estimate. The theoretical underpinings of an analysis of variance can be related to a regression analysis⁽¹⁶⁾; however, it is preferable to distinguish between variance- and regression-based methods. The primary difference is that the variance-based method is indifferent to the ordering of bins, whereas the ordering is important for a regression (or curve-fitting) approach. From Figure 4, one notes that the variability of bin means will be large because there is a trend between the input and output. As the trend becomes more pronounced, then R^2 will increase in magnitude.

If no trend exists, then the bin means will uniformly lie about a mean of one-half. A result of the central limit theorem in statistics⁽¹³⁾, is that the variance of bin means is equal to the total variance divided by N_{rep} , provided each bin contains random samples from the same population (i.e., no bin-to-bin variation for either the mean or variance). The explained variance is slightly over-estimated by including this variance attributed to a finite number of repetitions. Hence, an improved importance index has been suggested⁽⁷⁾ as an alternative variance ratio.

In words, Equation 7 is an improved estimate of the variance of bin means divided by the total variance. If the input is found to strongly affect the output, then the alternative variance ratio increases.

$$R_a^2 = \frac{\text{Nrep} \sum_{i=1}^{\text{Nvec}} (\bar{y}_i - \bar{y}_{..})^2 - \frac{\sum_{i=1}^{\text{Nvec}} \sum_{j=1}^{\text{Nrep}} (y_{ij} - \bar{y}_i)^2}{\text{Nrep}}}{\sum_{j=1}^{\text{Nrep}} \sum_{i=1}^{\text{Nvec}} (y_{ij} - \bar{y}_{..})^2} \quad (7)$$

Finally, the F-statistic can be used as an importance index based on a fixed-effects, one-way analysis of variance between input and output^(14,15). The 50 LHS bins act as distinct levels within which 40 samples are collected. The analysis determines if there is a statistically significant difference between the means of the distributions in each of the bins. In statistical terms, a null hypothesis is formed stating that the bin means are equal, $H_0: \mu_1 = \mu_2 = \dots = \mu_{\text{Nbin}}$. If the null hypothesis is true, then the input has a negligible effect on the output. Alternatively, data may indicate that all of the bin means are not equal, hence, the null hypothesis is rejected and the input is identified as being important.

In Figure 4(b), the bin means are shown for the infil parameter. Visually, one can identify a distinct difference among bin means. The F-statistic is used as an importance measure in an analysis of variance

$$F = \frac{\frac{\text{Nrep} \sum_{i=1}^{\text{Nvec}} (\bar{y}_i - \bar{y}_{..})^2}{(\text{Nvec} - 1)}}{\frac{\sum_{i=1}^{\text{Nvec}} \sum_{j=1}^{\text{Nrep}} (y_{ij} - \bar{y}_i)^2}{\text{Nvec} (\text{Nrep} - 1)}} \quad (8)$$

Both the numerator and denominator of the F-statistic are estimates of the total variance of y ,

assuming no trend exists between input and output. If a trend exists, then the numerator will overestimate the variance. Hence, the magnitude of F will increase.

It is noteworthy that the one-way analysis of variance and the resulting F-statistic are based on a fixed effects model⁽¹⁴⁾ which assumes all measurements are independent, random samples, drawn from normal distributions that have equal variances. The assumption of independent random samples is not true for this work. For a simple random sampling (sometimes called pure Monte Carlo), the y 's are independent. However, for an LHS scheme (stratified sampling without replacement), the set of x 's are not independent; hence, the y 's are not completely independent (see Appendix of Reference No. 8). The covariance between the y 's is not zero, yet predicting the covariances appears to be an insurmountable task for a large, complex computer program. The covariance is related to the complexity of the functional relationship between the inputs and output. From experience, the LHS can perform no worse than the simple random sampling, especially on problems where the input parameters have significant interactions. If minimal interactions exist, then the LHS can perform much better than simple random sampling. In general, the F-statistic generated using LHS will be larger than if it was generated using simple random sampling. Because the observations are derived using the LHS scheme, they are not independent and a heuristic cutoff value for F needs to be established that will be larger than a theoretically derived cutoff based on completely independent outputs.

Application Problem

The variance-based method was applied using a computer program that evaluates the performance of the proposed HLW repository at Yucca Mountain, Nevada^(3,11). The primary

purpose of the program is to predict the release of radionuclides and thereby assess compliance or exceedance of regulatory limits. As implemented in a recently completed exercise, the performance is measured using the U.S. Environmental Protection Agency (EPA) release limits to the accessible environment which are described in the remanded 40 CFR Part 191. The boundary of the accessible environment is the ground surface or at a 5 km radius from the repository. The normalized release is summed over the next 10,000 yr. The program uses the LHS scheme with 195 input parameters for the base case (neglecting external disruptive scenarios). More details are available in (3). A goal of this application problem was to identify the top ten parameters using a variance-based method and compare these results with regression-based results.

Comparing Importance Indices

The first step was to complete a set of LHS-50 runs. The number of LHS-50 runs was increased from 8 to 16, 24, 32, and finally 40. As more runs were completed, the importance indices were computed using all of the available data for all of the 195 input parameters.

In Figure 5, the importance indices are compared for the top ten parameters. In each of the three plots, the abscissa represents the quantity of output information and the ordinate represents the value of the importance index for various input parameters. The importance of a parameter is judged by a large value of the importance index, typically one exceeding a threshold. As the amount of output information increases, one would expect the magnitude of the importance index to increase in exceedance of a cutoff, thereby indicating increased statistical confidence that a parameter is truly important. Only one importance index has such

a property, the F-statistic. Both the R^2 and R_a^2 decrease with increasing output information, which is a counter-intuitive result. One would expect the importance index to provide a stronger indication of importance as more information becomes available. The R^2 index decreases nearly monotonically while the R_a^2 primarily decreases with increasing runs. Hence, we selected the F-statistic as the importance index to use.

After viewing the scatter plots of all input parameters, we heuristically selected a F-statistic cutoff value of 3.0 above which a parameter was identified as being important. The value of 3.0 was selected primarily because there is a natural break in the data with approximately 5 variables performing noticeably different than the other 190 variables. Heuristic selection of a cutoff value is suggested by McKay⁽⁷⁾. In the traditional one-way analysis of variance^(13,14,15), a theoretically derivable cutoff for a 1 or 5 percent level of significance can be derived. Embedded in the standard statistical one-way analysis of variance hypothesis testing are assumptions that each of the data points are independent, as well as being drawn from normal distributions with equal variances. Because we are using the LHS scheme, the data points are not completely independent.

Top Ten Parameters

Based on the initial 40 LHS-50 runs, five parameters were identified as being important. These parameters are the top five identified in Figures 5(c). In Figure 6, rank ordered plots show the F-statistic computed for each parameter. The parameters are sorted by F-statistic so that the first has the largest F. The F-statistic is plotted against the sorted (or ranked) order. These plots help identify individual or groups of important parameters. In Figure 6(a), the infil parameter is clearly important, as noted by the large F-statistic which

confirms the trend observed in the scatter plots in Figure 4.

After identifying at least one important parameter, McKay⁽⁷⁾ outlines a sequence of procedures in which additional parameters are found. The procedure is summarized by constructing sets of potentially important parameters. The sets of parameters are constructed by adding a previously unimportant parameter to the set of the previous step. The important parameters are set to a fixed value or to several specific values. The set of new Nrep by LHS-Nvec runs are completed and the variance ratios compared. A set of important parameters is thus increased by at least one at each cycle. The procedure is similar to step-up parameter selection described in regression-based methods⁽⁷⁾.

