Updated Code for Sensitivity Analysis Based on Regression

Yevgeniy V. Zagayevskiy and Clayton V. Deutsch

Complex earth science systems, such as the response of a petroleum reservoir, are often understood by Monte Carlo Simulation where input variables take different realizations. The uncertainty in the response variables is observed directly, but understanding the importance of each input variable requires further post processing. A methodology of local sensitivity analysis based on linear and quadratic regression models is developed with software. The program implements a sensitivity analysis algorithm and constructs response models that are appropriate to understand the contribution of each variable in complex problems. Sensitivity coefficients, predicted response values and summary statistics with model utility tests and analysis of variance tables for the regression models are evaluated. The result of sensitivity analysis is visually summarized in the form of extended tornado chart. Two case studies are examined. The setting of each case study is taken from representative petroleum and mining examples. The computational aspects of sensitivity analysis are presented with pros and cons for the conventional and extended tornado charts. The importance of standardized sensitivity coefficients and other measures are developed. The proposed methodology is robust, efficient, descriptive and straightforward. Minimization of uncertainty in relationship to the different input variables and a resulting greater understanding of how each input variable contributes to the response variables are useful for decision-making and further study.

1. Introduction

Sensitivity analysis is an integral part of the analysis of complex natural engineering systems. It enables examining the system more accurately and fully by investigating the relationship between input parameters and response variables. In experimental design and subsequent mathematical modeling, a sensitivity study helps to identify how model responses are related to model parameters. The basic relationship will likely be known to an investigator, but they may not know how to rank the relative importance of different input variables and they may want to know where to focus future attention to reduce uncertainty.

The goal of sensitivity analysis is to apportion contribution of uncertainty in input variables to uncertainty of model responses, or in other words, to detect and quantify changes in values of model responses caused by corresponding changes in values of input parameters (Saltelli et al, 2008). Results of the study indicate the input parameters that are important in the model and those of secondary or tertiary importance. Implementation of sensitivity analysis helps to (a) characterize the system by exploring relationships between model responses and input variables and in some cases also exploring the interaction between input variables, (b) assess the quality of the existing model, (c) simplify existing complex models, and (d) prioritize research objectives and set up an experimental design (Saltelli et al, 2008). Sensitivity analysis supplements uncertainty analysis and the two are usually performed together.

Sensitivity analysis is a well-established research area, hence numerous collection of literature can be found regarding this topic. Among several available approaches to sensitivity analysis this paper focuses on local sensitivity analysis, where sensitivity coefficients are defined as the partial derivative of a model response with respect to an input variable (Jorgensen, 1994; Gacuci, 2003; Balakrishnan et al, 2006; Conejo et al, 2006; Saltelli et al, 2008; Azmy and Sratori, 2010). In numerical application, the partial derivative is replaced with the ratio of response change to change of single input variable that caused this variation, with other input variables held constant at some values (usually at their mean values). Another method to derive sensitivity coefficients, which is used in this paper, is regression analysis. It has been proven for linear regression model that sensitivity coefficients and regression coefficients are identical (Chatterjee and Hadi, 1988). Linear regression coefficients are referred to as first order sensitivity coefficients. Higher order sensitivity coefficients can be derived from higher order regression models. For instance, second order sensitivity coefficients correspond to interaction terms from a quadratic regression model.

Sensitivity coefficients represented by partial derivatives are not able to fully characterize the influence of input variables on model responses. In order to improve information about the importance of input variables and eliminate dependence on the units of variables sigma-normalized sensitivity coefficients (also known as standardized sensitivity coefficients) are introduced (Oakley and O'Hagan, 2004; Saltelli et al, 2008). It is shown in subsequent sections of this paper that sensitivity coefficients are good for capturing the contribution of individual

input variables to a model value, while sigma-normalized coefficients are more suitable for revealing the contribution of uncertainty in input variables to uncertainty in model responses. Standardized sensitivity coefficients of first order from a linear regression model capture the linear influence of input variables on responses, but fail to characterize higher order influence. So, standardized interaction terms or second order standardized sensitivity coefficients from quadratic regression model are used to describe higher order (second order) influence of input variables on response values. Note that even though some input variables may not be influential by themselves, their cross products may have an influence on model responses. Thus, different order standardized sensitivity coefficients from linear and quadratic or even higher order regression models could be checked and an appropriate regression model selected for sensitivity analysis.

Other sensitivity analysis approaches can be found in the literature. Some of them are (1) variance based sensitivity analysis (Cukier et al, 1973; Cukier et al, 1978; Homma and Saltelli, 1996; Saisana et al, 2005; Saltelli et al, 2008; Ratto et al, 2009; Tong, 2010; Shahsavani and Grimvall, 2011), which is based on decomposition of response variance into conditional variances representing sensitivity of response on individual input variable, (2) sampling-based method or regionalized sensitivity analysis (Spear et al, 1994; Young, 1999; Saltelli et al, 2004; Helton et al, 2005), where sampling from the distribution of the response is carried out using Monte Carlo simulation (Metropolis and Ulam, 1949) and the values of input variables that cause specific values of the response are studied, and (3) emulator-based methods (Welch et al, 1992; Oakley and O'Hagan, 2004; Petropoulos et al, 2009; Petropoulos et al, 2010), where sensitivity analysis is carried out not on true complex unknown models of natural phenomena, but rather on simpler models derived from known values of input variables and model response. Comparison studies of sensitivity analysis techniques can also be found in the literature (Saltelli et al, 1993; Gacuci and Ionescu-Bujor, 2004a and 2004b; Chan et al, 2006).

Results of sensitivity study can be presented numerically and/or graphically. A graphical representation of sensitivity coefficients is easier to understand and interpret. Three main graphical plots for summarizing sensitivity analysis results are the tornado chart, spider diagram, and response surface maps (Welch et al, 1992; Deutsch et al, 2002). The first two are good for showing the results from local and global sensitivity study and the latter plot is only used in global sensitivity analysis. Conventional tornado charts present first-order sensitivity coefficients. Spider diagrams heavily depend on plausible variation ranges and units of input variables, and, therefore, a limited number of input variables may be plotted on spider diagrams. Moreover, construction of both conventional tornado charts and spider diagrams require a known relationship between the model responses and input variables, which is not always the case. For this reason an extended tornado chart is proposed, which is a modification of the conventional tornado chart. It can summarize not only first order, but also second order sensitivity coefficients from quadratic regression model, and is unit insensitive. Thus, among all discussed graphs for summarizing local sensitivity analysis results, the extended tornado chart based on regression approach is deemed the most effective one. Development of the extended tornado chart is an essential part of the paper. It is intended for quick and reliable visual assessment of the relationship between model responses and input variables and is used to rank input variables in order to decide whether an input variable should be retained in the model. Its application and properties are described in subsequent sections.

The algorithm of local sensitivity analysis based on regression approach and graphical representation of corresponding results are implemented in program SABOR (Sensitivity Analysis based on Regression). It utilizes the least squared error concept to compute regression coefficients. The program is robust, efficient and easy-to-use.

The paper is organized as follows. First, the objective of sensitivity analysis is defined. Then, the theoretical background of sensitivity analysis based on regression approach is presented. The extended tornado chart is developed, explained and compared to the conventional tornado chart. Second, the computational algorithm used in SABOR is explained. Three, the program SABOR, its parameter file and content of output files are described. Four, data from three different case studies on petroleum and mining engineering application are used to examine the proposed sensitivity analysis algorithm. Finally, conclusions are made and comments on possible future work are given. This paper is a continued and extended version of one year old paper on the same topic by the authors (Zagayevskiy and Deutsch; 2010).

2. Theoretical Background

Sensitivity analysis is a study that quantifies the influence of uncertainty in input variables X_i ($i = 1, ..., p$) on uncertainty in model responses Y_j ($j = 1, ..., r$), and, therefore, shows how uncertainty in input variables can explain uncertainty in responses (Saltelli et al, 2008). The results of sensitivity analysis rank input variables according to

their influence on model responses and help to identify whether an input variable should be retained or removed from the model, thus, more parsimonious models may be attained. Exclusion of the least important input variables does not worsen model quality (in some cases it even improves it) and keeps a model itself simple. For sake of the simplicity it is assumed that a model has only one response (*r = 1*). The program SABOR is also written to handle models with single response. Of course, the study could be repeated for multiple response variables and the relationship between response variables could be examined.

Local sensitivity analysis examines a limited region of variables' values in comparison to global sensitivity analysis (Saltelli et al, 2008). Local sensitivity analysis is favoured to global because there is often some information available on the expected value of the input variables. A sensitivity coefficient is represented as the partial derivative of the model response with respect to an input variable, with other input variables fixed at constant values (in our case at mean values). First and second order sensitivity coefficients are shown below (Equations (1) and (2)).

$$S_i = \frac{\partial Y}{\partial X_i}\bigg|_{X_k=\bar{x}_k, k=1,\dots,p, k\neq i} \tag{1}$$

$$S_{ij} = \frac{\partial^2 Y}{\partial X_i \partial X_j}\bigg|_{X_k=\bar{x}_k, k=1,\dots,p, k\neq i, k\neq j} \tag{2}$$

Sensitivity coefficients of every input variable can be derived simultaneously by regression analysis (Chatterjee and Hadi, 1988). A data set of sampled or simulated homotopic values is necessary in order to construct a regression model. The linear regression model or first order regression model is the simplest form of regression model that can be used to fit data. The next advanced model is the quadratic or second order regression model. We limit our analysis to quadratic regression models, since more complicated models often overfit the data and are complex to visualize. The general form of a regression model with a single response is shown in matrix notation in the Equation (3) and its estimate is shown in the Equation (4). The difference between models is the exclusion of an error term from and usage of coefficient estimates in the prediction model. A regression model is a probabilistic model whose coefficients must be estimated. The probabilistic part is depicted through error terms added to the deterministic model. For more details on regression analysis please refer to the literature (Devore and Peck, 2005; Johnson and Wichern, 2007).

$$\boldsymbol{Y} = \boldsymbol{Z} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3}$$

$$\hat{\boldsymbol{Y}} = \boldsymbol{Z} \cdot \boldsymbol{b} \tag{4}$$

where, **Y** is a column vector with the model response values; $\hat{Y}$ is a column vector with the predicted response values from the regression model; **Z** is a matrix with the first column consisting of 1's and other columns with values of input variables or predictors; **β** is the column vector with regression coefficients; **b** is the column vector with estimated regression coefficients; **ε** is the column vector with deviations from the regression fit (error vector). Detailed forms of these vectors and matrices are shown below for linear regression model. The model vectors and matrices look similar for the quadratic model.

