Proceedings of the 2005 Winter Simulation Conference
M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.

MULTIPLE PREDICTOR SMOOTHING METHODS FOR SENSITIVITY ANALYSIS

Curtis B. Storlie

Department of Statistics
Colorado State University
Fort Collins, CO 80523-1877, U.S.A.

Jon C. Helton

Department of Mathematics and Statistics
Arizona State University
Tempe, AZ 85287-1804, U.S.A.

ABSTRACT

The use of multiple predictor smoothing methods in sampling-based sensitivity analyses of complex models is investigated. Specifically, sensitivity analysis procedures based on smoothing methods employing the stepwise application of the following nonparametric regression techniques are described: (i) locally weighted regression (LOESS), (ii) additive models (GAMs), (iii) projection pursuit regression (PP_REG), and (iv) recursive partitioning regression (RP_REG). The indicated procedures are illustrated with both simple test problems and results from a performance assessment for a radioactive waste disposal facility (i.e., the Waste Isolation Pilot Plant). As shown by the example illustrations, the use of smoothing procedures based on nonparametric regression techniques can yield more informative sensitivity analysis results than can be obtained with more traditional sensitivity analysis procedures based on linear regression, rank regression or response surface regression when nonlinear relationships between model inputs and model predictions are present.

1 INTRODUCTION

Sampling-based approaches to uncertainty and sensitivity analysis are both effective and widely used (Helton and Davis 2000, 2002, 2003). Analyses of this type involve the generation and exploration of a mapping from uncertain analysis inputs to uncertain analysis results. The underlying idea is that analysis results $\mathbf{y}(\mathbf{x}) = [y_1(\mathbf{x}), y_2(\mathbf{x}), \ldots, y_{nY}(\mathbf{x})]$ are functions of uncertain analysis inputs $\mathbf{x} = [x_1, x_2, \ldots, x_{nX}]$. In turn, uncertainty in $\mathbf{x}$ results in a corresponding uncertainty in $\mathbf{y}(\mathbf{x})$. This leads to two questions: (i) What is the uncertainty in $\mathbf{y}(\mathbf{x})$ given the uncertainty in $\mathbf{x}$?, and (ii) How important are the individual elements of $\mathbf{x}$ with respect to the uncertainty in $\mathbf{y}(\mathbf{x})$? The goal of uncertainty analysis is to answer the first question, and the goal of sensitivity analysis is to answer the second question. In practice, the implementation of an uncertainty analysis and the implementation of a sensitivity analysis are very closely connected on both a conceptual and a computational level.

Five basic components underlie the implementation of a sampling-based uncertainty and sensitivity analysis: (i) Definition of distributions $D_1, D_2, \ldots, D_{nX}$ that characterize the epistemic uncertainty in the components $x_1, x_2, \ldots, x_{nX}$ of $\mathbf{x}$, (ii) Generation of a sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{nS}$ from the $\mathbf{x}$'s in consistency with the distributions $D_1, D_2, \ldots, D_{nX}$, (iii) Propagation of the sample through the analysis to produce a mapping $[\mathbf{x}_i, \mathbf{y}(\mathbf{x}_i)]$, $i = 1, 2, \ldots, nS$, from analysis inputs to analysis results, (iv) Presentation of uncertainty analysis results (i.e., approximations to the distributions of the elements of $\mathbf{y}$ constructed from the corresponding elements of $\mathbf{y}(\mathbf{x}_i)$, $i = 1, 2, \ldots, nS$), and (v) Determination of sensitivity analysis results (i.e., exploration of the mapping $[\mathbf{x}_i, \mathbf{y}(\mathbf{x}_i)]$, $i = 1, 2, \ldots, nS$).

The primary focus of this presentation is the sensitivity analysis component of a sampling-based uncertainty and sensitivity analysis. Traditional parametric regression procedures, often in conjunction with the use of rank transformations, are popular and usually effective sensitivity analysis tools (Section 2). However, such procedures can fail to identify the effects of influential variables when the underlying relationships between analysis inputs and analysis results are both nonlinear and nonmonotonic. Nonparametric regression procedures are presented as tools for use in sensitivity analyses when more traditional parametric regression procedures fail to identify the relationships that exist between analysis inputs and analysis results (Section 3). The application of nonparametric regression procedures in sensitivity analysis is illustrated with two analytic test functions (Campolongo 2000) and a result from a performance assessment (PA) for the Waste Isolation Pilot Plant (WIPP, Helton and Marietta 2000) (Section 4). The presentation then ends with a brief discussion (Section 5).

2 PARAMETRIC REGRESSION ANALYSIS

Traditional parametric regression analysis provides an algebraic representation of the relationships between a de-

pendent variable y (i.e., an element of **y**) and one or more independent variables (i.e., elements of **x**). Unless stated otherwise, regression analysis is usually assumed to involve the construction of linear models of the form

$$\hat{y} = b_0 + \sum_{j=1}^{nX} b_j x_j \,. \tag{1}$$

The regression coefficients in Equation (1) are usually determined such that the sum

$$\sum_{i=1}^{nS} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{nS} \left[y_i - \left(b_0 + \sum_{j=1}^{nS} b_j x_{ij} \right) \right]^2 \tag{2}$$

is minimized. As a result, the regression model in Equation (1) is often referred to as a least squares model.

An important property of least squares regression models is the equality

$$\sum_{i=1}^{nS} (y_i - \overline{y})^2 = \sum_{i=1}^{nS} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{nS} (\hat{y}_i - y_i)^2 \,, \tag{3}$$

where $\overline{y}$ denotes the estimated expected value for y. The ratio

$$R^2 = \sum_{i=1}^{nS} (\hat{y}_i - \overline{y})^2 \Big/ \sum_{i=1}^{nS} (y_i - \overline{y})^2 \tag{4}$$

provides a measure of the extent to which the regression model can match the observed data. Specifically, when the variation about the regression model is small, then the corresponding R^2 value is close to 1, which indicates that the regression model is accounting for most of the uncertainty in y. Conversely, an R^2 value close to 0 indicates that the regression model is not very successful in accounting for the uncertainty in y. When the individual x_j in the regression model in Equation (1) are linearly independent, the R^2 value for the regression model can be expressed as

$$R^2 = R_1^2 + R_2^2 + \ldots + R_{nX}^2 \,, \tag{5}$$

where R_j^2 is the R^2 value that results from regressing y on only x_j. Thus, R_j^2 is equal to the contribution of x_j to the R^2 value for the regression model in Equation (1) when the x_j's are independent.