For this work, we selected five parameters to be important from the first set of runs. A second set of runs was completed with each of the five important parameters set to their mean values. Thus we were searching for additional input parameters which are important over the range of other parameters, not necessarily only at high or low values of the output. If parameters which are important for high values of output were the only ones sought, then one might select fixed values at varying quantile levels in their respective CDFs.

One can also select a few values of the important parameters based, perhaps, on a coarse LHS discretization. For example, an LHS-5 can be used on the important parameters while an LHS-50 can be used on the remaining parameters. In this case, many more evaluations of the computer program are required. Because our program required a long time to complete calculations, it was prohibitive to explore more than the mean of the previously selected parameters.

In Figure 6(b), the results of the second set of runs indicated that three more

parameters are important. For the last set of runs, a total of eight parameters were set to their mean values and a new set of runs initiated. Based on our experience, if the F-statistic of a parameter exceeded a cutoff, then as Nrep increased it would only continue to increase in exceedance of the cutoff [see Figure 5(c)]. Hence, the last set of runs was terminated as soon as two parameters exceeded the cutoff so that a total of ten parameters were identified as important. These parameters are: infil, forwar1, ecorr6, ecorr7, rdiffl1, ecorr3, ecorr2, ecorr8, ecorr5, sol4Am.

In Figure 7, conditional CCDFs are plotted where the outputs have been conditioned by holding either 0, 5, or 10 input parameters fixed. Each set of curves is based on a set of 40 LHS-50 runs, and contains five curves derived by grouping eight independent LHS-50 runs into an equivalent LHS-400. The grouping of results was selected for historical reasons to facilitate comparison with earlier work⁽³⁾. As expected, Figure 7 shows that the variance of the output is reduced as the important parameters are fixed.

The reduction in output variability can also be quantified. For each set of five equivalent LHS-400 curves, a single equivalent LHS-2000 curve was constructed. Based on this curve, the output value was determined for 95 percent of the distribution. Thus the range of output was measured between the 2.5 and 97.5 percent probabilities. For the 0 fixed case, the CCDF equals 97.5 percent at an output of 0.02 and 2.5 percent at 18.0. Thus the output initially ranged over 2.95 orders-of-magnitude or equivalently by a factor of 890. By selecting and fixing the top 5 parameters, the range of the CCDF was decreased so that the output varied from 0.27 to 5.94 which is 1.34 orders-of-magnitude or equivalently a factor of 22.0. Fixing the top ten parameters yielded an output range from 0.29 to 1.86 which is 0.81

orders-of-magnitude or equivalently a factor of 6.5. By fixing the most important parameters to their mean values, the output variability was reduced from a spread of 890 to a spread of 6.5, or by over two orders-of-magnitude.

The mean value to which the output converges as more parameters are fixed is roughly the mean of the initial distribution of outputs. However, this mean value is not of interest. It is the reduction in output variability that is important.

Comparison with Regression Results

A multilinear regression analysis was completed using the original rank transformed 40 LHS-50 runs with no parameters fixed. A stepwise procedure was employed where each parameter was independently regressed with the output data. The parameter with the largest coefficient of determination, R^2 , was selected as an important parameter. The next step added one previously unimportant parameter and regressed. The set of parameters with the largest R^2 is selected as the new set of important parameters. The p-value for each regression coefficient is checked to determine if it is above a 5 percent threshold. The p-value is the probability of observing a nonzero regression coefficient from the finite sample when the true (population) coefficient is actually zero. If the p-value exceeds a significance level of 5 percent, then the parameter was excluded from the important set. In each step, one parameter is added to the set of important parameters. The process is stopped for one of a number of reasons: (i) no additional statistically significant parameters can be identified, or (ii) the addition of parameters yields a minimal improvement in R^2 , thus indicating an overfitting of the output data. In this work, the stepwise addition of parameters was stopped when the overall R^2 changed by less than 0.01.

In Table 1, the results of the regression analysis are presented. Fortuitously, the regression analysis terminated with ten parameters due to small incremental improvements in R^2 with the addition of new parameters. This is convenient for comparison purposes. The regression results are very similar to the variance-based results. The top five in both methods are the same, and seven of the parameters are the same in both sets. This agreement between sets of important parameters significantly increases confidence in both methods.

A comparison was made to determine why the variance- and regression-based methods differed in three parameters. It should be noted that the top five parameters control a significant amount of output variability and that other parameters are progressively less important. Scatter plots for the three parameters selected by the regression, yet missed by the variance-based method, are shown in Figure 8. Conversely, those selected by the variance, yet missed by regression, are shown in Figure 9. A trend in any of the scatter plots is difficult to detect. One can detect a trend in the $akr3$ scatter plot [Figure 8(a)]. Low values of $akr3$ tend to yield low normalized releases. Based on Figures 6(a) and 8(a), it was observed that a combination of low $akr3$ and low $infil$ leads to very low outputs.

In Figure 10, four conditional scatter plots for $akr3$ are shown for each of the four quartiles of $infil$. Only for the lowest quartile of $infil$, Figure 10(a), does $akr3$ show a correlation with the output. When $infil$ is larger, the correlation between $akr3$ and the output is overwhelmed by the correlation with $infil$. In Figure 11, conditional scatter plots for $infil$ with four quartiles of $akr3$ demonstrate similar behavior. It is observed that the general trend between $infil$ and the output is not strongly influenced by the value of $akr3$. From Figure 11(a), it is only when both $infil$ and $akr3$ have low values that the output becomes

very small.

It is interesting that *akr3* was the next most significant parameter identified in the variance-based method after the first set of runs [see Figure 6(a)]. If the cutoff value for the F-statistic were lower, then *akr3* would have been selected. In comparison, *ecorr8* was also nearly selected to be important in the variance-based method in the initial set of 40 LHS-50 runs. Although not selected in the first set of runs, *ecorr8* was selected in the second set of runs. An explanation for why *akr3* was not selected is that it is important only at low values of *infil*. From Figures 10(a) and 11(a), the output can be very small when both *akr3* and *infil* are near their minimum values. In the second step of the variance-based method, *infil* was set to its mean value. Thus, the importance of *akr3* was diminished in subsequent runs. Other than this observation, the distinction between the parameters selected by the regression- and variance-based methods appears minimal where both methods selected the same top five parameters.

A final comment on the regression results is that the last few parameters resulted in small improvements in the coefficient of determination which had a maximum final value of 63 percent. This is an indication that the multilinear model was beginning to overfit the data while only explaining 63 percent of the output variability. These results are similar to those from comparable problems reported in the literature. For example, a probabilistic risk assessment for a nuclear power plant was conducted using two independent LHS-200 runs with 161 input parameters⁽⁵⁾. The output was regressed using a multilinear model and typically less than 10 parameters were identified as being important for a single output of interest. Good agreement was achieved for the first few (approximately the top five)

parameters of the two independent LHS-200 runs, however, the last few important parameters varied between runs. In addition, the final coefficient of determination was from 0.6 to 0.7, indicating a large amount of unexplained variability. In theory, one can search for a better regression model, however multilinear models are used predominantly in practice.

Verification Runs

In Figures 12 and 13, conditional CCDFs are shown which serve as a check on selected important parameters. The verification runs are performed to determine if the set of ten parameters truly controlled the output. If the most important parameters control the output, then one would expect that fixing them to high or low values would strongly affect the location of the outputs yet the range of outputs remain narrow for any single run. Alternatively, fixing the less important parameters to different values would not be expected to significantly affect the range or distribution of the output.

