Machine Learning for Sensitivity Analysis of Probabilistic Environmental Models

29 May 2011

Prepared by

Neptune and Company, Inc.

This page is intentionally blank, aside from this statement.

CONTENTS

FIGURES

TABLES

1.0 Introduction

Complex models are useful for investigating dynamics of systems where multiple variables are interacting in a nonlinear manner. Increasingly these investigations are being conducted using probabilistic simulation approaches such that the uncertainty in the understanding of the system can be propagated through to the predicted response. Quantitatively assessing the importance of inputs becomes important when uncertainty in the response is deemed to be unacceptable for the decision at hand. Sensitivity analysis (SA) can be used to help identify those inputs for which uncertainty reduction through further information collection will have the most impact on reducing uncertainty in the model response. However, sensitivity analysis of high dimensional probabilistic models can be computationally challenging. These challenges can be met through machine learning methods applied to probabilistic simulation results.

Quantitative assessment of the importance of inputs is necessary when the level of uncertainty in the system response exceeds the acceptable threshold specified in the decision making framework. One of the goals of sensitivity analysis is to identify which variables have distributions that exert the greatest influence on the response.

2.0 Sensitivity Analysis Approaches

Sensitivity analysis deals with estimating influence measures for input variables for a given model. In general, this estimate can span the *qualitative* to *quantitative* spectrum, as well as the *local* to *global* spectrum. A *qualitative* SA attempts to provide a relative ranking of the importance of input factors without incurring the computational cost of *quantitatively* estimating the percentage of the output variation accounted for by each input factor. A *local* SA involves varying one input factor while holding all other input factors constant and assessing the impact on the model output. This is *local* in the sense that only a minimal portion of the full volume of the input factor space is explored (*i.e.*, the point at which the input factors are held constant). Although local sensitivity analysis is useful in some applications, the region of possible realizations for the model of interest is left largely unexplored. Global sensitivity analysis attempts to explore the possible realizations of the model more completely. The space of possible realizations for the model can be explored through the use of search curves or evaluation of multi-dimensional integrals using Monte Carlo methods. However, global sensitivity becomes more difficult as the dimensionality of the model increases.

An example of *quantitative local* SA approach is differential analysis based on the partial derivatives of the model with respect to each input factor. Given a model of the form $y = f(X)$, the *local* relative sensitivity measure, S_i, of each input factor, x_i, on model output y can be calculated as:

$$S_i = \left\{ E_x \left[\frac{\partial f(X)}{\partial x_i} \right]^2 \mathrm{var}_x[x_i] \right\}^{1/2} \quad \textbf{(1)}$$

Quantitative global SA attempts to explore the full volume defined by the input factors and then averages over the variation of all input factors to provide an estimate of sensitivity:

$$S_i = \frac{\mathrm{var}_{x_i}[E(y \mid x_i)]}{\mathrm{var}(y)} \tag{2}$$

The analysis is successful if $\sum_{i=1}^{p} S_{x_i} \cong 1$, where p is the number of model parameters.

2.1 Regression Based Methods

Quantitative SA approaches include squared standardized regression coefficients (SSRC) and squared standardized rank regression coefficients (SSRRC).

Given a linear regression model of the form

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i \tag{3}$$

the variance of the model output can be estimated as

$$\mathrm{var}(y) = \sum_{i=1}^{p} \beta_i^2 \,\mathrm{var}(x_i) \tag{4}$$

assuming the input factors are independent. If the model output and input factors are standardized to a mean of 0 and a variance of 1 then the square of the regression coefficients (β_i^2) provides an estimate of S_i. Alternatively, regressing the input factors on the *ranks* of the model output can help to fit nonlinearities in the model.

The coefficient of determination, R^2, provides a measure of how far along the continuum towards being both *quantitative* and *global* these SA methods achieve for a given application. The closer the R^2 is to one the closer the results are to being both *global* and *quantitative*. 1-R^2 represents the percentage of output variation not accounted for by the SA method. As this percentage increases, confidence in the influence estimates is reduced although the resulting relative ranking may still be of value. However, at some unknown percentage, the validity of the relative rankings also come into question as the potential increases that account for this output variation would change the relative rankings. Standardized regression and rank correlation methods assume a monotonic linear relationship between the input factors and the model output. A low R^2 may be reflective of a model structure that does not meet this assumption.