The regression coefficients b_j, $j = 1, 2, \ldots, nX$, are not very useful in sensitivity analysis because each b_j is influenced by the units in which x_j is expressed and also does not incorporate any information on the distribution assigned to x_j. Because of this, the regression model in Equation (1) is usually reformulated as

$$(\hat{y} - \overline{y})/\hat{s} = \sum_{j=1}^{nX} \left(b_j \hat{s}_j / \hat{s} \right) \left(x_j - \overline{x}_j \right) \Big/ \hat{s}_j \,, \tag{7}$$

where $\overline{y}$, $\overline{x}_j$, $\hat{s}$ and $\hat{s}_j$ denote estimated means and standard deviations for y and x_j. The coefficients $b_j\hat{s}_j/\hat{s}$ in Equation (7) are referred to as standardized regression coefficients (SRCs).

The SRC $b_j\hat{s}_j/\hat{s}$ provides a measure of variable importance based on the effect on y relative to the standard deviation of y of moving x_j away from its expected value by a fixed fraction of its standard deviation. Further, when the x_j's are independent, the inclusion or exclusion of an individual x_j from the regression model has no effect on the SRCs for the remaining variables in the model. Thus, as long as the x_j's are independent, the SRCs $b_j\hat{s}_j/\hat{s}$ provide a useful measure of variable importance, with (i) the absolute values of the coefficients $b_j\hat{s}_j/\hat{s}$ providing a comparative measure of variable importance (i.e., variable x_u is more important than variable x_v if $|b_u\hat{s}_u/\hat{s}| > |b_v\hat{s}_v/\hat{s}|$) and (ii) the sign of $b_j\hat{s}_j/\hat{s}$ indicating whether x_j and y tend to move in the same direction or in opposite directions. However, when x_j's are not independent, SRCs do not provide reliable indications of variable importance.

For purposes of sensitivity analysis, there is usually no reason to construct a regression model containing all the uncertain variables (i.e., x_1, x_2, …, x_{nX}) as indicated in Equation (1). Rather, a more appropriate procedure is to construct regression models in a stepwise manner. With this procedure, a regression model is first constructed with the most influential variable (e.g., $\tilde{x}_1$ as determined based on R^2 values for regression models containing only single variables). Then, a regression model is constructed with $\tilde{x}_1$ and the next most influential variable (e.g., $\tilde{x}_2$ as determined based on R^2 values for regression models containing $\tilde{x}_1$ and each of the remaining variables). The process then repeats to determine $\tilde{x}_3$ in a similar manner and continues until no more variables with an identifiable effect on y_k can be found. Variable importance (i.e., sensitivity) is then indicated by the order in which variables are selected in the stepwise process, the changes in cumulative R^2 values as additional variables are added to the regression model, and the SRCs for the variables in the final regression model. An example of a sensitivity analysis of this form is presented in Table 1.

Table 1: Example of Stepwise Regression Analysis to Identify Uncertain Variables Affecting the Uncertainty in WIPP Repository Pressure under Undisturbed Conditions at 10,000 yr Performed for a Sample of Size 300 from 31 Uncertain Variables (Table 8.6, Helton and Davis 2000)

Step[a]	Variable[b]	SRC[c]	R^{2d}
1	*WMICDFLG*	0.718	0.508
2	*HALPOR*	0.466	0.732
3	*WGRCOR*	0.246	0.792
4	*ANHPRM*	0.129	0.809
5	*SHRGSSAT*	0.070	0.814
6	*SALPRES*	0.063	0.818

[a] Steps in stepwise regression analysis.
[b] Variables listed in the order of selection in regression analysis.
[c] SRCs for variables in final regression model.
[d] Cumulative R^2 value with entry of each variable into regression model.

This section only considers linear regression models. However, linear regression models also include models of forms such as

$$\hat{y} = b_0 + \sum_{j=1}^{nX} b_j f_j\left(x_j\right) + \sum_{j=1}^{nX} \sum_{l=j+1}^{nX} b_{jl} f_{jl}\left(x_j, x_l\right). \quad (8)$$

This inclusion exists because the preceding model is still linear in its coefficients (i.e., b_0, the b_j, the b_{jl}); in essence, the indicated transformations involving the x_j (i.e., $f_j(x_j)$, $f_{jl}(x_j, x_l)$) are simply defining a new set of analysis inputs to be used in a regression-based sensitivity analysis. Results can be improved in some analyses by well-chosen variable transformations of the form indicated in Equation (8). However, in large analyses involving many uncertain analysis inputs (i.e., x_j) and many possibly time-dependent analysis results (i.e., y's), the a priori determination of suitable transformations can be difficult. Also, care must be taken to suitably account for any correlations that may be introduced by the chosen transformations (i.e., $f_j(x_j)$ and $f_{jl}(x_j, x_l)$ may be highly correlated).

Nonlinear regression provides an alternative to linear regression that can be useful in some analyses. In nonlinear regression, at least some of the model coefficients are operated on by nonlinear functions. For example,

$$\hat{y} = b_0 + b_1 \exp\left(b_2 x_1\right) + b_3 \sin\left(b_4 x_2\right) \quad (9)$$

is a nonlinear model because b_2 and b_4 appear in expressions that are operated on by nonlinear functions. A major challenge in the use of nonlinear regression in sensitivity analysis is the determination of a suitable form for the nonlinear regression model.