In Figure 12, a coarse LHS-5 was applied to the most important parameters and a fine LHS-400 was applied to the less important parameters. The LHS-400 was accomplished by combining eight independent LHS-50 runs. The conditional CCDFs (dashed lines) are compared with the original CCDF (solid line). The original CCDF is based on a combination of the original 40 LHS-50 runs. We note that the conditional CCDFs are rather steep, indicating a narrow range of output values. This is because only the less important parameters are being varied. The location of the conditional CCDFs are dictated by the specific values and combinations of the most important parameters. The most important parameters are noted to significantly affect the location where the conditional CCDFs break.

In Figure 13, similar calculations are presented. Here, the important parameters are

varied while the less important parameters are fixed. A fine LHS-400 is used for the most important parameters and a coarse LHS-5 is used for the less important parameters. The conditional CCDFs lie near the original CCDF, indicating that the range and distribution of the output is being controlled by the most important parameters. One of the conditional CCDFs, however, does vary significantly from the cluster of other curves. The main deviation is due to a number of very small outputs, where 15 percent of the outputs had values smaller than 0.01. After reviewing the input matrices, we attributed this to the parameter $akr3$ being small. Because our interests are more for higher values of the output, this effect at the low values was not explored further. Overall, the conditional CCDFs enhance confidence that the dominant parameters were identified in the variance-based method.

Conclusions

A variance-based method proposed by McKay⁽⁷⁾ for identifying the important input parameters in a complex computer program was investigated. The method represents an alternative approach to regression-based methods. A number of variance-based importance statistics were investigated, and the F-statistic was identified as having the desirable characteristic of becoming more significant as more runs are completed and more information is available with which to gauge an input parameter's significance.

Both a variance- and regression-based method were applied to data generated by a computer program which was recently employed in the performance assessment of a proposed HLW repository at Yucca Mountain, Nevada⁽³⁾. A subset of ten parameters was identified as being important from a total of 195 input parameters. Both methods agreed on

five of the top five, and seven of the top ten parameters. After reviewing scatter plots, an explanation was developed for why the variance-based method missed one apparently important parameter. This was attributed more to the application of the method than to a deficiency of the variance-based method.

The reduction in output variability was quantified as the important parameters were fixed to their median values. The range in output was reduced by over two orders-of-magnitude due to fixing the top ten parameters. A set of confirmation runs were completed in which the important parameters were noted to strongly affect the distribution of outputs, and the remaining 185 less important parameters have much smaller effects.

The variance-based method has potential benefits of not requiring any assumptions about the functional form of relationship between input parameters and the output. In addition, it readily quantifies the reduction in the output variability as input parameters are fixed. Two significant drawbacks of the method are the need to heuristically select critical (cutoff) values of an importance measure and the potentially prohibitive costs associated with the repetitive execution of the computer program.

The choice between variance- and regression-based methods is most probably influenced by the cost associated with computing time. The variance-based method requires many more computer runs than a regression-based method, hence this may prohibit its use in certain cases. Overall, the ideas which motivate the variance-based approach are sound and suggest new avenues for exploring the relationship between the input and output generated by a complex computer program with many parameters. Having identified the set of important parameters, the practice of computing conditional CCDFs based on a combination of

alternating coarse and fine LHS discretizations for important parameters enhances confidence in the selection of important parameters.

Acknowledgments

This report 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-02-93-005. The activities reported here were performed on behalf of the NRC Office of Nuclear Regulatory Research, Division of Regulatory Applications. The report is an independent product of the CNWRA and does not necessarily reflect the views or regulatory position of the NRC.

References

1. Nuclear Regulatory Commission. 1990. *Severe Accident Risks: An Assessment for Five U.S. Nuclear Power Plants*. NUREG-1150. Washington, DC: Nuclear Regulatory Commission.
2. Helton, J.C., J.W. Garner, R.D. McCurley, and D.K. Rudeen. 1991. *Sensitivity Analysis Techniques and Results for Performance Assessment at the Waste Isolation Pilot Plant*. SAND90-7103. Albuquerque, NM: Sandia National Laboratories.
3. Nuclear Regulatory Commission. 1995. *Phase 2 Demonstration of the NRC's Capability to Conduct a Performance Assessment for a High-Level Waste Repository*. NUREG-1464. Washington, DC: Nuclear Regulatory Commission.
4. Iman, R.L., and J.C. Helton. 1985. *A Comparison of Uncertainty and Sensitivity Analysis Techniques for Computer Models*. NUREG/CR-3904. Washington, DC: Nuclear Regulatory Commission.
5. Iman, R.L., and J.C. Helton. 1991. The repeatability of uncertainty and sensitivity analyses for complex probabilistic risk assessments. *Risk Analysis* 21(4): 591-606.
6. Wu, Y.T., A.G. Journel, L.R. Abramson, and P.K. Nair. 1991. *Uncertainty Evaluation Methods for Waste Package Performance Assessment*. NUREG/CR-5639. Washington, DC: Nuclear Regulatory Commission.

7. McKay, M.D. 1995. *Evaluating Prediction Uncertainty*. NUREG/CR-6311. Washington, DC: Nuclear Regulatory Commission.
8. McKay, M.D., W.J. Conover, and R.J. Beckman. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 21: 239-245.
9. Iman, R.L., and M.J. Shortencarier. 1984. *A FORTRAN 77 Program and User's Guide for the Generation of Latin Hypercube and Random Samples for Use with Computer Models*. NUREG/CR-3624. Washington, DC: Nuclear Regulatory Commission.
10. Draper, N.R., and H. Smith. 1966. *Applied Regression Analysis*. New York, NY: John Wiley and Sons.
11. Sagar, B., and R.W. Janetzke. 1993. *Total-System Performance Assessment (TPA) Computer Code: Description of Executive Module, Version 2.0*. CNWRA 93-017. San Antonio, TX: Center for Nuclear Waste Regulatory Analyses.
12. Iman, R.L., and W.J. Conover. 1979. The use of rank transform in regression. *Technometrics* 21: 499-509.
13. Mendenhall, W., R.L. Scheaffer, and D.D. Wackerly. 1981. *Mathematical Statistics with Applications*. Boston, MA: Duxbury Press.
14. Dunn, O.J., and V.A. Clark. 1987. *Applied Statistics: Analysis of Variance and Regression*. New York, NY: John Wiley and Sons.
15. Rice, J.A. 1988. *Mathematical Statistics and Data Analysis*. Pacific Grove, CA: Wadsworth & Brooks.
16. Montgomery, D.C. 1991. *Design and Analysis of Experiments*. New York, NY: John Wiley & Sons.