$$\boldsymbol{Y}^T = \begin{bmatrix} y_{(1)} & y_{(2)} & \dots & y_{(n)} \end{bmatrix}_{1 \times n} \tag{5}$$

$$\boldsymbol{Z} = \begin{bmatrix} 1 & x_{1(1)} - \bar{x}_1 & x_{2(1)} - \bar{x}_2 & \dots & x_{p(1)} - \bar{x}_p \\ 1 & x_{1(2)} - \bar{x}_1 & x_{2(2)} - \bar{x}_2 & \dots & x_{p(2)} - \bar{x}_p \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1(n)} - \bar{x}_1 & x_{2(n)} - \bar{x}_2 & \dots & x_{p(n)} - \bar{x}_p \end{bmatrix}_{n \times (p+1)} \tag{6}$$

$$\boldsymbol{\beta}^T = \begin{bmatrix} \alpha & \beta_1 & \beta_2 & \dots & \beta_p \end{bmatrix}_{1 \times (p+1)} \tag{7}$$

$$\boldsymbol{\varepsilon}^T = \begin{bmatrix} \varepsilon_{(1)} & \varepsilon_{(2)} & \dots & \varepsilon_{(n)} \end{bmatrix}_{1 \times n} \tag{8}$$

$$\boldsymbol{b}^T = \begin{bmatrix} a & b_1 & b_2 & \dots & b_p \end{bmatrix}_{1 \times (p+1)} \tag{9}$$

where, $y_{(m)}$ is the mth value of model response Y; $x_{i(m)}$ is the mth value of ith input variable or predictor X_i; $\bar{x}_i$ is the mean value of ith input variable X_i; p is the number of input variables; n is the total number of values (data, observations, measurements, or realizations); α and a are the intercept and its estimate of the regression equation; β_i and b_i are the regression coefficient and its estimate of ith input variable X_i; $\varepsilon_{(m)}$ is the mth error term associated with m^{th} value of response Y, it is normally distributed with zero mean and constant variance; ^{T} is the transpose matrix operator; kxl is the vector or matrix size, where k is the number of rows, l is the number of columns.

Algebraic forms of linear and quadratic regression models and their estimates are shown in Equations (10) – (13).

$$y_{(m)} = \alpha + \sum_{i=1}^{p} \beta_i \cdot (x_{i(m)} - \bar{x}_i) + \varepsilon_{(m)} \tag{10}$$

$$y_{(m)} = \alpha + \sum_{i=1}^{p} \beta_i \cdot (x_{i(m)} - \bar{x}_i) + \sum_{i=1}^{p}\sum_{j=i}^{p} \beta_{ij} \cdot (x_{i(m)} - \bar{x}_i) \cdot (x_{j(m)} - \bar{x}_j) + \varepsilon_{(m)} \tag{11}$$

$$\hat{y}_{(m)} = a + \sum_{i=1}^{p} b_i \cdot (x_{i(m)} - \bar{x}_i) \tag{12}$$

$$\hat{y}_{(m)} = a + \sum_{i=1}^{p} b_i \cdot (x_{i(m)} - \bar{x}_i) + \sum_{i=1}^{p}\sum_{j=i}^{p} b_{ij} \cdot (x_{i(m)} - \bar{x}_i) \cdot (x_{j(m)} - \bar{x}_j) \tag{13}$$

where, β_{ij} and b_{ij} are the interaction term and its estimate between the ith and j^{th} input variables X_i and X_j in the quadratic regression model.

From Equations (10) and (11) it can be shown that intercept α of regression equations is $\alpha \approx E[Y]$ for linear model and $\alpha \approx E[Y] - \sum_{i=1}^{p}\sum_{j=i}^{p} \beta_{ij} \cdot COV[X_i, X_j]$ for quadratic model, where $E[Y]$ is the expected value operator of model response Y, and $COV[X_i, X_j]$ is the covariance operator for input variables X_i and X_j.

From estimates of the regression models (Equations (12) and (13)) it is evident that the regression coefficients b_i's indicate a change in the response value with a unit change of a predictor. Thus, regression coefficients characterize the contribution of input variables to the response model. Estimates of interaction terms b_{ij}'s show mutual influence of two input variables on model response values. Moreover, if Equations (12) and (13) are substituted into Equations (1) and (2), the similarity between regression and sensitivity coefficients is found. See Equations (14) – (16) for more details.

For the linear regression model:

$$S_i = \left.\frac{\partial Y}{\partial X_i}\right|_{X_k = \bar{x}_k, k=1,\dots,p, k \neq i} = \left.\frac{\partial\left(\alpha + \sum_{i=1}^{p} \beta_i \cdot (X_i - \bar{x}_i) + \varepsilon\right)}{\partial X_i}\right|_{X_k = \bar{x}_k, k=1,\dots,p, k \neq i} = \beta_i \tag{14}$$

For the quadratic regression model:

$$S_i = \left.\frac{\partial Y}{\partial X_i}\right|_{X_k = \bar{x}_k, k=1,\dots,p, k \neq i} =$$

$$= \left.\frac{\partial\left(\alpha + \sum_{i=1}^{p} \beta_i \cdot (X_i - \bar{x}_i) + \sum_{i=1}^{p}\sum_{j=i}^{p} \beta_{ij} \cdot (X_i - \bar{x}_i) \cdot (X_j - \bar{x}_j) + \varepsilon\right)}{\partial X_i}\right|_{X_k = \bar{x}_k, k=1,\dots,p, k \neq i} = \beta_i \tag{15}$$

$$S_{ij} = \frac{\partial^2 Y}{\partial X_i \partial X_j}\bigg|_{X_j = \bar{x}_j, X_k = \bar{x}_k, k=1,\ldots,p, l=1,\ldots,p, k \neq i, l \neq i} =$$

$$= \frac{\partial^2 \left(\alpha + \sum_{i=1}^{p} \beta_i \cdot (X_i - \bar{x}_i) + \sum_{i=1}^{p} \sum_{j=i}^{p} \beta_{ij} \cdot (X_i - \bar{x}_i) \cdot (X_j - \bar{x}_j) + \varepsilon \right)}{\partial X_i \partial X_j}\Bigg|_{X_j = \bar{x}_j, X_k = \bar{x}_k, k=1,\ldots,p, l=1,\ldots,p, k \neq i, l \neq i} = \beta_{ij} \tag{16}$$

Thus, regression coefficients from linear and quadratic regression models are identical to first order sensitivity coefficients and their names are used interchangeable. The interaction terms from quadratic regression model are identical to second order sensitivity coefficients.

In order to estimate coefficients in regression models the least squared error (LSE) concept is used, where the sum of squared differences between measured and predicted values of the model response is minimized. LSE concept can be shown in matrix form as in the Equation (17). In order to minimize LSE, which is a function of the regression coefficients, all partial derivatives of LSE with respect to regression coefficients should be set to zero and the resulting system of equations should be solved leading to estimates of the regression coefficients (Equation (18)).

$$LSE(\boldsymbol{b}) = \left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right)^T \left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right) = \left(\boldsymbol{Y} - \boldsymbol{Z} \cdot \boldsymbol{b}\right)^T \left(\boldsymbol{Y} - \boldsymbol{Z} \cdot \boldsymbol{b}\right) \tag{17}$$

$$\frac{\partial(LSE(\boldsymbol{b}))}{\partial \boldsymbol{b}} = 0 \quad \Rightarrow \quad \boldsymbol{b} = \left(\boldsymbol{Z}^T \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^T \boldsymbol{Y} \tag{18}$$

It is claimed that sigma-normalized or standardized sensitivity coefficients $b_{stand,i}$ are true measures of the response sensitivity to input variables (Oakley and O'Hagan, 2004; Saltelli et al, 2008). Their expression is shown in Equations (19) and (20). Also it is shown in Case Study I that sensitivity coefficients or regression coefficients b_i are good for quantification of the contribution of input variables to model response values and standardized sensitivity coefficients $b_{stand,i}$ are better for characterization of the contribution of uncertainty of input variables into uncertainty of model response. And, thus, the standardized coefficients are deemed more representative results of sensitivity analysis.

$$\boldsymbol{b}_{stand} = \boldsymbol{V} \cdot \boldsymbol{b} \tag{19}$$

where, $\boldsymbol{b}_{stand}$ is the column vector with sigma-normalized or standardized sensitivity coefficients, $\boldsymbol{V}$ is the standardizing matrix. Note that the term standardized sensitivity coefficient is not applicable to the intercept of the regression model. Matrix $\boldsymbol{V}_{lin}$ for linear regression coefficient has the following form:

$$\boldsymbol{V}_{lin} = \begin{bmatrix} 0 & 0 & 0 & \ldots & 0 \\ 0 & \dfrac{\sigma_{X_1}}{\sigma_Y} & 0 & \ldots & 0 \\ 0 & 0 & \dfrac{\sigma_{X_2}}{\sigma_Y} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \dfrac{\sigma_{X_p}}{\sigma_Y} \end{bmatrix}_{(p+1)\times(p+1)} \tag{20}$$

where, $b_{stand,i} = 0$ for $i = 1$ and for other $i \neq 1$: $b_{stand,i} = \dfrac{\sigma_{Xi}}{\sigma_Y} \cdot b$. For the quadratic regression model

second order standardized coefficients are found as $b_{stand,ij} = \dfrac{\sigma_{Xi} \cdot \sigma_{Xj}}{\sigma_Y^2} \cdot b$.

Once the sensitivity and standardized sensitivity coefficients are computed, input variables are ranked according to these coefficients, and corresponding tornado charts are plotted. Conventional tornado charts utilize specified quantile values of the distribution of input variables as shown in the Figure 1 (Deutsch et al, 2002). In

order to plot tornado charts a mathematical model of the system should be known, which is not always the case. Lower and upper quantiles of the cumulative distribution function of input variables are selected (usually they are 0.10 and 0.90 or 0.05 and 0.95 probability quantiles) and corresponding values of input variables are defined. These quantile values of single input variable are used in the model equation to calculate corresponding values of the model response, while other input variables are fixed at their means. The resulting range of response values is measured and stored. The procedure is repeated for each input variable. In the end, variation ranges of model responses are compared, and input variables are ranked in descending order. These variations are plotted vertically as bars with names of the input variables indicated. The lengths of the bars are compared to assess the contribution of the input variables to the model response. The major drawback of conventional tornado chart is its dependence on the relationship between model response and predictors, which is expressed through a known equation.

