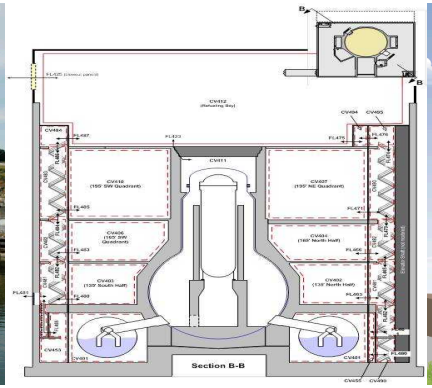
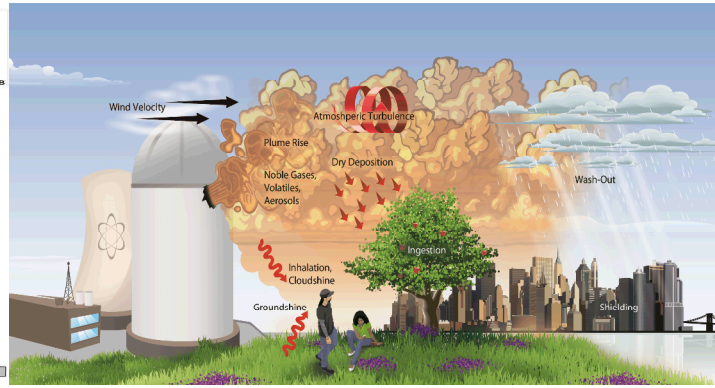
SAND #2013-3868 P

SAND2014-3253P

Exceptional service in the national interest

Sandia National Laboratories





SOARCA project uncertainty Analysis

Uncertainty and Sensitivity Analyses techniques

Presented by C. Sallaberry (6224)

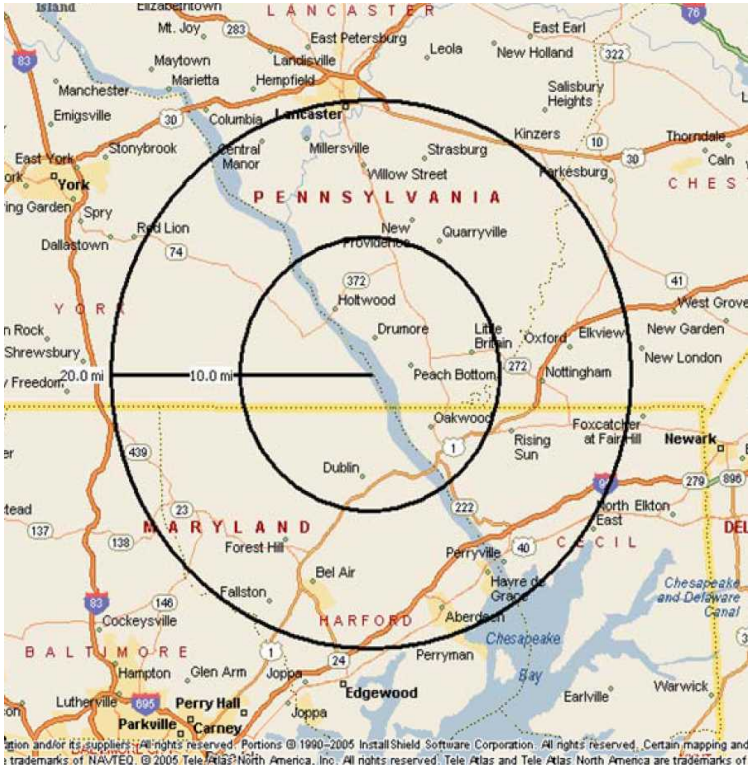
U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND2012-10584P

Overview of second part

- Uncertainty Analysis purpose
- Uncertainty Analysis Techniques
- Sensitivity Analysis definition and purpose
- Sensitivity Analysis Techniques
- Conclusions



Peach Bottom 10 and 20 mile analysis areas

Purpose of uncertainty analysis

- Study of the uncertainty in analysis results that derives from the collective uncertainty in analysis inputs.
- Primary source of information for the decision maker, as it answers the following questions:
  - What is the best strategy/choice ?
  - How confident am I in the choice I make based on the results?
  - What are the quantitative arguments for and against this choice?
- Such analysis is almost always recommended and often required in any complex analysis

> Preparation of fully documented written risk assessments that explicitly define the judgments made and attendant uncertainties clarifies the agency decision-making process and aids the review process considerably

Risk Assessment In The Federal Government: Managing The Process. National Academy Press, Washington, DC, 1983.