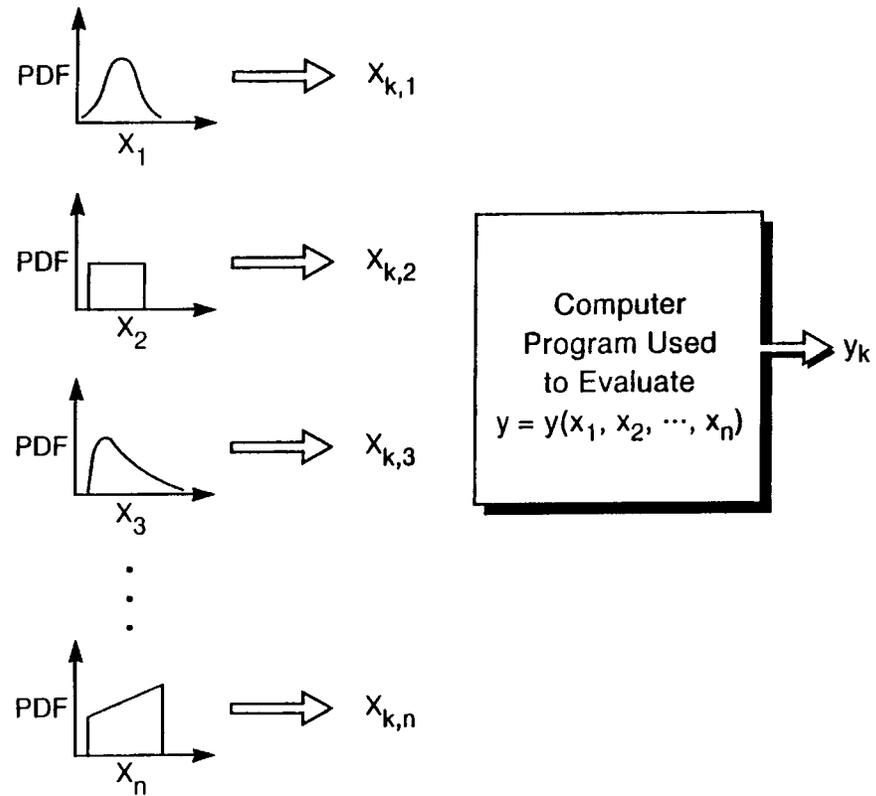
24/39

Table 1. Stepwise multilinear regression analysis results using the original 40 LHS-50 runs.

Parameter	SRC	R ²
infil	0.527	0.28
forwar1	0.283	0.35
rdiff11	0.260	0.42
ecorr7	-0.239	0.48
ecorr6	-0.236	0.54
akr3	0.173	0.57
ecorr8	0.150	0.59
retard3	-0.124	0.61
retard1	-0.111	0.62
ecorr2	0.098	0.63

SRC = standardized regression coefficient based on final stepwise multilinear regression.

R² = coefficient of determination based on regression with all previous parameters.



n = Number of input parameters
 k = One valid realization
 x = Input parameter
 y = Output

Figure 1: Input parameters are sampled from PDFs and used to compute the output.

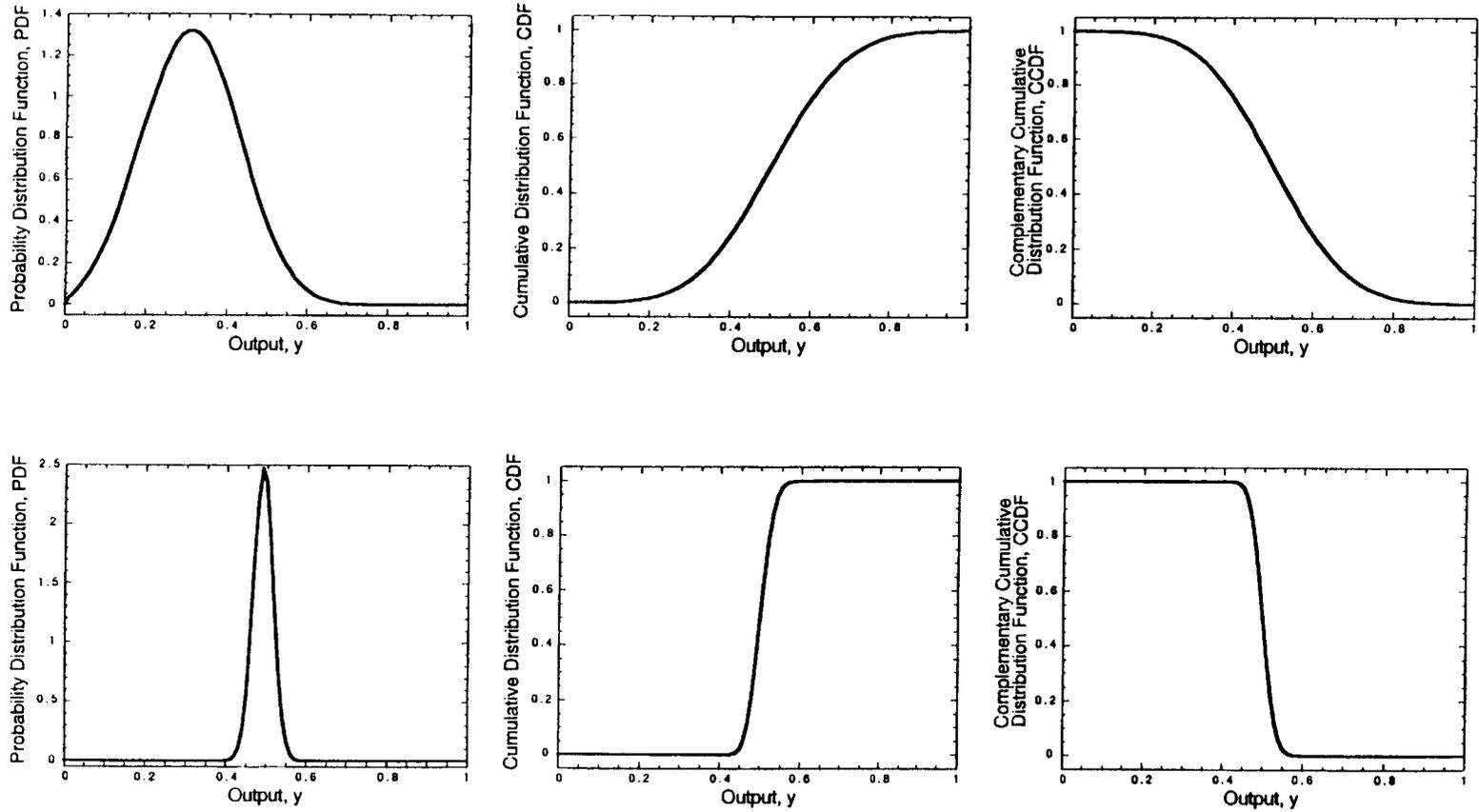
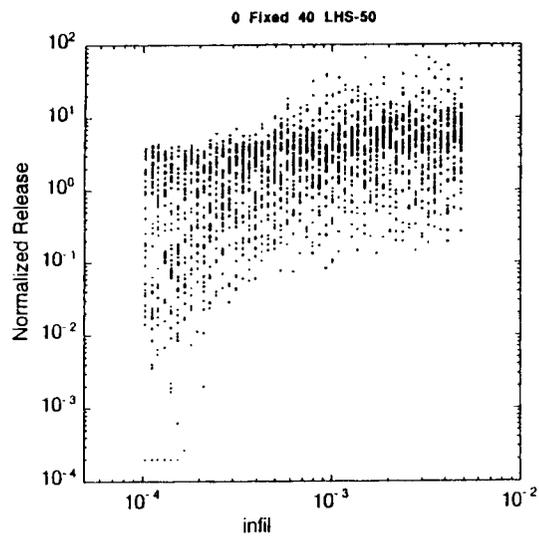
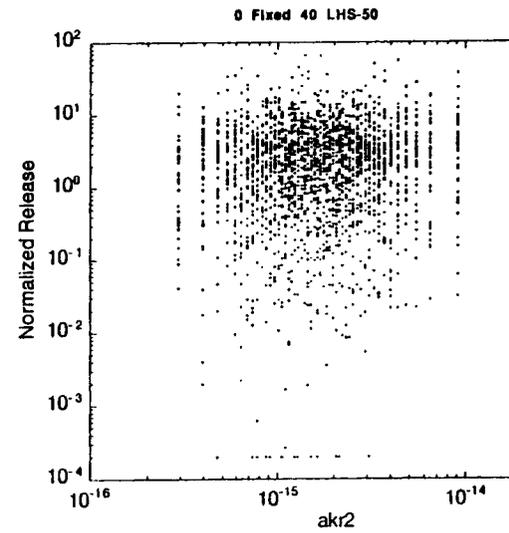


Figure 2: Due to input variability, the output has a distribution described using either a PDF, CDF, or CCDF (top). As the most important input variables are fixed, the range of output decreases (bottom).