The conventional tornado chart is developed further; see the sample extended tornado chart for plotting sensitivity coefficients and their standardized values from linear and quadratic models on the Figure 2. Note that charts for different regression models are slightly different due to interaction terms in the quadratic regression model. The difference in charts for different regression models can be tracked from case studies. The title of extended tornado chart and name of model response are depicted in the upper middle of the plot. The main middle part of the extended tornado chart is devoted to bars of two kinds. The bars on the right side of tornado represent either sensitivity coefficients or their standardized values and are plotted in descending order. Positive coefficients are shown in yellow color, and negative coefficients are drawn in orange. The red box on the bar end shows the confidence interval of the coefficient at a specified confidence level. The confidence interval provides information about uncertainty in the coefficients. Names of corresponding input variables are shown at the left. Other bars on the left side of the tornado represent either sensitivity coefficients or their standardized values without uncertainty, if a linear regression model is examined, or either interaction terms or their standardized values, when a quadratic regression model is examined. The bars are plotted in the same order as bars on the right side of tornado. Positive interaction terms are plotted in green, negative ones are in blue. Confidence intervals of interaction terms at a specified confidence level are shown as a grey box. All bars are scaled to the largest upper boundary of confidence level of any coefficient or interaction term. Usually the regression coefficients have the largest value, but interaction terms may have larger values in some cases. Input variables with relatively short bars of coefficients and interaction terms are candidates to be discarded from the model. If all interaction term bars are short, the linear model is appropriate for the given data set. Summary statistics including coefficient of determination, its adjusted value, standard error of model deviations, model utility test based on F statistic, and prediction power of the model (percentage ratio of standard deviations of predicted and actual values of model response) are presented in the upper left corner of the plot. The table with means, standard deviations, coefficients of correlation and variation, sensitivity coefficients and standardized sensitivity coefficients for model response and each input variable are tabulated in the table on the right of the chart. Coefficients, whose values are shown in blue in the table, are used for plotting bars. The specified confidence level is reported in the lower middle of the chart. Thus, the extended tornado chart visually summarizes results of sensitivity analysis and is useful for making decisions on the importance of input variables and the appropriateness of linear and quadratic models.

3. Algorithm of Sensitivity Analysis Based on Regression

The adopted algorithm for sensitivity analysis based on the regression approach for selection of important variables in linear or quadratic regression models is as follows (Devore and Peck, 2005; Johnson and Wichern; 2007):

1. Define the response variable and relevant input variables.
2. Generate correlation coefficients and scatter plots between model response and input variables. Input variables that have high correlation with response (correlation of 0.5 could be taken as a threshold) and input variables that exhibit any other nonlinear relationship with response should be retained in the model. Other input variables could be discarded. Input variables could be transformed to improve the relationship. Outliers and other extreme data should be considered for removal from the data set and the correlation should be recalculated for the values retained.
3. Construct the linear and quadratic regression models. The program SABOR could be used to compute coefficients for both linear and quadratic regression models simultaneously using LSE concept. The regression coefficients that correspond to sensitivity coefficients should be stored and the corresponding

standardized sensitivity coefficients should be calculated. When a regression model is used for prediction, extrapolation for values lying beyond the range of our data should be avoided, since most of the time weak predictions would be obtained.

4. Assess the goodness of model fit. There are several measures used to assess model quality.

4a. Coefficient of determination R^2, its adjusted value (adjusted R^2), and standard deviation of error s_e (standard deviation about least-squares or regression line) are calculated. For good model fit it is expected to have R^2 close to 1.0 or 100% and its adjusted value close to the unadjusted value and s_e should be low. Computed values for both models are compared and preliminary selection of linear or quadratic models is made. The model with higher R^2 and adjusted R^2 and lower s_e is desirable, but the simplest model should be favoured as well.

$$R^2 = 1 - \frac{SSResid}{SSTo} \quad \text{or} \quad R_{\%}^2 = 100\% \cdot \left(1 - \frac{SSResid}{SSTo}\right) \tag{21}$$

$$R_{adj}^2 = 1 - \left[\frac{n-1}{n-(k+1)}\right] \cdot \frac{SSResid}{SSTo} \quad \text{or} \quad R_{adj,\%}^2 = 100\% \cdot \left(1 - \left[\frac{n-1}{n-(k+1)}\right] \cdot \frac{SSResid}{SSTo}\right) \tag{22}$$

$$s_e = \sqrt{\frac{SSResid}{n-(k+1)}} \tag{23}$$

where, *SSResid* is the sum of squares of residuals; *SSTo* is the total sum of squares; $df = k+1$ is the degree of freedom of a model; k is the number of coefficients in a model. Thus, $k_1 = p$ for linear model and $k_2 = p \cdot (p+3)/2$ for quadratic model.

$$SSResid = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{24}$$

$$SSTo = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{25}$$

where, r_i is the residual of ith model response value y_i; $\bar{y}$ is the average of model response values.

4b. Make Q-Q plots for actual and predicted values of response from linear and quadratic regression models. The Q-Q plot with dots lying on straighter 45 degree line indicates better regression model.

4c. Residuals or even better their standardized values (Equation (26)) should be examined. Residuals are estimates of model error, and hence should satisfy the following four requirements in order to honour plausibility of the derived regression models:

- o Residuals should be normally distributed
- o Residuals should have zero mean
- o Residuals should have the same variance
- o Residuals should be independent from each other

$$r_{stand,i} = \frac{r_i}{s_{ri}} = \frac{y_i - \hat{y}_i}{\sqrt{s_e^2 \cdot (1 - (\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T)_{ii})}} \tag{26}$$

where, s_{ri} is the standard deviation of ith residual r_i.

Residual scatter plots and normal probability plots of the residuals or their standardized values are used to check conformity to the four assumptions listed above. Residuals should be scattered randomly on scatter plots between the response value and the residual and between the predictor value and the residual. No pattern should be present. Outlier and extreme values should be considered for removal from the data set. Normal probability plots of the residuals should show a straight line that indicates a normal distribution of the residuals. If these requirements are far from expected, the plausibility of the model is not established and another model should be found.

4d. Model utility test based on *F* statistic should be conducted for a model that has more than one predictor. The *F*-test checks null hypothesis H_o that all regression coefficients are zero ($\beta_1 = \beta_2 = ... = \beta_k = 0, k > 1$) against alternative hypothesis H_a stating that at least one coefficient is nonzero. Computed *F*-value, which is called observed *F*-value, is compared to critical *F*-value at specified confidence level alpha for two tailed tests. If observed

F-value is greater than critical *F*-value or if corresponding observed *P*-value is less than critical *P*-value, the null hypothesis H_o is rejected at level alpha. Otherwise H_o is not rejected.

$$Observed\ F\ value = \frac{SSRegr\ /\ df_1}{SSResid\ /\ df_2} \tag{27}$$

$$SSRegr = SSTo - SSResid \tag{28}$$

where, $df_1 = k$ is the degree of freedom of numerator or regression sum of squares (*SSRegr*), $df_2 = n - (k+1)$ is the degree of freedom of residual sum of squares (*SSResid*).

Once the model is deemed statistically plausible, each regression coefficient should be examined. This test is called *t* ratio test and is based on the *t* distribution. The *t* ratio test substitutes model utility test for a model with a single input variable. The test checks the statistical significance of H_o: $\beta_i = 0$ against H_a: $\beta_i \neq 0$ for every regression coefficient and intercept at specified confidence level alpha for two tailed tests at $df = n - (k+1)$ degrees of freedom. If observed *t* ratio is greater than critical *t* ratio or if corresponding observed *P*-value is less than critical *P*-value, the null hypothesis H_o is rejected at confidence level alpha.

$$Observed\ t_i\ value = \frac{b_i}{s_{b_i}} \tag{29}$$

where s_{bi} is the standard deviation of ith regression coefficient or intercept, whose value is defined as:

$$s_{b_i} = \sqrt{s_e^2 \left(\mathbf{Z}^T \mathbf{Z}\right)_{ii}^{-1}} \tag{30}$$

As a result of *t* ratio test input variables with statistically significant regression coefficients should be retained in the model, but other variables should be considered for removal. One could go back to step 3 and build the regression model based on the retained input variables. Once no more predictors can be discarded from a model and the appropriate order of the regression model is selected, it can be said that a final plausible model is found.

5. Calculate confidence intervals of means of regression coefficients and means of predicted response and prediction interval of response at given confidence level alpha and plot them.

In case the inference from the *t* ratio test suggests that the regression coefficient has a value different from zero, the confidence interval of its estimated mean at the specified confidence level alpha can be calculated as follows and is plotted as a box in the extended tornado chart:

$$CI\ for\ mean\ of\ \beta_i = b_i \pm t_{critical,alpha/2,n-(k+1)} \cdot s_{b_i} \tag{31}$$

The confidence and prediction intervals of the model response values are found using equations shown below and may be plotted against values of any input variable.

$$CI\ for\ mean\ of\ y = \hat{y} \pm t_{critical,alpha/2,n-(k+1)} \cdot s_{\hat{y}} \tag{32}$$

$$PI\ for\ y = \hat{y} \pm t_{critical,alpha/2,n-(k+1)} \cdot \sqrt{s_{\hat{y}}^2 + s_e^2} \tag{33}$$

$$s_{\hat{y}} = \sqrt{s_e^2 \mathbf{z_0} \left(\mathbf{Z}^T \mathbf{Z}\right)^{-1} \mathbf{z_0}} \tag{34}$$

$$\mathbf{z}_0^T = \begin{bmatrix} 1 & x_{1(0)} - \bar{x}_1 & x_{2(0)} - \bar{x}_2 & ... & x_{p(0)} - \bar{x}_p \end{bmatrix} \tag{35}$$

where, z_0 is the column vector with deviation of values from the input variables' mean.

6. Summarize all results in the form of the extended tornado chart for linear and quadratic models. Importance of input variables is shown on and appropriate regression model is derived from the extended tornado charts.

4. Description of a Program SABOR

The algorithm for sensitivity analysis described above is implemented in the program SABOR. The program is written in FORTRAN and run through the command line (Hahn, 1996). Several subroutines are adopted from literature (Press et al, 1997). A file with the data is used as input to the program and seven output files are generated. Brief summary statistics are presented in the command line window to facilitate analyzing the data and understanding the goodness of both models. The data set should be stored in GSLib format for input to SABOR (Deutsch and Journel, 1998).

The program SABOR generates two output files with sensitivity coefficients and standardized sensitivity coefficients with corresponding confidence intervals at the specified level for both linear and quadratic regression models, respectively. Two more output files are devoted for predicted response values from both models with confidence and prediction intervals, difference between actual and predicted model response and their standardized values. The fifth output file contains summary statistics and includes mean, standard deviation and coefficient of variation for every variable, covariance and correlation matrices between all variables, coefficient of determination, standard error and equations of constructed linear and quadratic regression models, checks on the utility of the models and plausibility of corresponding coefficients at specified confidence level through F and t test statistics and ANOVA table. The last two output files are designed for plotting extended tornado charts with table of summary statistics using computed sensitivity coefficients or their standardized values for both regression models. The first four output files are text files in GSLib format, the fifth file is in free format, and last two output files have PostScript format (Warnock and Geschke, 1999).