The essence of these approaches is an analysis of variance (ANOVA) decomposition, decomposing the response variance into partial variances of increasing dimensionality. The total number of terms involved in this type of decomposition is $2^p - 1$. The dimensionality of an ANOVA decomposition analysis becomes prohibitive at even moderate values of p. "the larger the number of factors, the higher the likelihood of non-negligible higher-order terms".

2.2 Fourier Amplitude Sensitivity Test (FAST)

Several approaches have been proposed to handle nonlinear, nonmonotonic models. Two of these approaches include the Fourier Amplitude Sensitivity Test (FAST) (Saltelli *et al*. 1999) and Sobol's design of experiment (SDOE) approach (Sobol 1993). These methods provide an estimate of the proportion of the variation in the model output due to an input factor through

an ANOVA-like decomposition of the output variation. The two methods use a different computational strategy for decomposing the partial variances of increasing dimensionality (main-effects, two-way interactions, three-way interactions, etc.). In the context of SA, this ANOVA decomposition can be described in terms of total sensitivity indices for each input factor, S_{Ti}. An S_{Ti} for input factor i is calculated as the sum across all main and interaction sensitivities that involve the i^{th} input factor:

$$S_{T_i} = S_i + \sum_{j \neq i}^{n} S_{ij} + \sum_{j,k \neq i}^{n} S_{ijk} + \cdots \quad (5)$$

where S_i is the first-order or (main effect) sensitivity index and S_{ij} is the second-order (interaction effect) sensitivity index and so on. The total number of sensitivity indices is $2^n - 1$, where n is the number of input factors. Because SDOE requires multi-dimensional integration to estimate the sensitivity indices, this method can be prohibitive computationally for moderately complex models (complexity is defined by the number of input factors, n).

FAST is a computationally elegant alternative to SDOE for side stepping this "curse of dimensionality". FAST involves simulating input factors using a *random phase-shift* sampling scheme:

$$x(s)_i = \frac{1}{2} + \frac{1}{\pi}\arcsin(\sin(\omega_i s + \zeta_i)) \quad (6)$$

where s varies from $(-\pi, \pi)$, the ω_i's are a linearly independent set of frequencies (a unique frequency for each input factor), and ζ_i is a random phase shift chosen uniformly in $[0,2\pi)$. The model output based on this sampling scheme for the input factors shows different periodicities at each of the ω_i. The amplitude of the oscillation at the ω_i's and their harmonics provides a measure of the model output sensitivity to the corresponding input x_i. Fast Fourier Transform of the resulting model output provides the pieces from which the S_T's can be computed.

This is accomplished by a Fourier series expansion of $y = f(x(s)_i)$

$$y = \sum_{-\infty}^{+\infty} \{A_j \cos js + B_j \sin js\} \quad (7)$$

where A_j and B_j are the Fourier coefficients and can be estimated via a fast Fourier transform algorithm.

The spectrum of the Fourier transform is

$$\Lambda_j = A_j^2 + B_j^2 \quad (8)$$

Summing all Λ_j provides an estimate of the total variance in y

$$\hat{D} = \sum_{j \in Z} \Lambda_j \quad (9)$$

Summing all Λ_j excluding the frequency embedded in x_i and its associated higher harmonics, Z^0, provides an estimate of the variance due to the uncertainty in x_i

$$\hat{D}_i = \sum_{j \in Z^0} \Lambda_j \quad (10)$$

The sensitivity of y to x_i is then given by

$$\hat{S}_i = \hat{D}_i / \hat{D} \tag{11}$$

Unfortunately, global sensitivity methods such as the FAST require construction of model simulations in which a signal is embedded in each input parameter and then the strength of the signal in the model realizations is a measure of parameter sensitivity. This requires construction of a separate model with distributions for input parameters constructed specifically for sensitivity analysis, rather than for uncertainty analysis. The space of possible realizations for the model can be explored through the use of search curves or evaluation of multi-dimensional integrals using Monte Carlo methods. However, these approaches to global sensitivity analysis become more computationally intensive as the dimensionality of the model (i.e., the number of model parameters) increases and can be prohitbitive for models that include hundreds or thousands of stochastic parameters.