28/39



(a)



(b)

Figure 3: Scatter plots showing strong (a) and weak (b) correlation between one of the inputs and the output.

68/ht

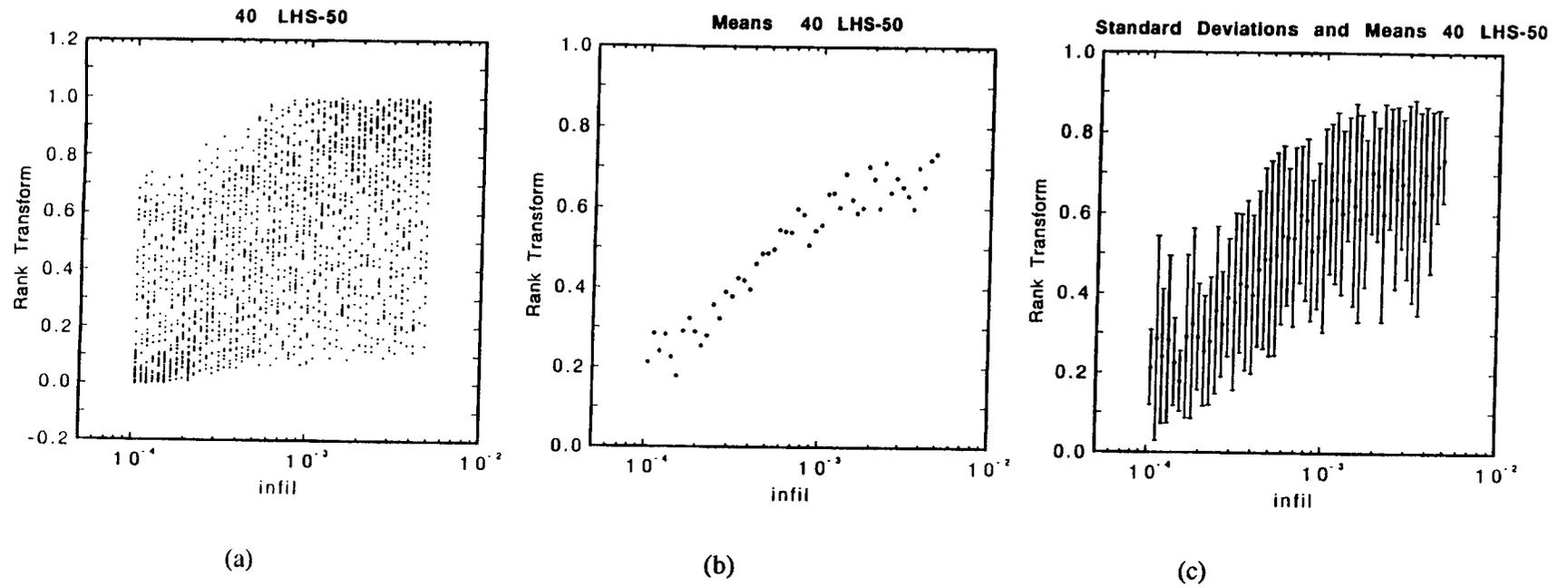
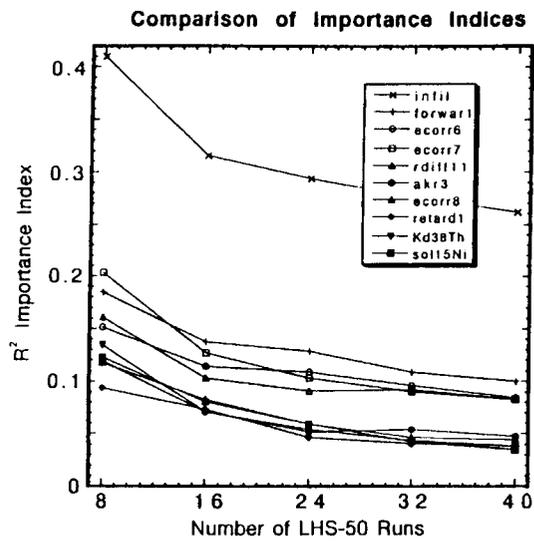
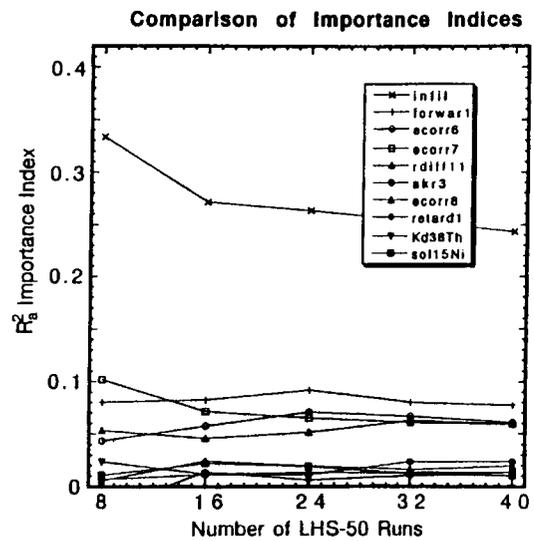


Figure 4: Scatter plot using (a) rank transformed output, (b) means, and (c) standard deviations within each of the 50 LHS bins.

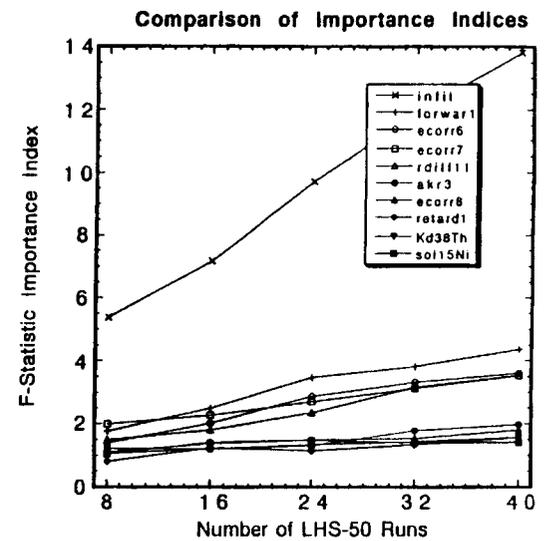
30/39



(a)



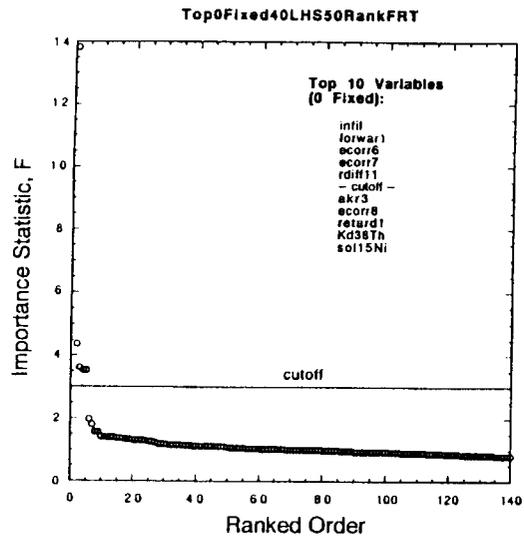
(b)



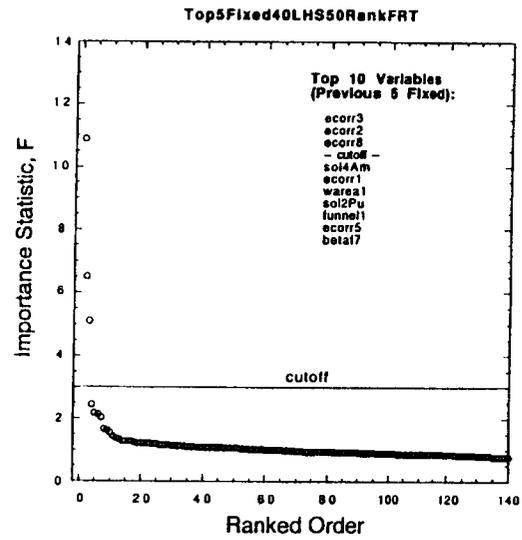
(c)

Figure 5: Comparison of convergence properties of R^2 , R_a^2 , and F importance indices.