A rank transformation can be used to convert a nonlinear but monotonic relationship between the x_j and y into a linear relationship (Iman and Conover 1979). With this transformation, the values for the x_j and y are replaced by their corresponding ranks. Specifically, the smallest value for a variable is assigned a rank of 1; the next largest value is assigned a rank of 2; tied values are assigned their average rank; and so on up to the largest value, which is assigned a rank of nS. Use of the rank transformation results in rank (i.e., Spearman) correlation coefficients (RCCs), rank regressions, standardized rank regression coefficients (SRRCs) and partial rank correlation coefficients (PRCCs). In the presence of nonlinear but monotonic relationships between the x_j and y, use of the rank transform can substantially improve the resolution of sensitivity analysis results (Table 2).

3 NONPARAMETRIC REGRESSION

There are drawbacks to the parametric regression techniques indicated in Section 2 that can reduce their effectiveness in some sensitivity analyses. First, it is necessary to provide an a priori specification of the form of the regression model (e.g., linear as in Equations (1) and (8), nonlinear as in Equation (9), or linear with rank transformed data). Unfortunately, when complex patterns of behavior are present, it can be difficult to determine the appropriate form for a regression model. Such determinations can be a particular challenge in exploratory analyses that can involve 10s or even 100s of analysis results, with each result potentially requiring the specification of a different regression model. Second, the specified form for the regression is required to hold across the entire mapping from analysis inputs to analysis results, which makes the representation of local behavior and/or asymptotes difficult. In addition, grid-based procedures (Kleijnen and Helton 1999) have the drawback that the associated sensitivity results can be dependent on the particular grid selected for use. Unfortunately, the most appropriate grid for use with these procedures is not always apparent.

Table 2: Comparison of Stepwise Regression Analyses with Raw and Rank-Transformed Data for Cumulative Brine Flow over 10, 000 yr under Undisturbed Conditions from Anhydrite Marker Beds to Disturbed Rock Zone Surrounding the WIPP Repository Performed for a Sample of Size 300 from 31 Uncertain Variables (Table 8.8, Helton and Davis 2000).

Step[a]	Raw Data			Rank-Transformed Data		
	Variable[b]	SRC[c]	R^{2d}	Variable[b]	SRRC[e]	R^{2d}
1	*ANHPRM*	0.562	0.320	*WMICDFLG*	–0.656	0.425
2	*WMICDFLG*	–0.309	0.423	*ANHPRM*	0.593	0.766
3	*WGRCOR*	–0.164	0.449	*HALPOR*	–0.155	0.802
4	*WASTWICK*	–0.145	0.471	*WGRCOR*	–0.152	0.824
5	*ANHBCEXP*	–0.120	0.486	*HALPRM*	0.143	0.845
6	*HALPOR*	–0.101	0.496	*SALPRES*	0.120	0.860
7				*WASTWICK*	–0.010	0.869

[a] Steps in stepwise regression analysis.
[b] Variables listed in order of selection in regression analysis.
[c] SRCs for variables in final regression model.
[d] Cumulative R^2 value with entry of each variable into regression model.
[e] SRRCs for variables in final regression model.

Nonparametric regression procedures provide an alternative to parametric regression procedures and grid-based procedures that can mitigate the potential problems indicated in the preceding paragraph. With nonparametric regression procedures, an a priori specification of the exact algebraic form of the regression model is not required. Rather, an iterative procedure is used to construct a model that captures the relationships that are present in the mapping between analysis inputs and a particular analysis result. This iterative construction procedure does not require the use of a grid and produces a model that can represent local patterns of behavior. Nonparametric regression is often referred to as smoothing. Popular nonparametric regression procedures include (i) locally weighted regression (LOESS), (ii) generalized additive models (GAMs), (iii) projection pursuit regression (PP_REG), and (iv) recursive partitioning regression (RP_REG). These procedures are briefly described below.

The LOESS technique (Cleveland 1979) is based on the assumption that the relationship between y and $\mathbf{x}$ is of the form

$$y = f(\mathbf{x}) = \alpha(\mathbf{x}) + \boldsymbol{\beta}(\mathbf{x})\mathbf{x}, \tag{10}$$

where $\boldsymbol{\beta}(\mathbf{x}) = [\beta_1(\mathbf{x}), \beta_2(\mathbf{x}), \ldots, \beta_{nX}(\mathbf{x})]$ and $\mathbf{x} = [x_1, x_2, \ldots, x_{nX}]^T$. In turn, an approximate relationship of the form

$$\hat{y} = \hat{f}(\mathbf{x}) = \hat{\alpha}(\mathbf{x}) + \hat{\boldsymbol{\beta}}(\mathbf{x})\mathbf{x} \tag{11}$$

is sought with LOESS. The quantities $\hat{\alpha}(\mathbf{x})$ and $\hat{\boldsymbol{\beta}}(\mathbf{x})$ for a given value of $\mathbf{x}$ are defined to be the values for α and $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_{nX}]$ that minimize the sum

$$\sum_{i=1}^{nS} (\alpha + \boldsymbol{\beta}\mathbf{x}_i - y_i)^2 \left[1 - \left(\frac{\|\mathbf{x} - \mathbf{x}_i\|}{d_r(\mathbf{x})}\right)^3\right]^3 I_{[0, d_r(\mathbf{x}))}(\|\mathbf{x} - \mathbf{x}_i\|), \tag{12}$$

where (i) $d_r(\mathbf{x})$ is the distance to the r^{th} nearest neighbor of $\mathbf{x}$ in nX-dimensional Eulidean space, (ii) $I_{[0,dr(\mathbf{x}))}(\|\mathbf{x} - \mathbf{x}_i\|)$ equals 1 if $\|\mathbf{x} - \mathbf{x}_i\| < d_r(\mathbf{x})$ and equals 0 otherwise, and (iii) the individual independent variables (i.e., $x_1, x_2, \ldots, x_{nX}$) are normalized to mean zero and standard deviation one so that the value for the norm $\|\cdot\|$ is not dominated by the units used for these variables. The determination of α and $\boldsymbol{\beta}$ is straightforward with the use of appropriate matrix techniques (p. 139, Simonoff 1996).