A parameter file is required to run the program. A template is created automatically if no file is present. An example is shown in the Figure 1. On the first line, the name of the input file with data is specified. The number of input variables is on the next line. Then, column numbers of input variables in data file are specified according to their order in the input file. The column number of the single response is defined on the fourth line. The confidence level alpha for model utility, t ratio tests and confidence and prediction intervals are entered on the fifth line. The name for the output file for regression/sensitivity coefficients is entered on sixth line. The seventh line has the name of the output file for predicted response values. Next, eighth line is used to specify name of the output file for summary statistics. Ninth line has the name of the output file for extended tornado charts. Coefficients used in plotting of extended tornado charts are specified on the last tenth line: (1) sensitivity coefficients or (2) standardized sensitivity coefficients. Prefix `lin_` is added to output files for linear regression model and `quad_` – for quadratic regression model.

```
              Parameters for SABOR
              ********************

       START OF PARAMETERS:
1-     data.dat            -input file with data
2-     2                   -   number of predictors/input parameters
3-     1   2               -   column number(s) with predictor(s)
4-     3                   -   column number with single response
5-     0.05                -confidence level alpha
6-     coef.out            -output file for regression/sensitivity coefficients
7-     response.out        -output file for actual and predicted response values
8-     sumstat.out         -output file for summary statistics
9-     etrnplt.ps          -output file for extended tornado chart
10-    1                   -   plot coefficients (1) or standardized values (2)
```

Figure 1: Sample parameter file of program SABOR

5. Case Studies

Two different data sets are examined in order to demonstrate the efficacy of the algorithm for sensitivity analysis based on regression approach, to assist in selection of appropriate regression model (linear or quadratic) that fits the data the best and to show that sensitivity coefficients are good for understanding the contribution of input variables into model, standardized sensitivity coefficients are proper measure for revealing how uncertainty in an

individual input variable contributes to uncertainty in model response. Although the examined data sets cover petroleum and mining concepts, the proposed algorithm of sensitivity analysis and associated program can be applied to data from other industrial sectors and scientific branches.

5.1 Case Study I

This case study examines the relationship between iron content and other minerals deposited in a lateritic deposit. The model response is iron content *Fe* in mass percentage (%) and six input variables are nickel (*Ni*), quartz (*SiO₂*), magnesia (*MgO*), cobalt (*Co*), alumina (*Al₂O₃*), and chromium (*Cr₂O₃*), which has same units. A total of 12439 samples are used in modeling.

Scatter plots (Figure 4) show from fair to strong linear relationships between model response and input variables. There is a very weak linear relationship between iron and nickel. However, a nonlinear trend is seen. It is found that linear and quadratic regression models and all coefficients and interaction terms except cobalt – cobalt, are statistically significant at a confidence level of 0.05. So, we can consider the regression models significant.

From the extended tornado chart for the linear regression model it is obvious that cobalt has the highest contribution to the model and all other variables have almost the same model contribution. In the standardized coefficients plot, magnesia and quartz are the most influential variables. All other variables have almost equal influence on the model response. From the tornado charts for the quadratic regression models based on sensitivity coefficients it is evident that beside cobalt only the magnesia – magnesia interaction term plays an important role, though not the largest. From the extended tornado chart based on standardized coefficients it is evident that only the magnesia – magnesia interaction term is influential. So, the data set probably can be described well by the linear regression model, since only single interaction term is significant. Also, no big difference is found between the coefficients of determination of linear and quadratic models (R^2_{linear} = 98.49% against $R^2_{quadratic}$ = 99.97%). The equality of the models is seen on Q-Q plots, where values of actual response are plotted against predicted values from linear and quadratic regression models (Figure 5).

A small sensitivity study is conducted in order to understand what variables may be discarded from the model. If a linear model is selected for modeling of iron, *Co* is deemed important based on sensitivity coefficients and *MgO* and *SiO₂* are deemed important based on standardized sensitivity coefficients. Also from scatter plots it is said that *Ni* and *Fe* do not possess a linear relationship. Thus, six different cases are examined in order to understand the importance of the input variables (Table 1).

Table 1: Influence of inclusion of input variables on quality of linear regression model of iron *Fe*

Input variables included into model	Coefficient of Determination, R^2	Standard Deviation of Error, s_e
All	98.48%	1.416%
Co	76.92%	5.543%
MgO and *SiO₂*	96.85%	2.049%
MgO, *SiO₂* and *Co*	96.89%	2.035%
All except *Ni*	97.92%	1.662%
Ni	2.03%	11.419%

The input variables that have higher standardized sensitivity coefficients are more important for model quality than input variables with higher sensitivity coefficients. Exclusion of *Ni* from the linear regression model did not change the results very much. So, as a general rule, the more variables we include in the model, the better prediction results are. It is always safe to include all variables that are deemed significant based on sensitivity coefficients and their standardized values. However, for the sake of simplicity only *MgO*, *SiO₂* and *Co* would be included in the model.

The same table is constructed for the quadratic model. *Co* and *MgO* – *MgO* have the highest contribution to the model, while the response is the most sensitive to the interaction *MgO* – *MgO*. The importance of *Ni* is also examined here, which was deemed unimportant in the linear regression model. From the Table 2 it is found that all variables contribute to model quality, if they are somehow related to the model response (linearly or not).

Table 2: Influence of inclusion of input variables on quality of quadartic regression model of iron *Fe*

Input variables included into model	Coefficient of Determination, R^2	Standard Deviation of Error, s_e
All	99.97%	0.207%
MgO and *Co*	98.87%	1.225%
MgO	97.47%	1.834%
All except *Ni*	99.65%	0.688%
Ni	95.55%	2.434%

As a conclusion, linear regression models with input variables *MgO*, *SiO$_2$* and *Co* are selected to construct a model for iron. Its expression is shown in the Equation 36. Other input variables, which are *Ni*, *Al$_2$O$_3$*, and *Cr$_2$O$_3$*, are discarded from the model.

$$Fe = 16.41 - 0.62 \cdot (MgO - 21.93) - 0.49 \cdot (SiO_2 - 37.07) + 13.52 \cdot (Co - 0.05) \tag{36}$$

5.2 Case Study II

This case study is devoted to parameters of a Steam Assisted Gravity Drainage (SAGD) bitumen extraction method. The model response is bitumen production in millions of cubic meters after 10 years. Input variables are base continuous bitumen (*BaseContBitumen*), continuous bitumen (*ContBitumen*), reservoir porosity (*Porosity*) and water saturation (*Water Saturation*). The data are retrieved from a simulation study of 5778 experiments.

Scatter plots suggest that the volume of continuous bitumen has a good linear relationship with bitumen production, while the base continuous bitumen and water saturation exhibit weak linear relationships (Figure 8). Although the correlation between the model response and porosity is pretty high it is more likely that a quadratic model would fit better. These results suggest that base continuous bitumen might not be a useful predictor in linear regression model, and the interaction term of porosity – porosity may be influential in quadratic regression model.

It is found that linear and quadratic regression models and all coefficients and interaction terms are statistically significant at a confidence level 0.05. Extended tornado charts show the input variables that have a higher contribution to the model and which ones have the highest influence on the model response change. Porosity and water saturation have the highest contribution (positive contribution of porosity and negative contribution of water saturation) to the linear regression model, while production is more sensitive to continuous bitumen and porosity (Figure 10). So it is deemed that all variables except base continuous bitumen should be included in the linear regression model. Values of coefficients of determination for linear regression model are compared for different cases and tabulated in the Table 3. Predictors with highest values of standardized sensitivity coefficients are more important than predictors with highest sensitivity coefficients. It is safe to include all influential predictors regardless of what chart they are coming from. Predictors that have smallest bars in all tornado charts and do not exhibit linear relationship with model response can be safely removed from the linear regression model.

Table 3: Influence of inclusion of input variables on quality of linear regression model of bitumen production

Input variables included into model	Coefficient of Determination, R^2	Standard Deviation of Error, s_e
All	89.19%	249.390
Porosity and *Water Saturation*	53.33%	518.009
ContBitumen and *Porosity*	86.30%	280.667
All except *BaseContBitumen*	88.93%	252.330
BaseContBitumen	23.71%	662.241

From the extended tornado chart (Figure 11) it is evident that the quadratic regression model does much better than the linear. The coefficient of determination is higher and standard error is lower for quadratic regression model: $R^2_{quadratic}$ = 99.83% and *Std.Err.$_{quadratic}$* = 30.847 against R^2_{linear} = 89.19% and *Std.Err.$_{linear}$* = 249.39. If Q-Q plots are compared (see the Figure 9), it is clear that the quadratic model is fitted to response values better.

Keep in mind that we want to avoid danger of extrapolation and overfitting. Again porosity and continuous bitumen seem to be the major contributors to the model, while continuous bitumen alone is the major influential factor. The most interesting scenarios are compared and summarized in the Table 4. Inclusion of any extra input variable, which is somehow related to the response, in quadratic regression model will improve prediction of model response. As a conclusion, quadratic regression model is selected for this data set with all variables being predictors.

Table 4: Influence of inclusion of input variables on quality of quadratic regression model of bitumen production

Input variables included into model	Coefficient of Determination, R^2	Standard Deviation of Error, s_e
All	99.83%	30.847
Porosity and *ContBitumen*	97.00%	131.467
ContBitumen	92.17%	212.196
All except *BaseContBitumen*	97.48%	120.540
BaseContBitumen	91.79%	217.637

6. Conclusion

This paper presents details of the application of sensitivity analysis based on regression through a developed program `sabor.exe` and emphasizes the appropriateness of an extended tornado charts for summarizing sensitivity analysis results. The approach and program should be useful for practitioners and theorists, since it generates statistical test results and visualizes computed coefficients. Linear and quadratic regression models are constructed for examined data set in two case studies and a procedure for selecting a proper model is shown. Coefficients of determination R^2, their adjusted values, standard deviation of error s_e, and Q-Q plots of both regression models should be compared. If the difference in R^2 is small it is better to go with the simpler linear regression model, otherwise the model with the highest R^2 should be selected. The importance of input variables should be judged based on both types of tornado charts for linear and quadratic regression models, which are based on sensitivity coefficients and their standardized values. Tornado charts plotted using sensitivity or regression coefficients rank the input variables in descending order according to their contribution to the model, which are unit sensitive. Tornado charts plotted using standardized sensitivity coefficients rank input variables in descending order according to sensitivity level of model response change to change in input variables, which are unit insensitive. The input variables with the highest ranks in both charts should be retained in the model, other variables may be discarded. Predictors with highest standardized sensitivity coefficients are the most important for the linear regression model quality; hence, they should be always retained in the model. For quadratic regression model it is slightly different, any extra variable, which is related to model response in any fashion, improves the model quality.

Sensitivity analysis based on regression approach is claimed to be efficient, robust, and simple. The program `sabor.exe` visually summarizes results of sensitivity analysis and suggests what type of regression model, linear or quadratic, should be selected for modeling of response and what input variables should be retained and discarded from the model. Sensitivity analysis algorithm, program and extended tornado charts can be extended for higher order regression models.