Results from a typical probabilistic model run design for uncertainty analysis can be difficult to use in FAST. A probabilistic model run could be designed for both uncertainty analysis and FAST if the cumulative distribution functions (cdf) for all input factors are available analytically. Analytic cdf's are available for distributions such as the uniform and Weibull but not for the normal. Thus FAST may not be feasible when model run times are long and when uncertainty and sensitivity analysis are both part of the decision process.

2.3 Machine Learning Approaches

Because of the computational cost, sensitivity analysis of high-dimensional probabilistic models requires efficient algorithms for practical application. Machine learning provides tools that allow for the partitioning of the variance in the model response to the input parameters by exploration of the realizations from a model run for uncertainty analysis. Two common machine learning approaches that could be brought to bear for sensitivity analysis are bagging (Breiman 2001) and boosting (Friedman 2001) of regression trees. The advantages of machine learning approaches include the ability to fit non-monotonic and non-linear effects, the ability to fit parameter interaction effects, and the ability to visualize these effects and their interaction across the range of the response and input parameters. Bagging, boosting and other machine learning approaches typically produce similar results for noisy data. In the case of realizations from a probabilistic process model, each realization is a deterministic evaluation of the model and all the stochastic predictor variables are available. As such there is no unexplainable variation in the process model response (as is the case with observed data) and the choice of machine learning algorithm should have negligible impact of the results of the sensitivity analysis.

2.3.1 Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Splines (MARS) is a recursive partitioning approach (similar to Classification and Regression Trees) that helps deal with the ANOVA decomposition "curse of dimensionality", making estimation of sensitivity indices computationally achievable for large n (Friedman, 1991). MARS accomplishes this by optimally partitioning or, splitting the model output and input factors into subsets, from which splines are fit. The recursive nature of the algorithm results in increasingly local splits of the model results in which all significant interaction effects in subregions are found. MARS is

able to find and fit only the significant nonlinear and thresholds relationships between the model input and output. An input factor's influence is calculated as the sum of the partial residuals removing all main and interaction effects that variable enters.

$$S_{x_i} = \sum \left[f - a_o + \sum_{i \neq 1} f(x_i) + \sum_{i,j \neq 1} f(x_i, x_j) + \sum_{i,j,k \neq 1} f(x_i, x_j, x_k) + \cdots \right]^2 \qquad (12)$$

2.3.2 Gradient Boosting Machines (GBM)

Boosting of decision trees provides a technique that adds to the flexiblity offered by recusive partioning methods such as MARS. The Gradient Boosting Machines (GBM) approach utilizes boosting of binary recursive partitioning algorithms that deconstruct a response into the relative influence from a given set of explanatory variables (stochastic model input parameters). That is, the collection of results (the process model response) is broken up into parts, and each part is examined separately. This process is repeated with smaller and smaller parts, each analyzed for the relationship between the model inputs (explanatory variables) and the results. This sensitivity analysis methodology identifies which stochastic model input parameters are most influential in determining the results, such as media concentrations or future potential doses. It also identifies the ranges over which the influence is strongest.

Variance decomposition of the GBM fit is then used to estimate SIs. Under this decomposition approach, the goal is to identify the most influential explanatory variables that are identified within a model. The necessary degree of model complexity is assessed using validation metrics, based on comparison of model predictions, with randomly selected subsets of the data. This approach uses the "deviance" of the model as a measure of goodness of fit. The concept of deviance is fundamental to classical statistical hypothesis tests (e.g., the common *t*-test can be derived using a deviance-based framework) and guides the model selection process applied here.