For GAMs (Hastie and Tibshirani 1990), the function $f(\mathbf{x})$ is assumed to have the form

$$f(\mathbf{x}) = \sum_{j=1}^{nX} f_j(x_j), \tag{13}$$

where the f_j are arbitrary functions that will be determined as part of the analysis process. In turn, the observed values for y are assumed to be of the form

$$y_i = f(\mathbf{x}_i) = \sum_{j=1}^{nX} f_j(x_{ij}). \tag{14}$$

Given initial estimates $\hat{f}_2, \hat{f}_3, \ldots, \hat{f}_{nX}$ for $f_2, f_3, \ldots, f_{nX}$, an estimate $\hat{f}_1$ for f_1 can be obtained through use of the relationship

$$y_i - \sum_{j=2}^{nX} \hat{f}_j\left(x_{ij}\right) \cong f_1\left(x_{i1}\right) \tag{15}$$

for $i = 1, 2, \ldots, nS$. In particular, a scatterplot smoother (e.g., LOESS with only one independent variable) can be used to smooth the partial residuals on the left hand side of Equation (15) across x_1. This produces an estimate $\hat{f}_1$ for f_1 defined across the range of values for x_1. Given this estimate for f_1, the estimate $\hat{f}_2$ for f_2 can be refined in the same manner across the range of values for x_2 with $\hat{f}_1$, $\hat{f}_3$, $\hat{f}_4$, …, $\hat{f}_{nX}$. This procedure then continues and repetitively cycles through the variables. The cycling continues until convergence is achieved. The result is $\hat{f}_j$ defined at $x_{1j}, x_{2j}, \ldots, x_{nS,j}$ for $j = 1, 2, \ldots, nX$. Additional detail is available elsewhere (pp. 90 – 91, Hastie and Tibshirani 1990; pp. 300 – 302, Chambers and Hastie 1992).

The PP_REG procedure (Friedman and Stuetzle 1981) involves both dimension reduction and additive modeling and is based on the assumption that $f(\mathbf{x})$ has the form

$$f\left(\mathbf{x}\right) = \sum_{s=1}^{nD} g_s\left(\boldsymbol{\alpha}_s \mathbf{x}\right), \tag{16}$$

where $\boldsymbol{\alpha}_s = [\alpha_{1s}, \alpha_{2s}, \ldots, \alpha_{nX,s}]$, $\mathbf{x} = [x_1, x_2, \ldots, x_{nX}]^{\mathrm{T}}$, $\boldsymbol{\alpha}_s \mathbf{x}$ corresponds to a linear combination of the elements of $\mathbf{x}$, and g_s is an arbitrary function. Values for g_s, $\boldsymbol{\alpha}_s$ and nD are determined as part of the analysis procedure. The expression in Equation (16) is an additive model with the quantities $\boldsymbol{\alpha}_s \mathbf{x}$ replacing the elements x_j of $\mathbf{x}$ as the independent variables. Further, this expression involves a reduction in dimension as nD is usually smaller than nX. The entities $\hat{\boldsymbol{\alpha}}_1$, $\hat{\boldsymbol{\alpha}}_2$, …, $\hat{\boldsymbol{\alpha}}_{nD}$ and $\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_{nD}$ are estimated as part of the construction process. This is accomplished by first estimating $\boldsymbol{\alpha}_1$ and g_1. Specifically, $\hat{\boldsymbol{\alpha}}_1$ and $\hat{g}_1$ are defined to be the values for $\boldsymbol{\alpha}$ and $g_{\boldsymbol{\alpha}}$ that minimize the sum

$$\sum_{i=1}^{nS} \left[y_i - g_{\boldsymbol{\alpha}}\left(\boldsymbol{\alpha}\mathbf{x}_i\right) \right]^2, \tag{17}$$

where $\boldsymbol{\alpha} \in R^{nX}$, $\|\boldsymbol{\alpha}\| = 1$, and $g_{\boldsymbol{\alpha}}$ is the outcome of using a scatterplot smoother (e.g., LOESS) on the points $[y_i, \boldsymbol{\alpha}\mathbf{x}_i]$, $i = 1, 2, \ldots, nS$. Once $\hat{\boldsymbol{\alpha}}_1$ and $\hat{g}_1$ are estimated, the partial residuals $y_i - \hat{g}_1(\hat{\boldsymbol{\alpha}}_1 \mathbf{x}_i)$, $i = 1, 2, \ldots, nS$, are used to obtain $\hat{\boldsymbol{\alpha}}_2$ and $\hat{g}_2$. Specifically, $\hat{\boldsymbol{\alpha}}_2$ and $\hat{g}_2$ are defined to be the values for $\boldsymbol{\alpha}$ and $g_{\boldsymbol{\alpha}}$ that minimize the sum

$$\sum_{i=1}^{nS} \left\{ \left[y_i - \hat{g}_1\left(\hat{\boldsymbol{\alpha}}_1 \mathbf{x}_1\right) - g_{\boldsymbol{\alpha}}\left(\boldsymbol{\alpha}\mathbf{x}_i\right) \right] \right\}^2, \tag{18}$$

where $\boldsymbol{\alpha} \in R^{nX}$, $\|\boldsymbol{\alpha}\| = 1$, and $g_{\boldsymbol{\alpha}}$ is the outcome of using a scatterplot smoother on the points $[y_i - \hat{g}_1(\hat{\boldsymbol{\alpha}}_1 \mathbf{x}), \boldsymbol{\alpha}\mathbf{x}_i]$, $i = 1, 2, \ldots, nS$. This process continues until no appreciable improvement based on a relative error criterion is observed.

The RP_REG procedure (Breiman et al. 1984) is based on splitting the data into subgroups where observations within each subgroup are more homogeneous than they are over the set of all observations. Then, $f(\mathbf{x})$ is estimated with regression models defined for each subgroup. Specifically, $f(\mathbf{x})$ is estimated by

$$\hat{f}\left(\mathbf{x}\right) = \sum_{s=1}^{nP} \left(\hat{\alpha}_s + \hat{\boldsymbol{\beta}}_s \mathbf{x} \right) I_s\left(\mathbf{x}\right), \tag{19}$$

where (i) $\mathcal{A}_s$, $s = 1, 2, \ldots, nP$, designate the subgroups into which the data are partitioned, (ii) $\hat{y} = \hat{\alpha}_s + \hat{\boldsymbol{\beta}}_s \mathbf{x}$ is the least squares approximation to y associated with $\mathcal{A}_s$, and (iii) I_s is the indicator functions such $I_s(\mathbf{x}) = 1$ if $\mathbf{x}$ is associated with $\mathcal{A}_s$ and $I_s(\mathbf{x}) = 0$ otherwise. The subgroups $\mathcal{A}_s$, $s = 1, 2, \ldots, nP$, are developed algorithmically from the observations $[\mathbf{x}_i, y_i]$, $i = 1, 2, \ldots, nS$.