References

Azmy, Y., Sratori, E., 2010, *Nuclear Computational Science: A Century in Review*. Springer, New York

Balakrishnan, N., Castillo, E., Srabia, J.M., 2006, *Advances in Distribution Theory, Order Statistics, and Inference*, Birkhäuser, Boston

Chan, K., Saltelli, A., Tarantola, S., 1997, Sensitivity Analysis of Model Output: Variance-based Methods Make the Difference, *Proceeding of the 1997 Winter Simulation Conference*, 261 – 268

Chatterjee, S., Hadi, A.S., 1998, *Sensitivity Analysis in Linear Regression,* John Wiley & Sons, USA

Conejo, A.J., Castillo, E., Mínguez, R., García-Bertrand, R., 2006, *Decomposition Techniques in Mathematical Programming: Engineering and Science Applications,* Springer, The Netherlands

Cukier, R.I., Fortuin, C.M., Shuler, K.E., Petschek, A.G., Schaibly, J.H., 1973, Study of the Sensitivity of Coupled Reaction Systems to Uncertainties in Rate Coefficients. I Theory. *The J. of Chem. Phys*., 59(8), 3873 – 3878

Cukier, R.I., Levine, H.B., Shuler, K.E., 1978, Review: Nonlinear Sensitivity Analysis of Multiparameter Model Systems. *J. of Comput. Phys.*, 26, 1 – 42

Deutsch, C.V., Journel, A.G., 1998, *GSLIB: Geostatistical Software Library and User's Guide, 2nd Ed.*, Oxford University Press, New York

Deutsch, C.V., Zanon, S., Leuangthong, O., 2002, Procedures and Guidelines for Assessing and Reporting Uncertainty in Geostatistical Reservoir Modeling, *Cent. for Comput. Geostatistics* 4, 18 – 22

Devore, J., Peck, R., 2005, *Statistics: The Exploration and Analysis of Data, 5th Ed.,* Thomson Brook/Cole, Belmont, California

Gacuci, D.G., Ionescu-Bujor, M., 2004, A Comparative Review of Sensitivity and Uncertainty Analysis of Large-Scale Systems – I: Deterministic Methods, *Nucl. Sci. and Eng.*, 147(3), 189 – 203

Gacuci, D.G., Ionescu-Bujor, M., 2004, A Comparative Review of Sensitivity and Uncertainty Analysis of Large-Scale Systems – II: Statistical Methods, *Nucl. Sci. and Eng.*, 147(3), 204 – 217

Gacuci, D.G., 2003, *Sensitivity and Uncertainty Analysis, Volume 1: Theory,* Chapman & Hall/CRC, USA

Hahn, B.D., 1996, *Introduction to Fortran 90 for Scientists and Engineers*, J.W. Arrowsmith Ltd, Bristol

Helton, J.C., Johnson, J.D., Salaberry, C.J., Storlie, C.B., 2005, Survey of sampling based methods for uncertainty and sensitivity analysis, *Reliab. Eng. and Syst. Saf.*, 91, 1175–1209

Homma, T., Saltelli, A., 1996, Importance Measures in Global Sensitivity Analysis of Nonlinear Models, *Reliab. Eng. and Syst. Saf.*, 52, 1 – 17

Johnson, R.A., Wichern, D.W., 2007, *Applied Multivariate Statistical Analysis, 6th Ed.,* Pearson Education International, New Jersey

Jorgensen, S.E., 1994, *Fundamental of Ecological Modeling,* Elsevier, Amsterdam

Metropolis, N., Ulam, S., 1949, The Monte Carlo Method, *J. of the Am. Stat. Assoc.*, 44(247), 335 – 341

Oakley, J.E., O'Hagan, A., 2004, Probabilistic Sensitivity Analysis of Complex Models: A Bayesian Approach, *J. of the R. Stat. Soc.*, 66, 751 – 769

Petropoulos, G.P., Wooster, M.J., Carlson, T.N, Kennedy, M.C., Scholze, M., 2009, A Global Bayesian Sensitivity Analysis of the 1D SimSphere Soil-Vegetation-Atmospheric Transfer (SVAT) Model using Gaussian Model Emulation, *Ecol. Model.*, 220(19), 2427 – 2440

Petropoulos, G.P., Ratto, M., Tarantola, S., 2010, A Comparative Analysis of Emulators for the Sensitivity Analysis of a Land Surface Process Model, *Sixth International Conference on Sensitivity Analysis of Model Output*, 7716 – 7717

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1997, *Numerical Recipes in Fortran 77: The Art of Scientific Computing, Vol. 1, 2nd Ed.*, Press Syndicate, Cambridge

Ratto, M., Pagano, A., Young, P.C., 2009, Non-Parametric Estimation of Conditional Moments for Sensitivity Analysis, *Reliab. Eng. and Syst. Saf.,* 94, 237 – 243

Saisana, M., Saltelli, A., Tarantola, S., 2005, Uncertainty and Sensitivity Analysis Techniques as Tools for the Quality Assessment of Composite Indicators, *J. of the R. Stat. Soc.*, 168(2), 307 – 323

Saltelli, A., Andres, T.H., Homma, T., 1993, Sensitivity Analysis of Model Output: An Investigation of New Techniques, *Comput. Stat. and Data Anal., 15*, 211 – 238

Saltelli, A.M., Ratto, M., Andres, T., 2008, *Global Sensitivity Analysis: The Primer*, John Willey & Sons Ltd, Chichester

Saltelli, A.M., Tarantola, S., Campolongo, F., Ratto, M., 2004, *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models,* John Willey & Sons Ltd, Chichester

Shahsavani, D., Grimvall, A., 2011, Variance-based Sensitivity Analysis of Model Outputs using Surrogate Models, *Environ. Model. & Softw.*, 26(6), 723 – 730

Spear, R.C., Grieb, T.M., Shang, N., 1994, Parameter Uncertainty and Interaction in Complex Environmental Models, *Water Resour. Res.*, 30(11), 3159 – 3169

Tong, C., 2010, Self-Validated Variance-based Methods for Sensitivity Analysis of Model Outputs, *Reliab. Eng. and Syst. Saf. 95*, 301- 309

Warnock, J., Geschke, C., 1999, *PostScript: Language Reference, Adobe Systems Incorporated, 3rd Ed.*, Addison-Wesley Publishing Company, USA

Welch, W.J., Buck, R.J., Sacks, J., Wynn, H.P., Mitchell, T.J., Morris, M.D., 1992, Screening, Predicting, and Computer Experiments, *Technometrics*, 34(1), 15 – 25

Young, P., 1999, Data-based Mechanistic Modelling, Generalised Sensitivity and Dominant Mode Analysis, *Comput. Phys. Commun.*, 117(1- 2), 113 – 129