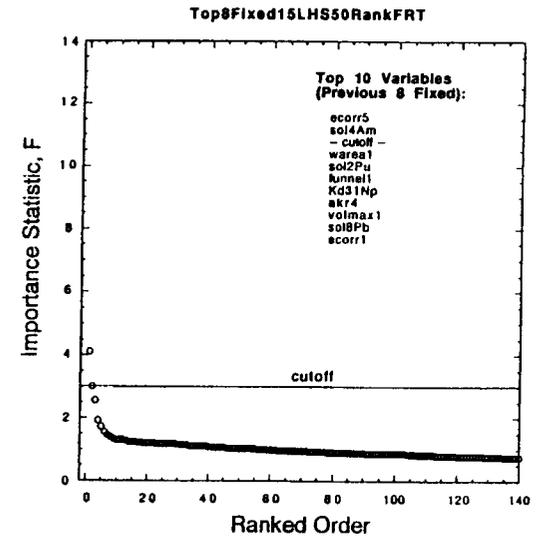
3/39



(a)



(b)



(c)

Figure 6: Ranked order plots used to select top ten variables.

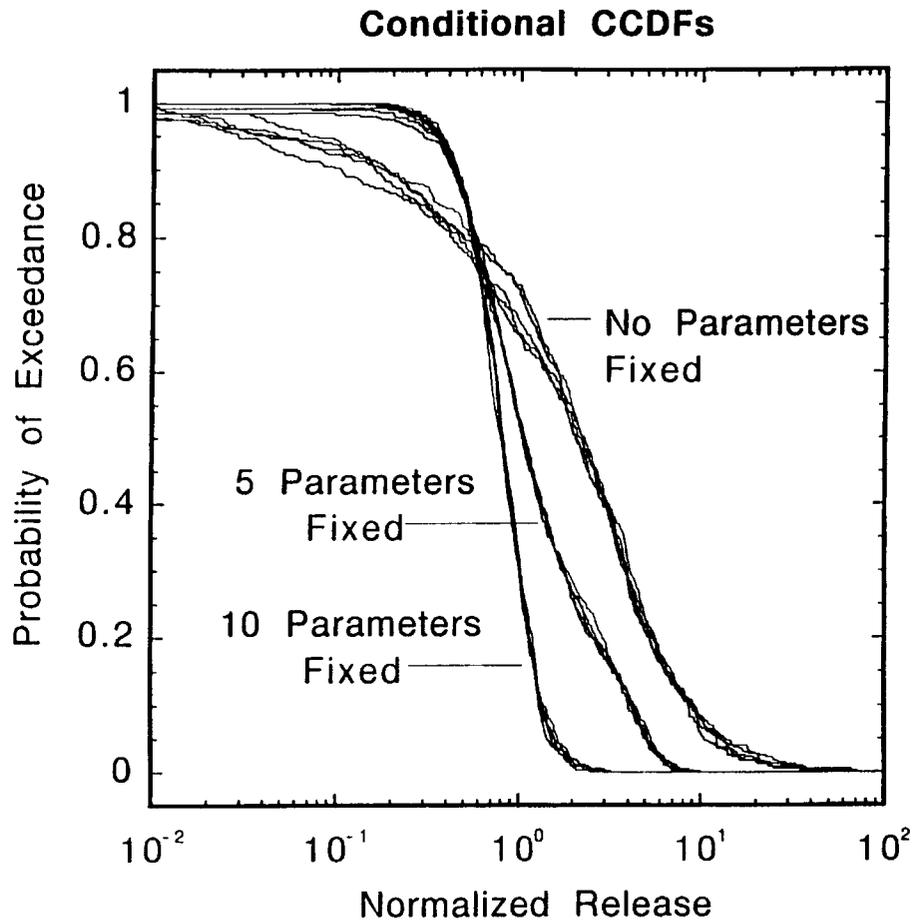
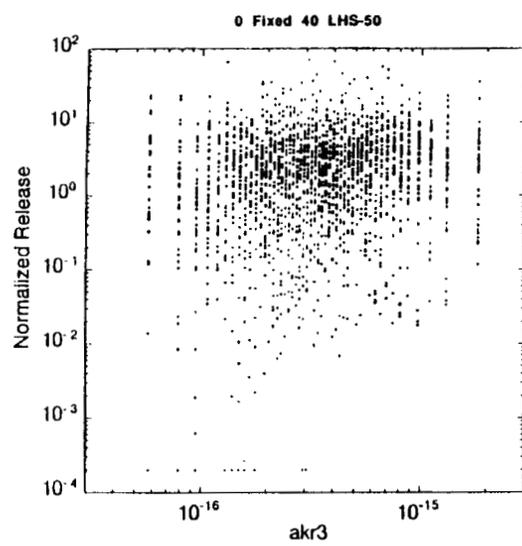
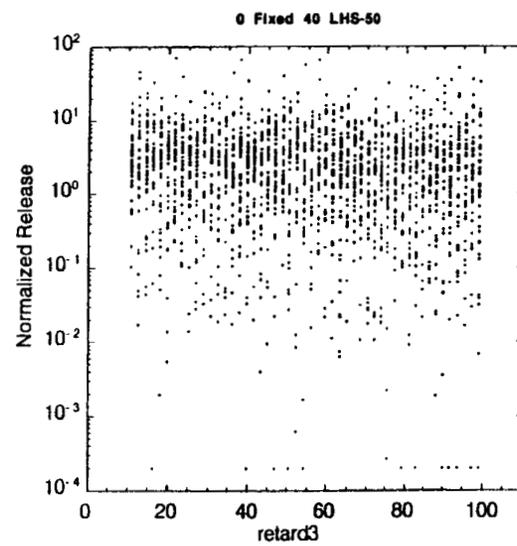


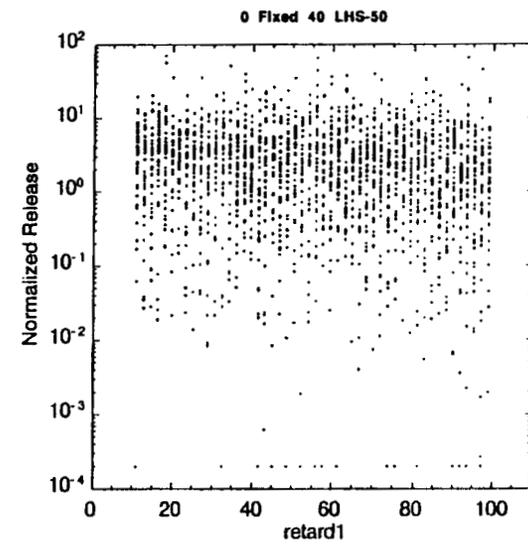
Figure 7: Conditional CCDFs showing how the output variability is reduced by fixing important input parameters.