The preceding procedures can all be carried out in a stepwise manner to determine variable importance, with (i) the most important variable $\tilde{x}_1$ being the variable that results in the single-variable model with the most predictive capability, (ii) the second most important variable $\tilde{x}_2$ being the variable that in conjunction with $\tilde{x}_1$ results in the two-variable model with the most predictive capability, and so on until (iii) some stopping criteria is reached that indicates that the consideration of additional variables does not produce models with improved predictive capability. Order of selection in the stepwise construction process and fraction of variability explained (i.e., R^2 as defined in Equation (8)) can be used to indicate variable importance. The F-statistic with appropriate degrees of freedom (a topic too complicated for consideration here; see Section 3.9 in Hastie and Tibshirani (1990) and Section 3.13 in Ruppert et al. (2003)) can be used to determine a stopping point in the stepwise variable selection procedure.

The R^2 value is the primary quantity used in this presentation to assess the contribution of the uncertainty associated with a group of variables to the uncertainty in an analysis result. In particular, if $\tilde{\mathbf{x}} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_p]$ is a vector of variables taken from the variables $x_1, x_2, \ldots, x_{nX}$ under consideration in a particular analysis (i.e., $\mathbf{x} = [x_1, x_2, \ldots, x_{nX}]$ is the vector of uncertain inputs under considera-

tion), $\hat{f}(\tilde{\mathbf{x}}) = \hat{f}(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_p)$ is an approximation to the real model $f(\mathbf{x}) = f(x_1, x_2, \ldots, x_{nX})$ estimated with a particular procedure from a mapping $[y_i, \mathbf{x}_i]$, $i = 1, 2, \ldots nS$, from analysis inputs to analysis results, and $\tilde{\mathbf{x}}_i = [\tilde{x}_{i1}, \tilde{x}_{i2}, \ldots, \tilde{x}_{ip}]$ for $i = 1, 2, \ldots, nS$, then

$$R^2 = 1 - \sum_{i=1}^{nS} \left[y_i - \hat{f}(\tilde{\mathbf{x}}_i) \right]^2 \Bigg/ \sum_{i=1}^{nS} [y_i - \bar{y}]^2 \qquad (20)$$

provides an estimate of the fraction of the uncertainty in y that derives from the uncertainty associated with the variables in $\tilde{\mathbf{x}}$.

In the following, R^2 is calculated in a stepwise manner for use in determining variable importance. The most important variable, designated $\tilde{x}_1$, is the element of $\mathbf{x} = [x_1, x_2, \ldots, x_{nX}]$ that gives the largest value for R^2. That is, $\tilde{\mathbf{x}} = [x_1]$, $\tilde{\mathbf{x}} = [x_2]$, …, $\tilde{\mathbf{x}} = [x_{nX}]$ are considered in the definition of R^2 in Equation (20), and the x_j that gives the highest value for R^2 is deemed to be the most important variable and taken to be $\tilde{x}_1$. The second most important variable, designated $\tilde{x}_2$, is the element of $\mathbf{x} = [x_1, x_2, \ldots, x_{nX}]$ that gives the largest value for R^2 when all possible values for $\tilde{\mathbf{x}} = [\tilde{x}_1, x_j]$, $\tilde{x}_1 \neq x_j$, are considered. The third most important variable, designated $\tilde{x}_3$, is determined in like manner from consideration of vectors of the form $\tilde{\mathbf{x}} = [\tilde{x}_1, \tilde{x}_2, x_j]$, $\tilde{x}_1 \neq x_j$ and $\tilde{x}_2 \neq x_j$, and so on through all nX elements of $\mathbf{x}$.

The contribution of $\tilde{\mathbf{x}}$ to the uncertainty in y that is estimated by R^2 is formally defined by

$$\rho^2 = 1 - E\left(\left[y - E(y|\tilde{\mathbf{x}}) \right]^2 \right) \Big/ E\left(\left[y - E(y) \right]^2 \right), \quad (21)$$

where (i)

$$E(y) = \int_{\mathbb{X}} f(\mathbf{x}) d_X(\mathbf{x}) \, \mathrm{d}X$$

$$E(y|\tilde{\mathbf{x}}) = \int_{\tilde{\mathbb{X}}^c} f(\tilde{\mathbf{x}}^c, \tilde{\mathbf{x}}) \, d_{\tilde{X}^c}(\tilde{\mathbf{x}}^c) \, \mathrm{d}\tilde{X}^c$$

$$E\left(\left[y - E(y) \right]^2 \right) = \int_{\mathbb{X}} \left[f(\mathbf{x}) - E(y) \right]^2 d_X(\mathbf{x}) \, \mathrm{d}X$$

$$E\left(\left[y - E(y|\tilde{\mathbf{x}}) \right]^2 \right) = \int_{\mathbb{X}} \left[f(\mathbf{x}) - E(y|\tilde{\mathbf{x}}) \right]^2 d_X(\mathbf{x}) \, \mathrm{d}X,$$

(ii) $(X, \mathbb{X}, p_X)$, $(\tilde{X}, \tilde{\mathbb{X}}, p_{\tilde{X}})$ and $(\tilde{X}^c, \tilde{\mathbb{X}}^c, p_{\tilde{X}^c})$ are the probability spaces associated with $\mathbf{x}$, $\tilde{\mathbf{x}}$, and $\tilde{\mathbf{x}}^c$, where $\tilde{\mathbf{x}}^c$ contains the elements of $\mathbf{x}$ not contained in $\tilde{\mathbf{x}}$, and (iii) $d_X(\mathbf{x})$, $d_{\tilde{X}}(\tilde{\mathbf{x}})$ and $d_{\tilde{X}^c}(\tilde{\mathbf{x}}^c)$ are the corresponding density functions for $\mathbf{x}$, $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}^c$ (Saltelli et al. 1999). For the simple test functions considered in the next section, ρ^2 can be calculated and used in comparisons with its corresponding estimate R^2 defined in Equation (20). However, the direct calculation of ρ^2 is too computationally demanding to be practical for large models of the type used in the WIPP PA.