The GBM fitting approach is based on finding the values of each explanatory variable that result in the greatest difference in mean for the corresponding subsets of the response. For example, if there were only a single explanatory variable, the GBM would identify the value of the explanatory variable that corresponds to a split of the response into two parts. This will ensure that no other split would result in corresponding groups of the response variable with a greater difference in means. When multiple explanatory variables are present, these multiple splits are referred to as "trees." Each tree results in an estimate (e.g., prediction) of the response. As multiple potential trees are evaluated, they are compared to the observed data using a loss function. The selection of the loss function is an influential aspect of the GBM process, and depends on the distribution of the response variable. For data that are sufficiently skewed (e.g., non-normal), the absolute error loss function typically produces more reliable results.

There is a trade-off that exists when considering which loss function to use. The squared-error loss function results in better fitting models, but can do so at the expense of introducing spurious variables into the model selection process when the response distribution is sufficiently skewed. The absolute error loss function produces model predictions with more variability, but is less likely to result in the selection of spurious variables in the model. For this application, the focus has been on using a deviance-based method to obtain models that

identify the most important explanatory variables with respect to the observed variability in the response. Therefore, the squared-error function was used in these applications.

With standard linear regression techniques, it is assumed that the relationship between the response and the explanatory variable is a constant (e.g., the parameter estimates in the linear model). With the GBM approach, this relationship is not constrained by assumptions of linearity, and the partial dependence plots show the data-based estimate of the relationship between the response and the explanatory variable. This is useful for understanding the influence of changes in a single explanatory variable, when integrating across all other explanatory variables.

2.4 Example

2.4.1 "Sobol g-function"

The Sobol *g*-function (Saltelli *et. al.* 1999) provides an analytic non-monotonic test function for evaluating the performance of various sensitivity analysis approaches. This function is defined as:

$$f = \prod_{i=1}^{p} g_i(x_i) \tag{13}$$

where p is the total number of input factors and $g_i(x_i)$ is given by

$$g_i(x_i) = \frac{|4x_i - 2| + a_i}{1 + a_i}, \tag{14}$$

with

$$x_i = \frac{1}{2} + \frac{1}{\pi}\arcsin(\sin(\omega_i s + \phi_i)), \tag{15}$$

and s varing along $(-\pi,\pi)$, $\varphi_i \sim U[0,2\pi)$, and ω_i are specified frequencies.

The Sobol *g* function was simulated for $p = 8$ and frequencies $\{\omega_i\} = \{23, 55, 77, 97, 107, 113, 121, 125\}$ for a specific set of a_i's. Table 1 provides a comparison of sensitivity indices calculated analytically (S) and using of GBM, MARS, FAST, differential analysis (DERIV), squared standardized regression coefficients (SSRC), and squared standardized rank regression coefficients (SSRRC).

Note that the GBM, MARS and FAST methods return sensitivity indices that are close to the actual sensitivities for the Sobol function (S). The Sobol function is highly non-liniear, hence the standardized regression approaches do not work very well. As described, FAST is computationally challenging. The difference between MARS and GBM is close, but preference is given overall to the GBM approach.

A goodness-of-fit statistic is also presented in the bottom row of Table 1. This is calculated as the standard chi-square good ness –of-fit statistic – the sum of the square of the observed (SA method) minus the expected (S value) all divided by the expected value, in which case a small value implies a better fit. These goodness-of-fit statistics show that the GBM method outperforms the other methods, although the difference is small for GBM and FAST.

	a	*S*	*GBM*	*MARS*	*FAST*	*DERIV*	*SSRC*	*SSRRC*
x_1	99	0.0001	0.0003	0.0000	0.0043	0.0037	0.6880	0.7805
x_2	0	0.4227	0.4146	0.4397	0.4287	0.3151	0.0137	0.0036
x_3	9	0.0058	0.0011	0.0084	0.0190	0.0401	0.0003	0.0000
x_4	0	0.4227	0.4200	0.4239	0.4269	0.3169	0.0163	0.0098
x_5	99	0.0001	0.0001	0.0000	0.0006	0.0037	0.0350	0.1152
x_6	4.5	0.0182	0.0335	0.0239	0.0141	0.0787	0.0012	0.0554
x_7	1	0.1304	0.1303	0.1041	0.1063	0.2382	0.0574	0.0344
x_8	99	0.0001	0.0000	0.0000	0.0002	0.0037	0.1881	0.0012
Goodness-of-Fit statistic			3.3	14.8	3.6	470	7,250	535