(a)



(b)



(c)

Figure 8: Scatter plots for parameters identified in the stepwise multilinear regression, yet missed in the variance-based method.

34/39

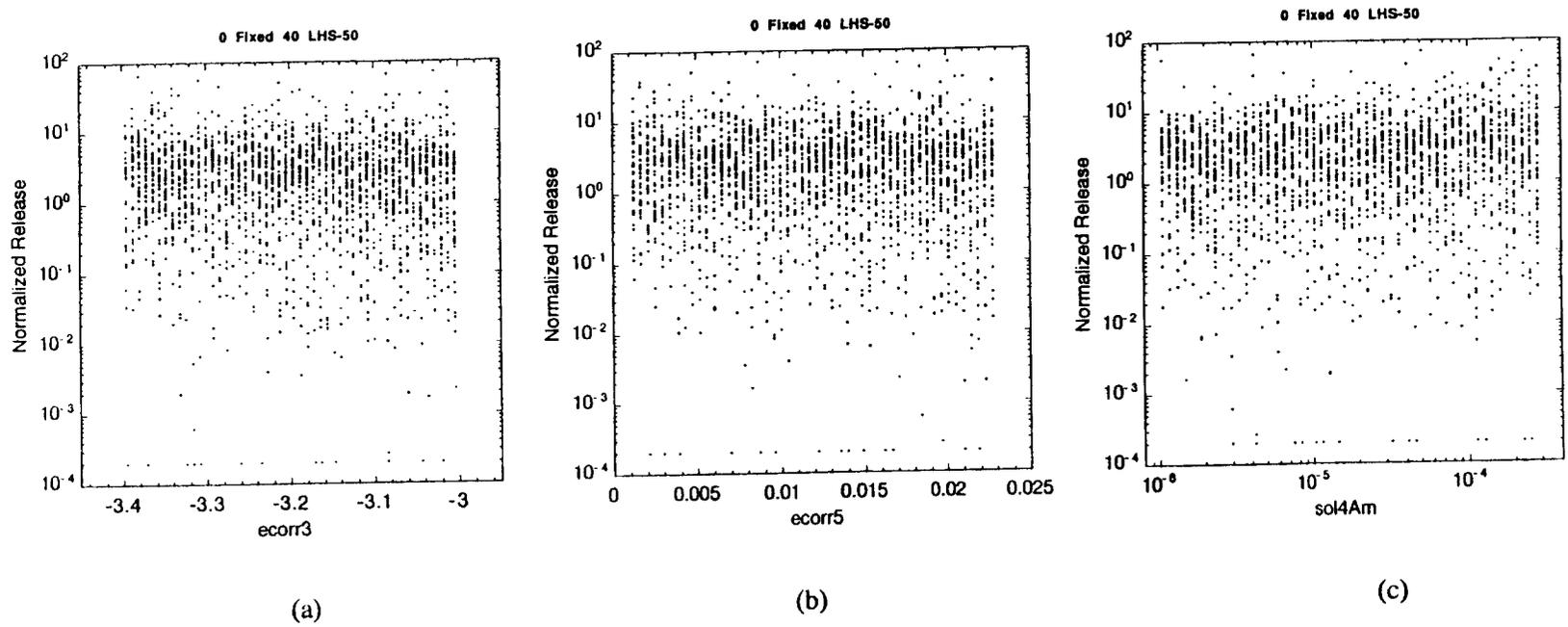


Figure 9: Scatter plots for parameters identified in the variance-based method, yet missed in the stepwise multilinear regression.

35/39

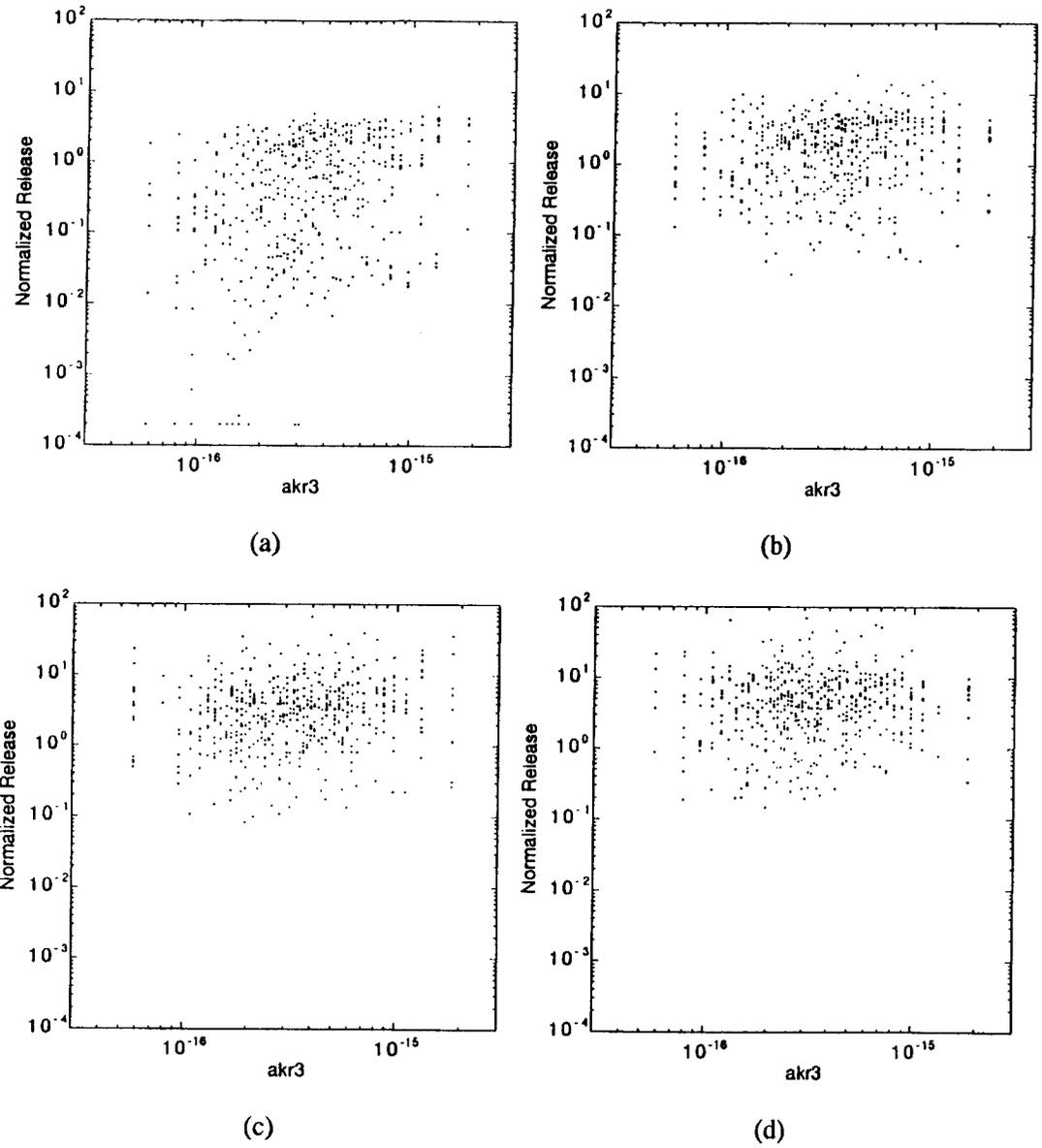


Figure 10: Scatter plots for akr3 when infil is restricted to (a) first, (b) second, (c) third, and (d) fourth quartile.