A more detailed discussion of the use of nonparametric regression in sensitivity analysis is given in Storlie and Helton (2005). General discussions of nonparametric regression procedures are give by Hastie and Tibshirani (1990), Chambers and Hastie (1992), Simonoff (1996), Bowman and Azzalini (1997) and Ruppert et al. (2003).

4 EXAMPLE RESULTS

As indicated in the Introduction, the application of nonparametric regression procedures in sensitivity analysis is illustrated with two analytic test functions and a result from a PA for the WIPP. The two test functions are given by

$$y = f(x_1, x_2, \ldots, x_8) = \prod_{j=1}^{8} \left\{ \frac{\left| 4x_j - 2 \right| + a_j}{1 + a_j} \right\} \qquad (22)$$

with $[a_1, a_2, \ldots, a_8] = [0, 1, 4.5, 9, 99, 99, 99, 99]$, and

$$\begin{aligned} y &= f(x_1, x_2, x_3) \\ &= \sin(2\pi x_1 - \pi) + 7 \sin^2(2\pi x_2 - \pi) \\ &\quad + 0.1(2\pi x_3 - \pi)^4 \sin(2\pi x_1 - \pi). \end{aligned} \qquad (23)$$

The independent variables $x_1, x_2, \ldots, x_8$ are assumed to be independent and uniformly distributed on [0, 1]. The result from the WIPP PA is pressure (*WAS_PRES*) in the repository at 10, 000 yr subsequent to a drilling intrusion at 1000 yr. The underlying model involves 31 uncertain variables and is based on the numerical solution of a system of nonlinear partial differential equations (Vaughn et al. 2000).

The analyses for the test functions in Equations (22) and (23) use a random sample of size 300 from $x_1, x_2, \ldots, x_{10}$, where x_9 and x_{10} are spurious variables included in the sample that, like $x_1, x_2, \ldots, x_8$, are independent and uniformly distributed on [0, 1]. The analyses for *WAS_PRES* use a sample of size 300 obtained by pooling three independent Latin hypercube samples of size 100 from 31 uncertain variables.

The analyses for the three examples are presented in Tables 3–5. In these tables, LIN_REG, RANK_REG and RS_REG are used to indicate linear regression, rank

Table 3: Sensitivity Analyses for Test Function $y = f(x_1, x_2, \ldots, x_8)$ in Equation (22)

Var[a]	R^2[b]	df[c]	p-val[d]	Var[a]	R^2[b]	df[c]	p-val[d]	Var[a]	R^2[b]	df[c]	p-val[d]	Var[a]	R^2[b]	df[c]	p-val[d]
LIN_REG				**RANK_REG**				**RS_REG**				**LOESS**			
x_{10}	0.0231	1.0	0.0084	x_{10}	0.0239	1.0	0.0073	x_1	0.6970	2.0	0.0000	x_1	0.7373	3.3	0.0000
GAM				**PP_REG**				x_2	0.8560	3.0	0.0000	x_2	0.8008	4.4	0.0000
x_1	0.7513	6.0	0.0000	x_1	0.7486	5.4	0.0000	x_3	0.8682	4.0	0.0000	**TRUE MODEL**			
x_2	0.9143	6.0	0.0000	x_2	0.9141	5.4	0.0000	x_6	0.8745	5.0	0.0159	x_1	0.7115	NA[e]	NA
x_3	0.9292	6.0	0.0000	x_8	0.9449	5.8	0.0000	**RP_REG**				x_2	0.9546	NA	NA
x_4	0.9324	2.0	0.0017	x_3	0.9610	5.3	0.0000	x_1	0.7500	3.0	0.0000	x_3	0.9891	NA	NA
								x_2	0.9654	32.0	0.0000	x_4	0.9996	NA	NA
								x_3	0.9792	32.0	0.0000	x_5	0.9997	NA	NA
								x_4	0.9808	-3.0	0.0000	x_6	0.9998	NA	NA
								x_9	0.9886	49.0	0.0000	x_7	0.9999	NA	NA
												x_8	1.0000	NA	NA

[a] Variables listed in order of selection with sample of size $nS = 300$.
[b] Cumulative R^2 value with entry of each variable into model (see Equation (21) for True Model and Equation (20) for all other cases).
[c] Incremental degrees of freedom with entry of each variable into model.
[d] p-value for model with addition of each new variable.
[e] NA indicates that result is not applicable.

Table 4: Sensitivity Analyses for Test Function $y = f(x_1, x_2, x_3)$ in Equation (23)

Var[a]	R^2[b]	df[c]	p-val[d]	Var[a]	R^2[b]	df[c]	p-val[d]	Var[a]	R^2[b]	df[c]	p-val[d]	Var[a]	R^2[b]	df[c]	p-val[d]
LIN_REG				**RANK_REG**				**RS_REG**				**LOESS**			
x_1	0.1579	1.0	0.0000	x_1	0.1442	1.0	0.0000	x_1	0.1595	2.0	0.0000	x_1	0.2685	3.3	0.0000
x_3	0.1735	1.0	0.0185	**PP_REG**				x_2	0.2026	3.0	0.0014	x_3	0.3613	4.3	0.0000
GAM				x_2	0.3572	5.3	0.0000	**RP_REG**				**TRUE MODEL**			
x_2	0.3775	8.0	0.0000					x_2	0.3785	9.0	0.0000	x_2	0.4463	NA[3]	NA
x_1	0.7449	8.0	0.0000					x_1	0.7722	38.0	0.0000	x_1	0.7593	NA	NA
								x_9	0.8367	20.0	0.0000	x_3	1.0000	NA	NA
								x_3	0.8579	2.0	0.0000				