Table 1. ***Sensitivity Indices by Sensitivity Analysis Method for Sobol g-function application with p = 8.***

GBM is run on the realizations themselves, whereas FAST requires set up interms of an embedded signal. This makes FAST cumbersome to deal with comparatively. Also, GBM outperforms MARS, which is not as flexible and runs slower. GBM tends to provide the best fit, is flexible and is applied directly to the realizations. Consequently, it is the preferred method, and the one that is used for the sensitivity analyses for the Clive DU PA.

2.4.2 Visualization

Once a GBM has been is constructed, each of the explanatory variables that exist in the model can be assigned an SI. The SI is obtained through variance decomposition and can be interpreted as the percentage of variability explained in the model by a given explanatory variable. The sum of the SI's across the entire set of explanatory variables in the machine will approximately equal the R^2 of the linear regression of the process model predictions versus the machine learning predictions. The R^2 values for this version of the model indicate the high degree of predictive power of the machine learning in fitting the process model predictions.

For a GBM model, the partial dependence is determined through the integration across the joint density to obtain a marginal distribution. The integration is performed using a "weighted tree traversal" measure that is analogous to more common integration procedures performed with Riemann or Lebesgue measures. The vertical axis of the partial dependence plot shows the change in the response variable as a function of the changes in the explanatory variable.

In order to assess the relationship between an individual explanatory variable and the response of interest, partial dependence plots are used (**Figure 1**). The first panel depicts a density estimate of the simulated response from the process model as well as the machine goodness-

of-fit and summary statistics for the response. The percentiles of the response distribution in this panel are shaded to provide a context for the partial dependence plots presented in the remaining panels. The colors indicate the percentile range of the response as follows:

1. The 0th - 25th percentile region is shaded orange-brown
2. The 25th - 50th percentile region is shaded dark yellow-green
3. The 50th - 75th percentile region is shaded light green
4. The 75th - 100th percentile region is shaded light blue

The y-axis scale of the partial dependence plots is in units of the response distribution (the x axis of the first panel). Given that each parameter has a different range and strength of influence on the response, the y axes of the partial dependence panels depict only the range of the response over which a particular parameter is influential. If the original scale of the response were maintained on each partial dependence panel, then the influence of the least influential parameter would not be visible in many cases. To counteract this scale issue, the background of the partial dependence panels is shaded to depict the percentile of the response over which the parameter is influential. For example, if the background of the partial dependence plot under the partial dependence line is light blue, then that indicates the parameter's influence on the upper end of the response distribution (i.e., the 75th to 100th percentile of the response).

The partial dependence panels in each figure show the distributions of the explanatory variables (black line), and the partial dependence curve (blue line) shows changes in the response as a function of each explanatory variable.

The plots show that the distributions for the three input parameters are uniform, and that the effects show sensitivity across the entire range of the inputs. The effects are first negative, and then positive, which is to be expected given Equation 15. Also note that the linear regression methods would not be able to track the non-linearity, and instead fits a straight, horizontal line for these parameters, which shows them to be non-sensitive. This is a prime example of why methods such as GBM are advantageous.

Note:

An implementation of Friedman's gradient boosting machine approach is available in the R statistical software in the gbm package. The gbm package functions were tailored to generate global sensitivity indices and partial dependence visualization of the impact of model input parameters on the model response based on a set of realizations from the probabilisticly run model.