36/39

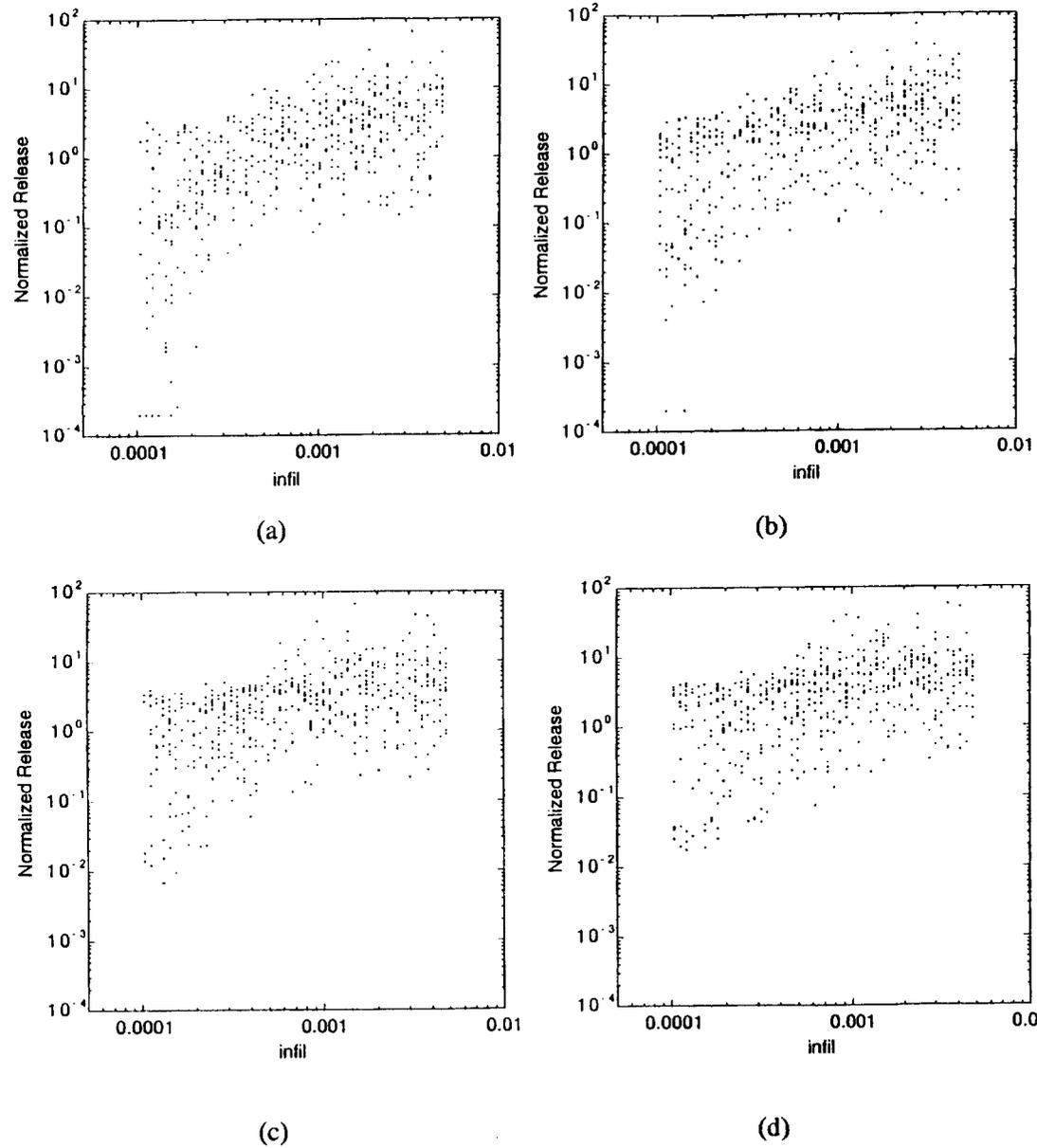


Figure 11: Scatter plots for infil when *akr3* is restricted to (a) first, (b) second, (c) third, and (d) fourth quartile.

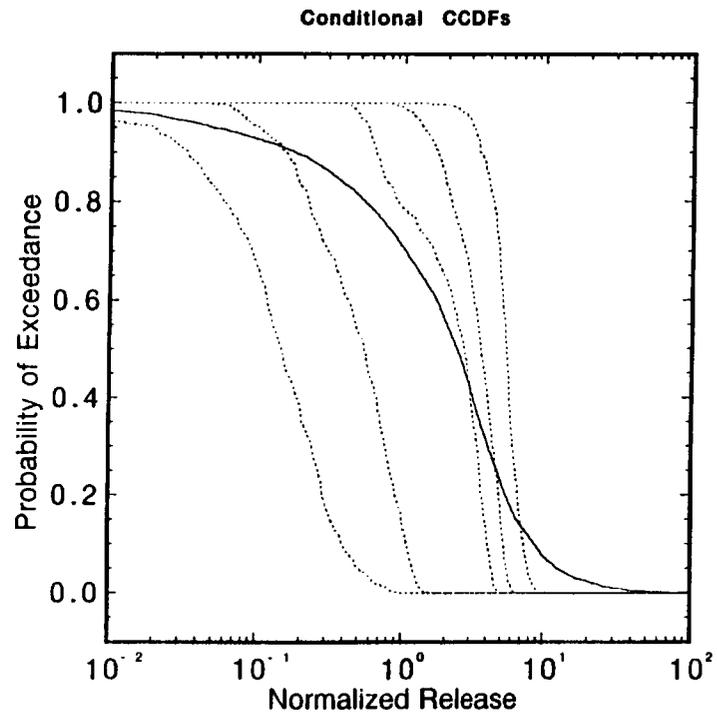


Figure 12: Conditional CCDFs using coarse LHS-5 on most important parameters and fine LHS-400 on less important parameters.

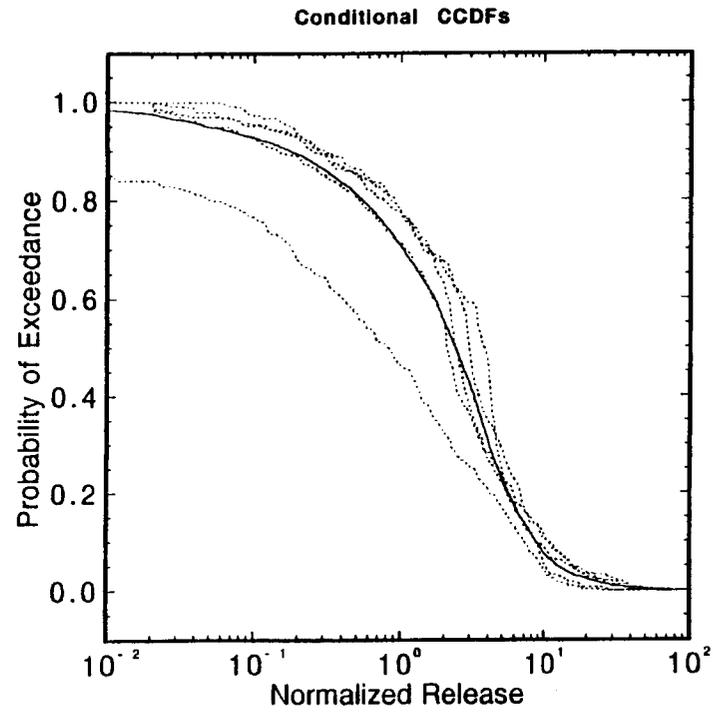


Figure 13: Conditional CCDFs using fine LHS-400 on most important parameters and coarse LHS-5 on less important parameters.

50/39