[a] Variables listed in order of selection with sample of size $nS = 300$.
[b] Cumulative R^2 value with entry of each variable into model (see Equation (21) for True Model and Equation (20) for all other cases).
[c] Incremental degrees of freedom with entry of each variable into model.
[d] p-value for model with addition of each new variable.
[a] NA indicates that result is not applicable.

regression and response surface regression, respectively, where RS_REG denotes a model of the form

$$\hat{y} = b_0 + \sum_{j=1}^{nX} b_j x_j + \sum_{j=1}^{nX} \sum_{l=j}^{nX} b_{jl} x_j x_l. \qquad (24)$$

Further, the designators LOESS, GAM, PP_REG and RP_REG remain as in Section 3, and the designator TRUE MODEL indicates results obtained with the test functions and the variance decomposition described in Equation (21).

Methods based on LIN_REG and RANK_REG perform poorly for the test function $y = f(x_1, x_2, \ldots, x_8)$ in Equation (22) and result in very low R^2 values (Table 3). In contrast, the remaining methods perform well and result in R^2 values between 0.80 and 0.99. The RP_REG procedure performed best as its R^2 values are in close agreement with results from an analytic variance decomposition (i.e., TRUE MODEL) for the four dominant variables (i.e., x_1, x_2, x_3, x_4). After RP_REG, the GAM procedure performs best but does not match the R^2 values from TRUE MODEL quite as well.

Methods based on LIN_REG, RANK_REG, RS_REG, LOESS and PP_REG all perform poorly for the test function $y = f(x_1, x_2, x_3)$ in Equation (23) (Table 4). Again, results obtained with the RP_REG and GAM procedures compare best with the TRUE MODEL results. However, the comparisons are not as good as those in Table 3, with the GAM procedure failing to identify the effect of x_3 and the RP_REG procedure indicating an effect for the spurious variable x_9.

Methods based on LIN_REG, RANK_REG and LOESS perform poorly for the variable *WAS_PRES* from the WIPP PA (Table 5) and result in final models with R^2 values between 0.26 and 0.52. Further, LIN_REG and RANK_REG fail to identify the dominant variable *BHPRM*. The RS_REG, GAM, PP_REG and RP_REG procedures perform reasonably well and result in final models with R^2 values between 0.81 and 0.93. Based on R^2 values and knowledge with respect to the actual effects of

Table 5: Sensitivity Analyses for Pressure (*WAS_PRES*) in the Repository at 10, 000 yr Subsequent to a Drilling Intrusion at 1000 yr

Var[a]	R^{2}[b]	*df*[c]	p-val[d]	Var[a]	R^{2}[b]	*df*[c]	p-val[d]	Var[a]	R^{2}[b]	*df*[c]	p-val[d]	Var[a]	R^{2}[b]	*df*[c]	p-val[d]
LIN-REG				**RANK_REG**				**RS_REG**				**LOESS**			
HALPRM	0.1188	1.0	0.0000	*HALPRM*	0.1207	1.0	0.0000	*BHPRM*	0.4550	2.0	0.0000	*BHPRM*	0.4610	3.3	0.0000
BPCOMP	0.1724	1.0	0.0000	*BPCOMP*	0.1716	1.0	0.0000	*HALPRM*	0.5499	3.0	0.0000	*HALPRM*	0.5202	4.4	0.0000
ANHPRM	0.2168	1.0	0.0001	*ANHPRM*	0.2023	1.0	0.0008	*BPCOMP*	0.6201	4.0	0.0000				
HALPOR	0.2428	1.0	0.0016	*BPVOL*	0.2258	1.0	0.0030	*ANHPRM*	0.6873	5.0	0.0000				
BPVOL	0.2679	1.0	0.0017	*HALPOR*	0.2494	1.0	0.0026	*HALPOR*	0.7299	6.0	0.0000				
GAM				*SHRGSSAT*	0.2636	1.0	0.0182	*WGRCOR*	0.7713	7.0	0.0000				
BHPRM	0.4906	8.0	0.0000	**PP_REG**				*WMICDFLG*	0.8030	8.0	0.0000				
ANHPRM	0.5622	4.0	0.0000	*BHPRM*	0.4794	5.4	0.0000	*BPMAP*	0.8273	9.0	0.0001				
BPCOMP	0.6246	2.0	0.0000	*HALPRM*	0.5564	1.7	0.0000	*BPINTPRS*	0.8456	10.0	0.0018				
HALPRM	0.6865	2.0	0.0000	*ANHPRM*	0.6373	4.2	0.0000	**RP_REG**							
HALPOR	0.7287	4.0	0.0000	*BPCOMP*	0.7234	13.4	0.0000	*BHPRM*	0.4906	9.0	0.0000				
WGRCOR	0.7559	4.0	0.0000	*WMICDFLG*	0.7898	5.8	0.0000	*HALPRM*	0.6053	8.0	0.0000				
BPVOL	0.7681	1.0	0.0002	*WGRCOR*	0.8135	6.6	0.0000	*ANHPRM*	0.7041	10.0	0.0000				
SHRBRSAT	0.7826	6.0	0.0076	*BPVOL*	0.8632	7.3	0.0000	*BPCOMP*	0.8307	27.0	0.0000				
WMICDFLG	0.7920	2.0	0.0029	*ANHBCEXP*	0.8886	-0.1	0.0000	*WGRCOR*	0.8382	-13.0	0.0000				
BPINTPRS	0.7989	1.0	0.0028	*HALPOR*	0.9077	12.3	0.0000	*HALPOR*	0.9003	28.0	0.0000				
SHRGSSAT	0.8053	1.0	0.0034	*SALPRES*	0.9253	6.4	0.0000	*BPINTPRS*	0.9285	18.0	0.0000				

[a] Variables listed in order of selection with sample of size $nS = 300$.
[b] Cumulative R^2 value with entry of each variable into model (see Equation (20)).
[c] Incremental degrees of freedom with entry of each variable into model.
[d] *p*-value for model with addition of each new variable.

the sampled variables, the PP_REG and RP_REG procedures performed best.