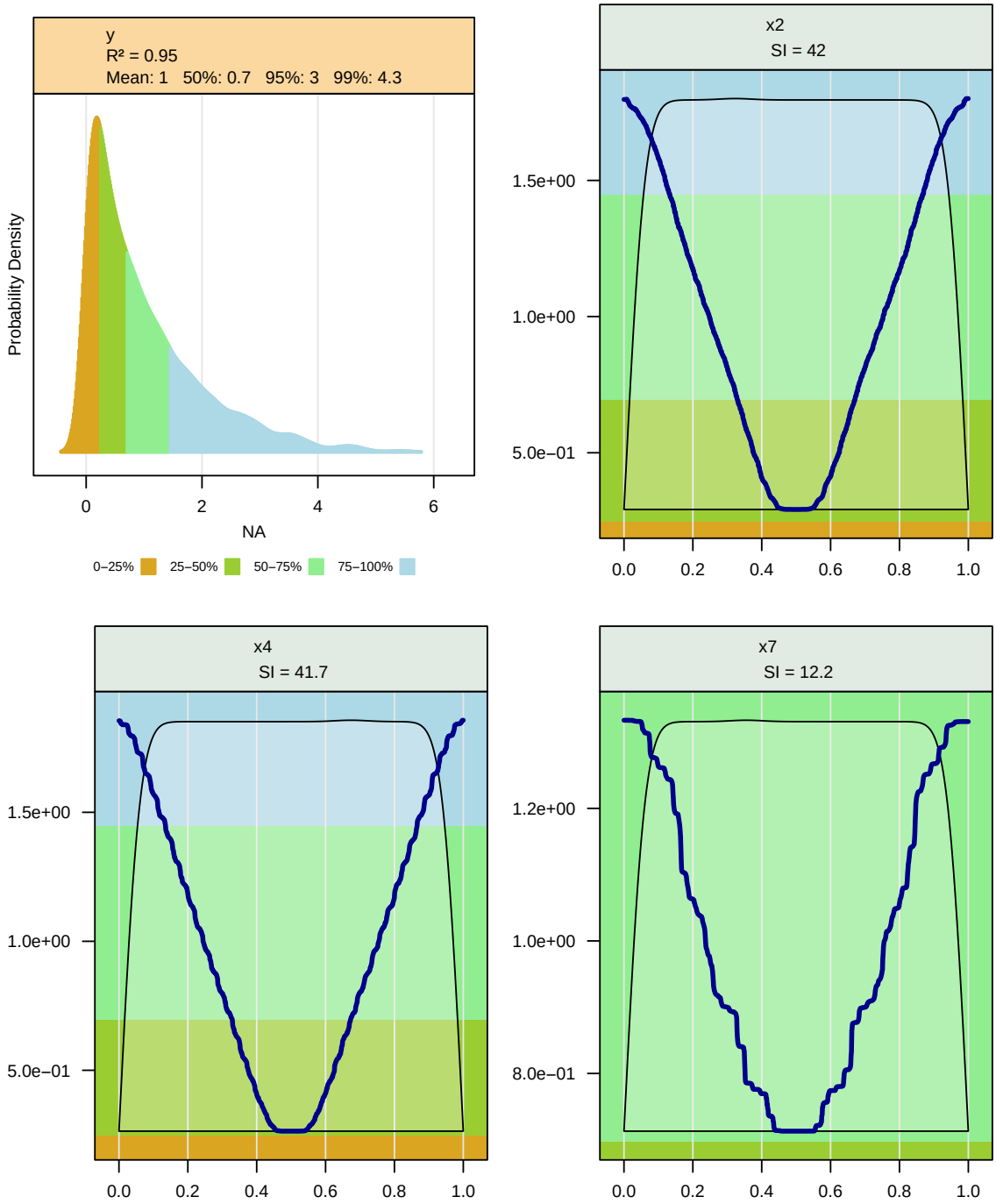

Figure 1. Sensitivity and Partial Dependence Plots for the GBM fit to the Sobol Function.

3.0 References

Friedman, J.H. (1991), "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, 19, 1-141.

Friedman, J.H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics 29(5):1189-1232.

Saltelli A., Tarantola S., and Chan K.P.-S. (1999), "A Quantitative Model-Independent Method for Global Sensitivity Analysis of Model Output," *Technometrics*, 41, 39-55.

Sobol, I.M. (1993), "Sensitivity Analysis for Nonlinear Mathematical Models," *Mathematical Modeling & Computational Experiment*, 1, 407-414.