Zagayevskiy, Y. and Deutsch, C.V., 2010, Calculation and Display of Sensitivity Coefficients, *Centre for Computational Geostatistics* 12, 127-1 – 127-13
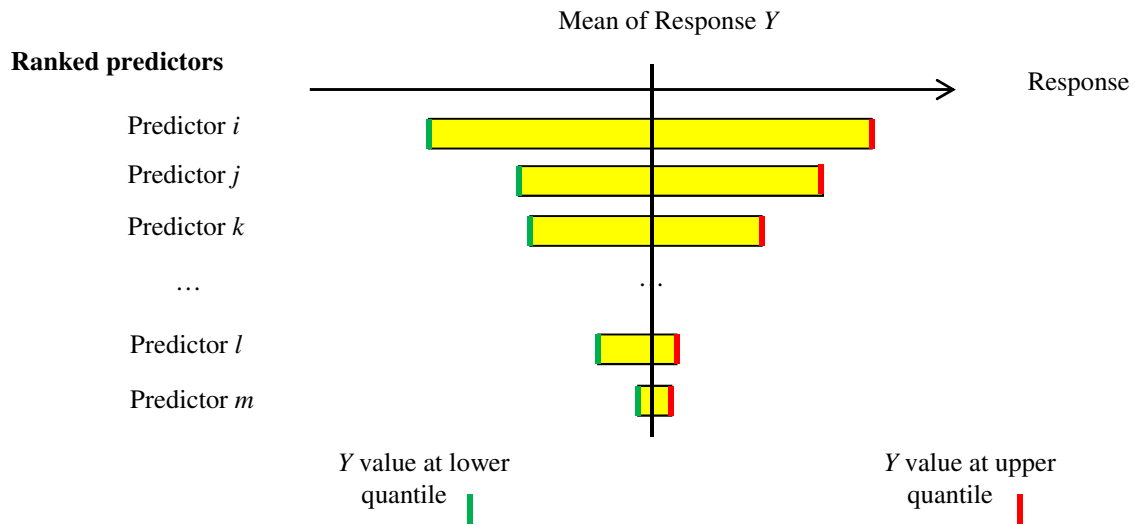
Figure 2: Schematic of conventional tornado chart
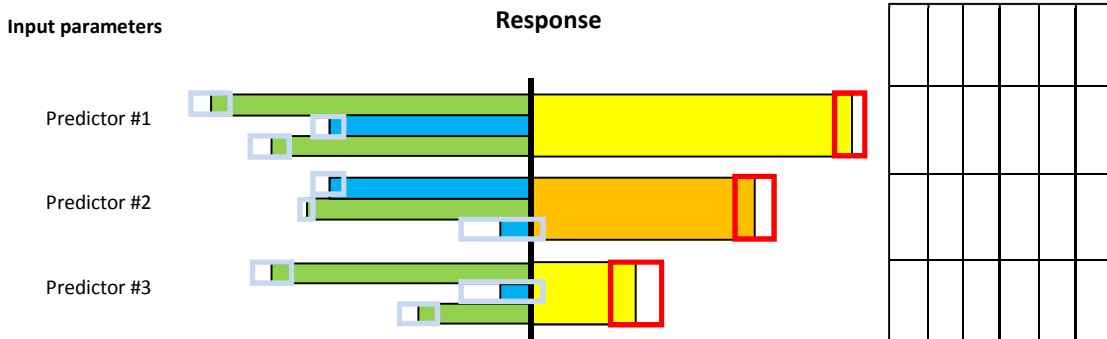
Figure 3: Schematic of extended tornado chart for quadratic regression model
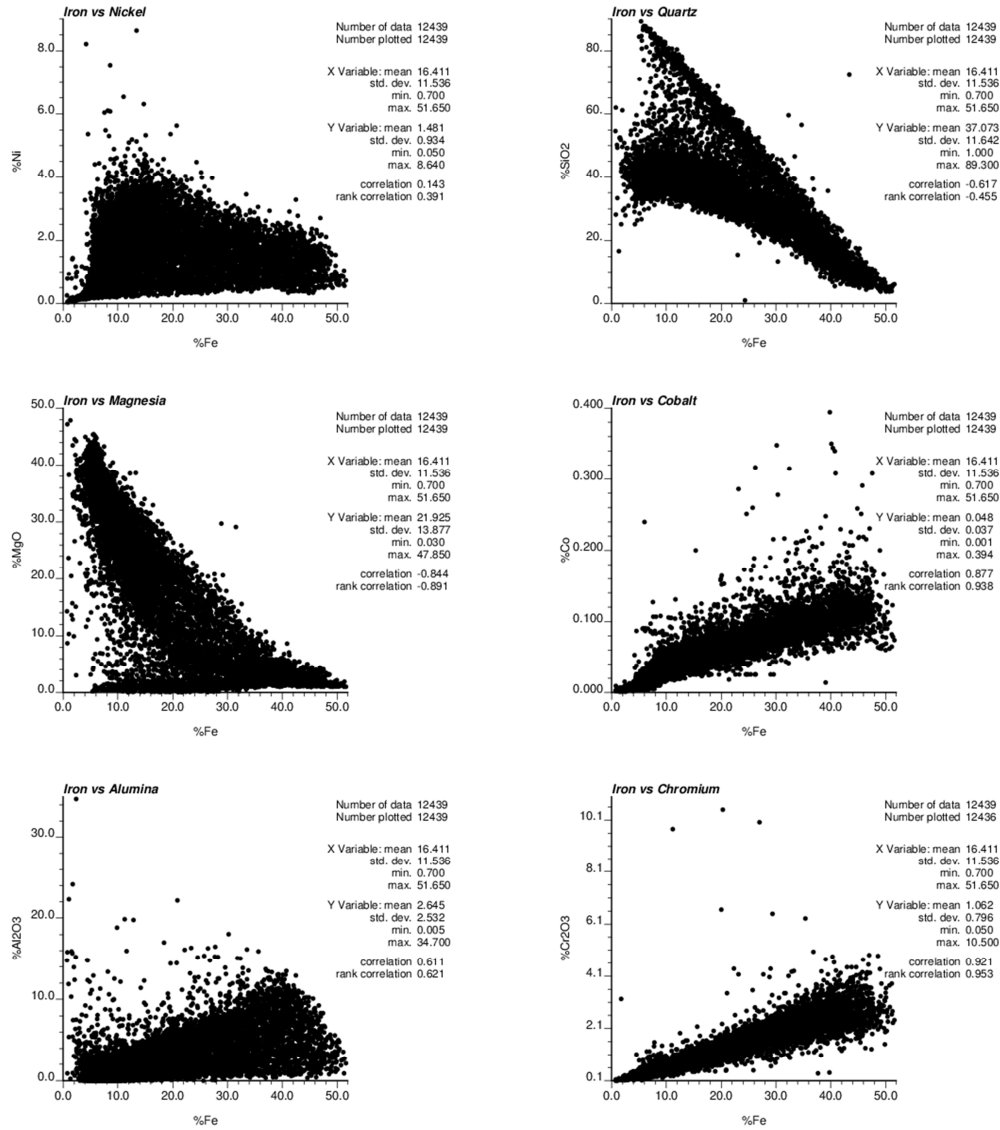
Figure 4: Scatter plots between model response (iron) and input variables (nickel, quartz, magnesia, cobalt, alumina, and chromium)
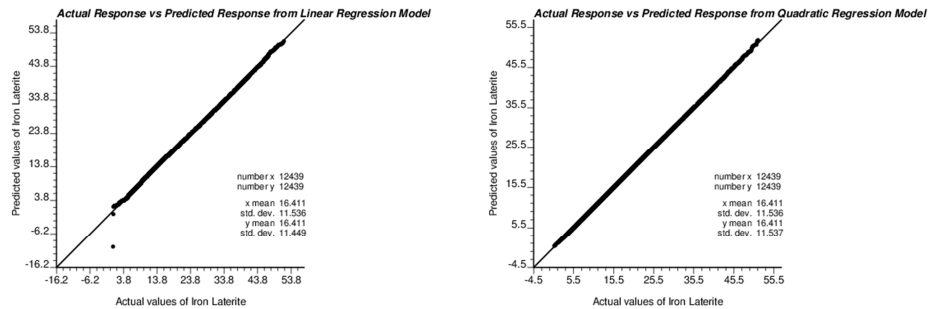
Figure 5: Q-Q plots for actual and predicted values of model response *Fe* from linear (left) and quadratic (right) regression models
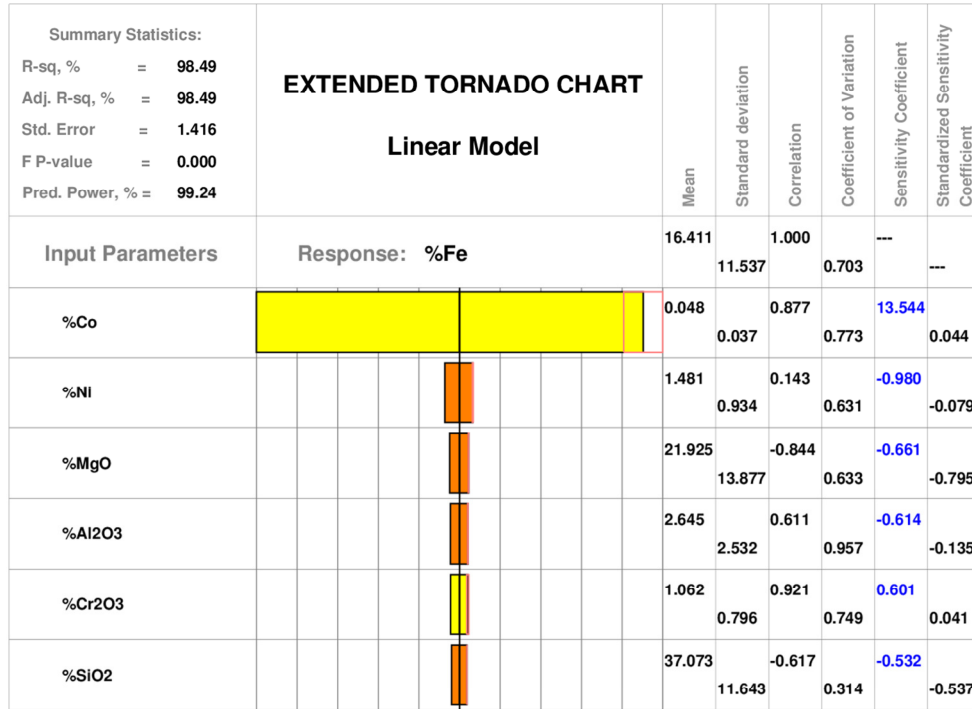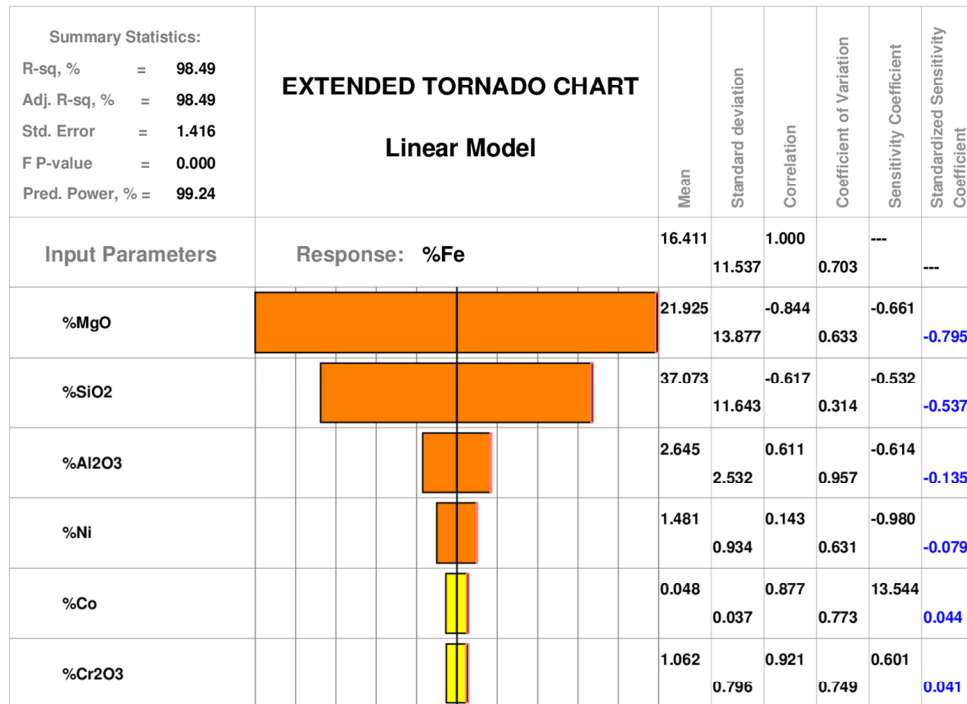
Figure 6: Extended tornado charts for linear regression model of *Fe*. Bars on upper plot are based on values of sensitivity coefficients, bars on lower plot are based on values of standardized sensitivity coefficients
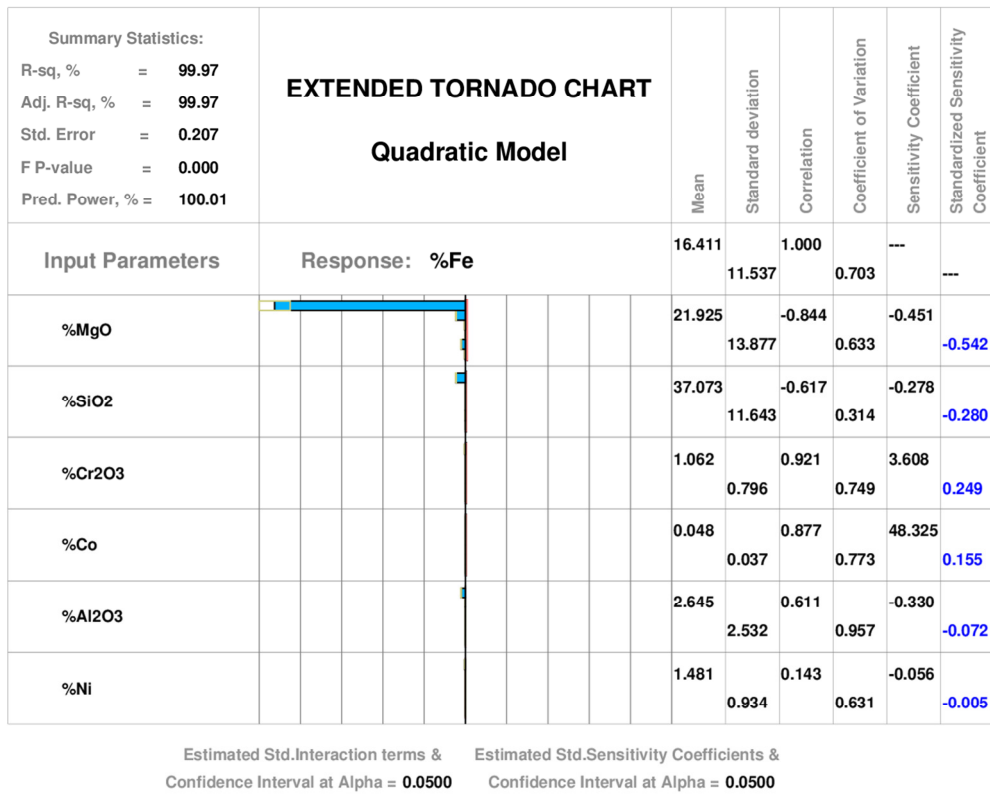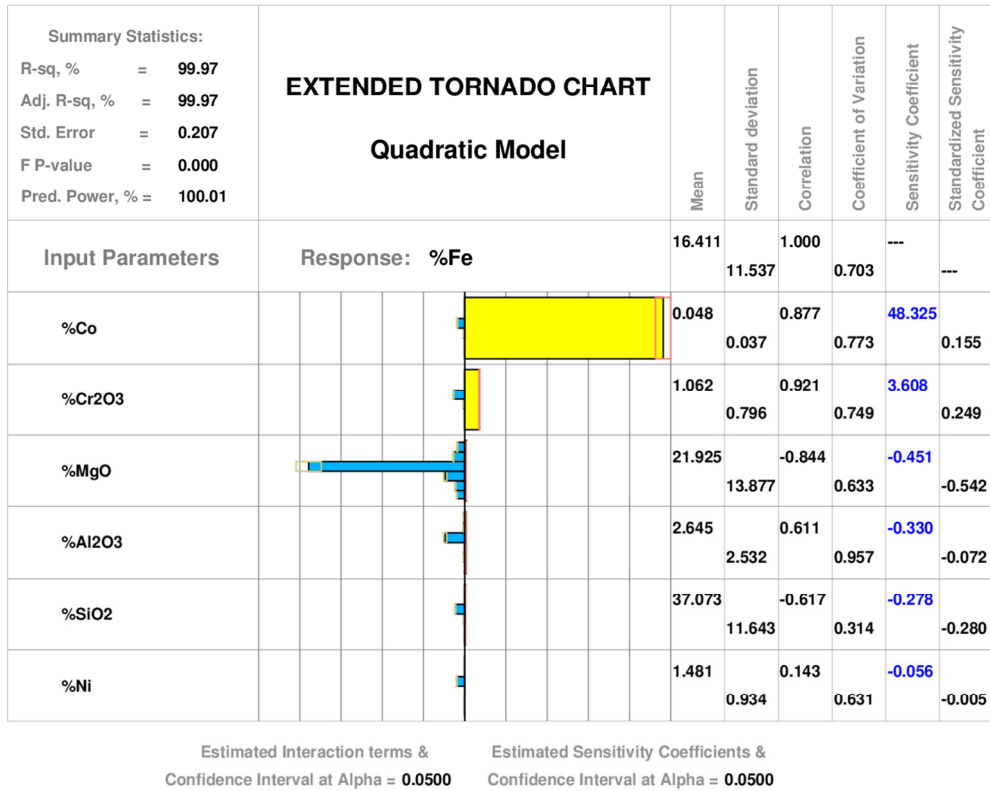
Figure 7: Extended tornado charts for quadratic regression model of *Fe*. Bars on upper plot are based on values of sensitivity coefficients, bars on lower plot are based on values of standardized sensitivity coefficients
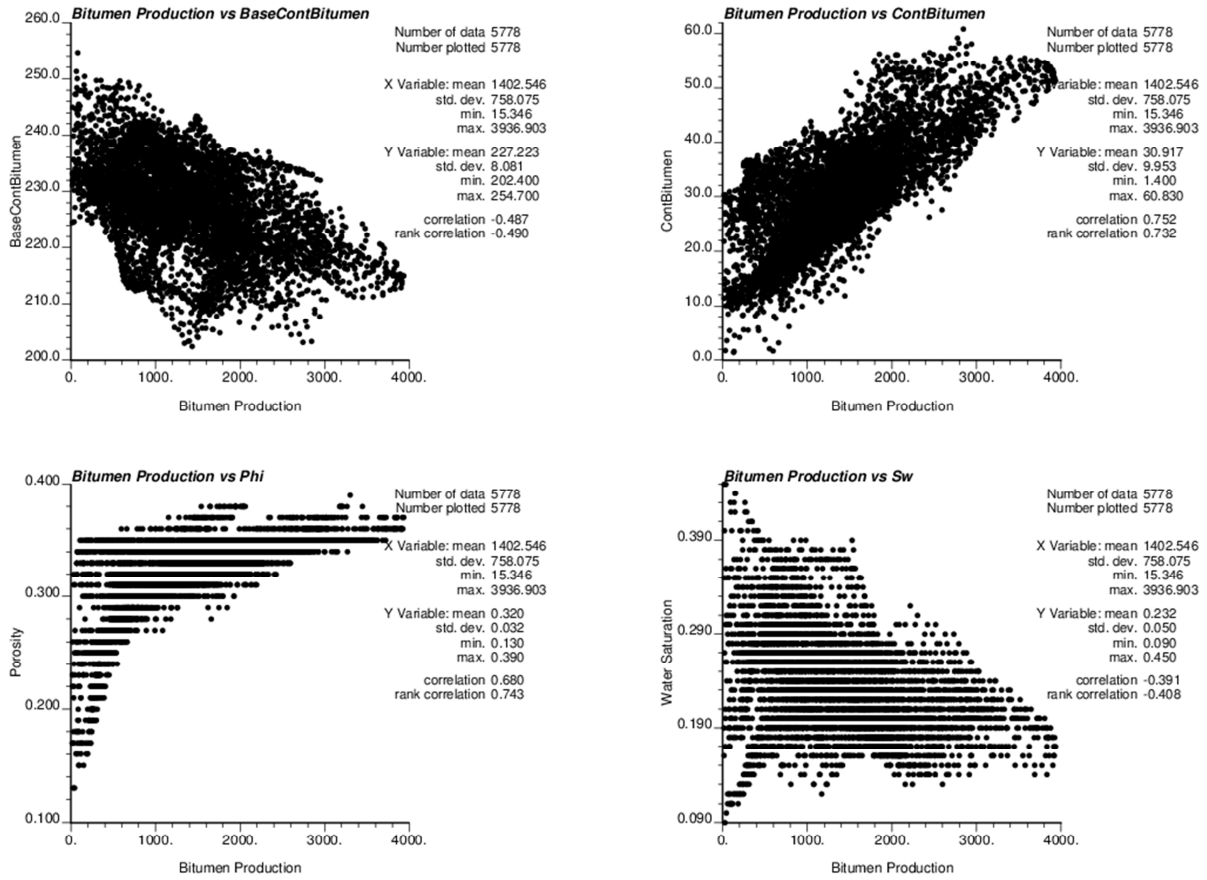
Figure 8: Scatter plots between model response (bitumen production) and input variables (base continuous bitumen, continuous bitumen, porosity, and water saturation)
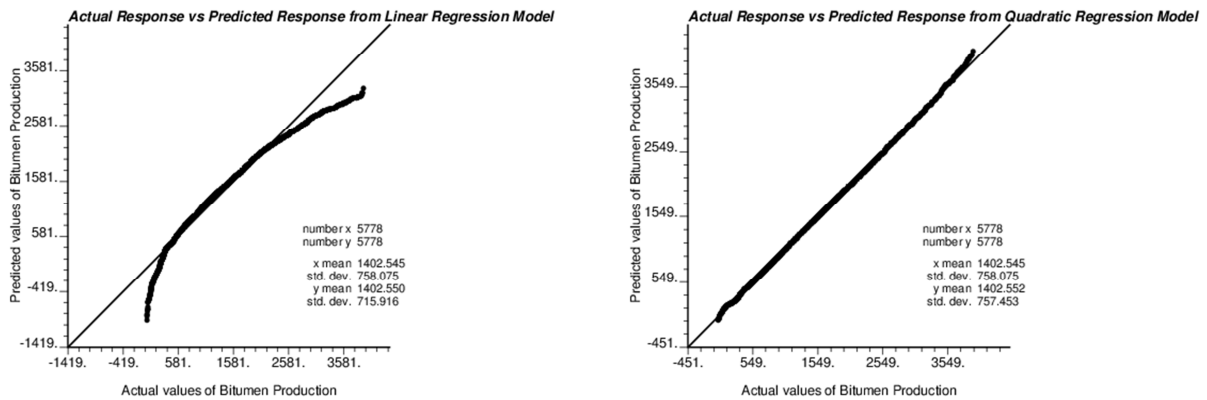
Figure 9: Q-Q plots for actual and predicted values of model response *bitumen production* from linear (left) and quadratic (right) regression models