In addition to the two test functions and the variable *WAS_PRES*, the complete study (Storlie and Helton 2005) considered two additional test functions and five additional variables from the WIPP PA.

5 OBSERVATIONS AND INSIGHTS

The following observations and insights are based on the three examples described in this presentation and on the additional seven examples contained in the complete study (Storlie and Helton 2005). Nonparametric methods worked quite well for sensitivity analysis and provide a useful addition to currently employed sampling-based sensitivity analysis procedures.

The overall best method considered in this study is RP_REG. In the test cases, it almost always ordered the input variables correctly and estimated the contributions to the R^2 accurately. The drawback is that it takes longer to apply than any of the other methods.

The GAM and RS_REG procedures had good performance on the test data and are fast computationally. The RS_REG procedure can model a certain degree of interaction while GAM does not. However, GAM can model more general nonlinearity than RS_REG. Also, multiplicative interaction terms could be used in GAM to make it a more general method.

The LOESS and PP_REG procedures displayed some problems that could reduce their usefulness for sensitivity analysis. Specifically, LOESS sometimes failed to identify important input variables, although it usually identified the two most important variables. The PP_REG procedure showed a tendency to err in the opposite direction and often included insignificant input variables in the model.

Given the nonlinear relationships that can be present in analyses with complex computer models, one should be cautious about using only linear methods for sensitivity analysis. However, when a linear regression with raw or rank-transformed data is appropriate, it should be used as it is the easiest method to implement and interpret.

A reasonable analysis strategy is initially to fit linear regressions with raw and rank-transformed data and ob-

serve the R^2 values. If these values are below 0.9, then fit a RS_REG surface. If RS_REG also has an R^2 below 0.9, then fit a GAM surface. If the GAM surface still has a low R^2, then fit a RP_REG model. This approach restricts the use of the more computationally demanding RP_REG procedure to situations where its use is necessary. This is important because real analyses can involve carrying out sensitivity analyses for hundreds of time-dependent analysis results (e.g., see the sensitivity analyses summarized in Helton and Marietta (2000)). The authors' experience is that linear regression with rank-transformed data and examination of associated scatterplots is usually sufficient to carry out a successful sensitivity analysis. However, there are situations where this approach will not be successful. Then, nonparametric regression procedures can often provide the needed techniques to determine the relationships between uncertain analysis inputs and analysis results.

ACKNOWLEDGEMENTS

Work performed for Sandia National Laboratories (SNL), which is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Security Administration under contract DE-AC04-94AL-85000. Editorial support provided by F. Puffer and J. Ripple of Tech Reps, a division of Ktech Corporation.

REFERENCES

Bowman, A.W., and A. Azzalini. 1997. *Applied smoothing techniques for data analysis*. Oxford: Clarendon.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone. 1984. *Classification and regression trees*. Belmont, CA: Wadsworth Intl.

Campolongo, F., A. Saltelli, T. Sorensen, and S. Tarantola. 2000. Hitchhiker's guide to sensitivity analysis. In *Sensitivity Analysis*, ed. A. Saltelli, K. Chan, and M. Scott. 15-47, New York, NY: John Wiley & Sons.

Chambers, J.M., and T.J. Hastie. 1992. *Statistical models in S*. Pacific Grove, CA: Wadsworth & Brooks.

Cleveland, W.S. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of American Statistical Association* 14 (368): 829-836.

Friedman, J.H., and W. Stuetzle. 1981. Projection pursuit regression. *Journal of the American Medical Association* 76 (376): 817-823.

Hastie, T.J., and R.J. Tibshirani. 1990. *Generalized additive models*. London: Chapman & Hall.

Helton, J.C., and F.J. Davis. 2000. Sampling-based methods for uncertainty and sensitivity analysis. Albuquerque, NM: Sandia National Laboratories.

Helton, J.C., and F.J. Davis. 2002. Illustration of sampling-based methods for uncertainty and sensitivity analysis. *Risk Analysis* 22 (3): 591-622.

Helton, J.C., and F.J. Davis. 2003. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering and System Safety* 81 (1): 23-69.

Helton, J.C., and M.G. Marietta. 2000. Special issue: The 1996 performance assessment for the Waste Isolation Pilot Plant. *Reliability Engineering and System Safety* 69 (1-3): 1-451.

Iman, R.L., and W.J. Conover. 1979. The use of the rank transform in regression. *Technometrics* 21 (4): 499-509.

Kleijnen, J.P.C., and J.C. Helton. 1999. Statistical analyses of scatterplots to identify important factors in large-scale simulations, 1: Review and comparison of techniques. *Reliability Engineering and System Safety* 65 (2): 147-185.

Ruppert, D., M.P. Wand, and R.J. Carroll. 2003. *Semiparametric regression*. New York, NY: Cambridge University Press.

Saltelli, A., S. Tarantola, and K.P.-S. Chan. 1999. A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics* 41 (1): 39-56.

Simonoff, J.S. 1996. *Smoothing methods in statistics*. New York, NY: Springer-Verlag.

Storlie, C.B., and J.C. Helton. 2005. Multiple predictor smoothing methods for sensitivity analysis. Albuquerque, NM: Sandia National Laboratories.

Vaughn, P., J.E. Bean, J.C. Helton, M. E. Lord, R.J. MacKinnon, and J.D. Schreiber. 2000. Representation of two-phase flow in the vicinity of the repository in the 1996 performance assessment for the Waste Isolation Pilot Plant. *Reliability Engineering and System Safety* 69 (1-3): 205-226.

AUTHOR BIOGRAPHY

CURTIS B. STORLIE received a Ph.D. in statistics from Colorado State University in 2005 and currently holds a post doctoral appointment in the Department of Statistics at North Carolina State University. His research interests involve nonparametric regression and stochastic modeling.

JON C. HELTON is a Professor Emeritus of Mathematics at Arizona State University. His research interests involve sampling-based procedures for uncertainty and sensitivity analysis and the performance of analyses for complex engineered facilities such as nuclear power plants and radioactive waste disposal sites.