| Summary Statistics:		EXTENDED TORNADO CHART	Mean	Standard deviation	Correlation	Coefficient of Variation	Sensitivity Coefficient	Standardized Sensitivity Coefficient
R-sq, %	= 89.19							
Adj. R-sq, %	= 89.18	Linear Model						
Std. Error	= 249.390							
F P-value	= 0.000							
Pred. Power, % =	94.44							
Input Parameters		Response: Bitumen Production	1402.545		1.000		---	
				758.141		0.541		---
Porosity			0.320		0.680		12227.615	
				0.032		0.101		0.521
Water Saturation			0.232		-0.391		-2757.184	
				0.050		0.217		-0.184
ContBitumen			30.917		0.752		43.973	
				9.954		0.322		0.577
BaseContBitumen			227.223		-0.487		-6.153	
				8.081		0.036		-0.066

Estimated Sensitivity Coefficients &
Confidence Interval at Alpha = **0.0500**

Summary Statistics:		EXTENDED TORNADO CHART	Mean	Standard deviation	Correlation	Coefficient of Variation	Sensitivity Coefficient	Standardized Sensitivity Coefficient
R-sq, %	= 89.19							
Adj. R-sq, %	= 89.18	Linear Model						
Std. Error	= 249.390							
F P-value	= 0.000							
Pred. Power, % =	94.44							
Input Parameters		Response: Bitumen Production	1402.545		1.000		---	
				758.141		0.541		---
ContBitumen			30.917		0.752		43.973	
				9.954		0.322		0.577
Porosity			0.320		0.680		12227.615	
				0.032		0.101		0.521
Water Saturation			0.232		-0.391		-2757.184	
				0.050		0.217		-0.184
BaseContBitumen			227.223		-0.487		-6.153	
				8.081		0.036		-0.066

Estimated Std.Sensitivity Coefficients &
Confidence Interval at Alpha = **0.0500**

Figure 10: Extended tornado charts for linear regression model of *bitumen production*. Bars on upper plot are based on values of sensitivity coefficients, bars on lower plot are based on values of standardized sensitivity coefficients

Summary Statistics:

R-sq, % = 99.83
Adj. R-sq, % = 99.83
Std. Error = 30.847
F P-value = 0.000
Pred. Power, % = 99.92

EXTENDED TORNADO CHART

Quadratic Model

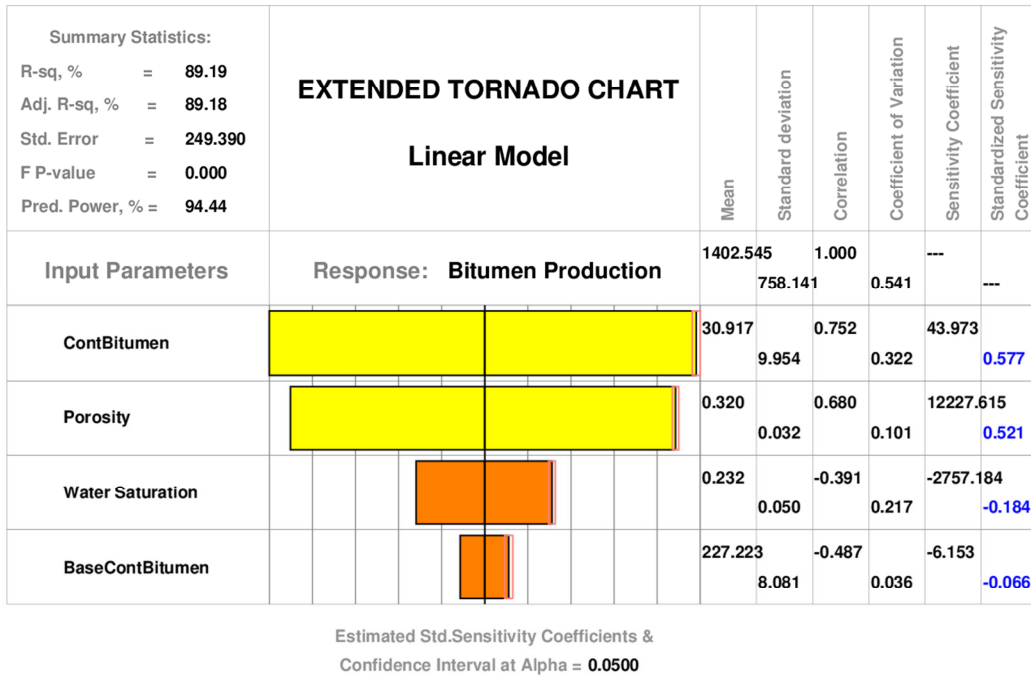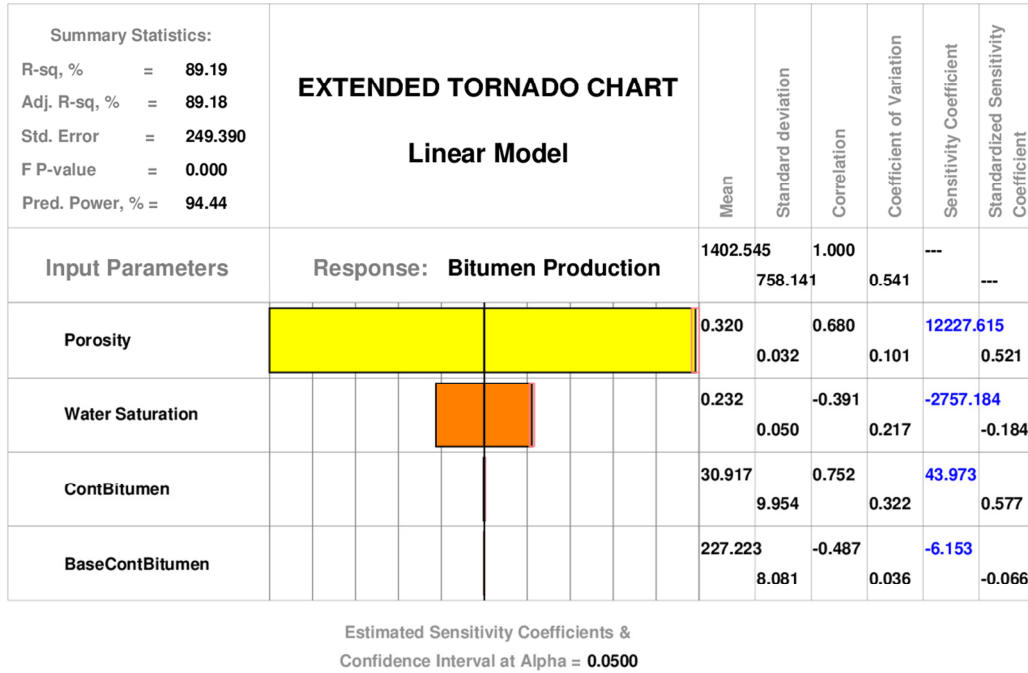| Input Parameters	Response: Bitumen Production	Mean	Standard deviation	Correlation	Coefficient of Variation	Sensitivity Coefficient	Standardized Sensitivity Coefficient
		1402.545 / 758.141		1.000 / 0.541		--- / ---	
Porosity		0.320	0.032	0.680	0.101	13059.500	0.556
Water Saturation		0.232	0.050	-0.391	0.217	-2163.328	-0.144
ContBitumen		30.917	9.954	0.752	0.322	47.400	0.622
BaseContBitumen		227.223	8.081	-0.487	0.036	-0.312	-0.003

Estimated Interaction terms &
Confidence Interval at Alpha = **0.0500**

Estimated Sensitivity Coefficients &
Confidence Interval at Alpha = **0.0500**

Summary Statistics:

R-sq, % = 99.83
Adj. R-sq, % = 99.83
Std. Error = 30.847
F P-value = 0.000
Pred. Power, % = 99.92

EXTENDED TORNADO CHART

Quadratic Model

Input Parameters	Response: Bitumen Production	Mean	Standard deviation	Correlation	Coefficient of Variation	Sensitivity Coefficient	Standardized Sensitivity Coefficient
		1402.545 / 758.141		1.000 / 0.541		--- / ---	
ContBitumen		30.917	9.954	0.752	0.322	47.400	0.622
Porosity		0.320	0.032	0.680	0.101	13059.500	0.556
Water Saturation		0.232	0.050	-0.391	0.217	-2163.328	-0.144
BaseContBitumen		227.223	8.081	-0.487	0.036	-0.312	-0.003

Estimated Std.Interaction terms &
Confidence Interval at Alpha = **0.0500**

Estimated Std.Sensitivity Coefficients &
Confidence Interval at Alpha = **0.0500**

Figure 11: Extended tornado charts for quadratic regression model of *bitumen production*. Bars on upper plot are based on values of sensitivity coefficients, bars on lower plot are based on values 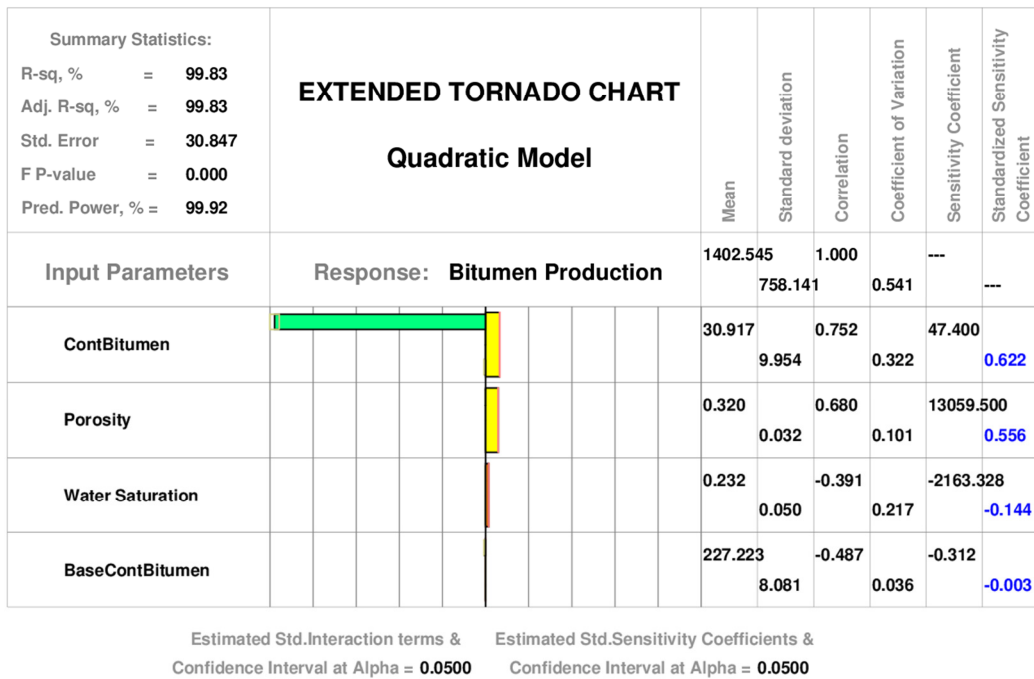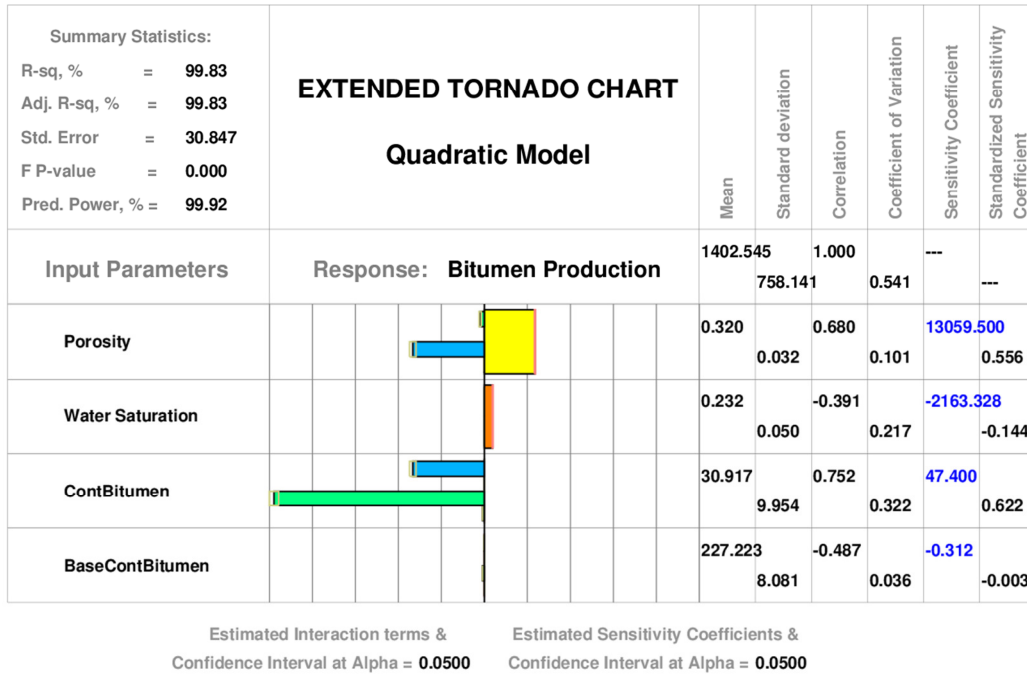of standardized sensitivity coefficients