Uncertainty analysis setup

Sandia National Laboratories

- Before starting the analysis it is important to define the framework of the analysis.
  - What is the chosen risk metric?
  - Is randomness (aleatory uncertainty) considered ? Is lack of knowledge (epistemic uncertainty) considered ?
  - Is the regulation defined for a physical result or a statistic on this result (expected value, quantile value) ?
  - Are the results time-dependent ? Is the regulation defined for a certain time? Up to a certain time ? Should the maximum over time be considered or another value?
- The definition of the metric of interest is crucial as it affects the whole analysis structure and the way results are presented.

Classical statistics used in UA

Sandia National Laboratories

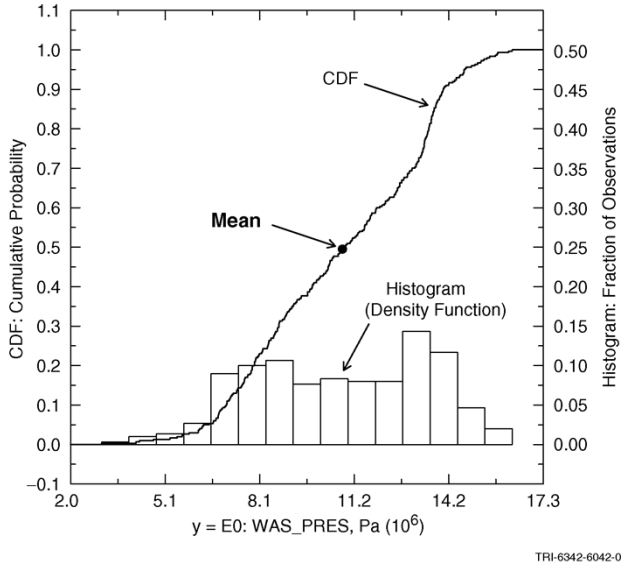
- Mean (over aleatory uncertainty, epistemic uncertainty or both) is always presented as a central tendency. It is often a good summary that included both the consequence of an event and the likelihood
- While the mean is usually a required value it is **Not sufficient** as it groups all the information in a single number. Risk analysis is not only probability x consequence. To have a full picture of the risk, **the decision maker needs to be presented with probability AND consequence**

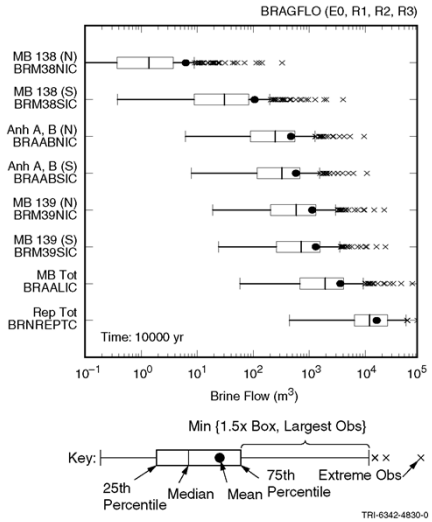
- Calculation or higher order moments (standard deviation, skewness, kurtosis…) gives information on the shape of the output distribution

- Calculation of quantiles gives a better estimate of risk and help making risk informed decisions.

Uncertainty analysis techniques and graphical representations

- Classical representation include CDF or CCDF, histogram of density function (PDF), with inclusion of the most relevant statistics (mean, median, significant quantiles …)
- When distributions need to be compared, a more compact and useful representation can be used, such as boxplots.





Importance of defining accurately the purpose of the analysis

Sandia National Laboratories

- Regulatory requirements need to be read carefully and understood.
- Often times, the words uncertainty, randomness, probability and equivalent will be used ambiguously in the requirements.
- Such condition may lead the people responsible for the analysis to interpret the wording and such interpretation may result in inappropriate uncertainty representation.
- It is therefore really important to clearly define the high level characterization in a robust mathematical framework including the role of uncertainty within it.
- ***Even if the regulatory requirement is a single number*, the analysis that support this number has to be unambiguous and defensible.**

Conclusions on Uncertainty Analysis

Sandia National Laboratories

- Uncertainty analysis is often seen as the conclusion of the complex analysis as it reflects the information presented to the decision maker

- However, even if the regulatory requirement are based on a single number, this number has to be defended in a well defined and unambiguous fashion and supported by a more complete demonstration.

- It is more appropriate to educate the decision maker and the public into understanding the analysis than simply giving a number.

- One has to give people what they need to know even if it is a lot more than what they asked for.

Sensitivity analysis definition

Sandia National Laboratories

- Along with uncertainty analysis, sensitivity analysis forms the last step of uncertainty treatment in complex systems
- The type of sensitivity analysis is dependent on the technique used to propagate uncertainty (deterministic, local, global)
- This presentation will focus mostly on the global sensitivity analysis techniques

Purpose of sensitivity analysis (1/2)

Sandia National Laboratories

- Involve generation and exploration of mapping from analysis inputs to analysis results
- Analysis input: $\boldsymbol{x} = [x_1, x_2 \ldots, x_{nX}]$
- Analysis results: $\mathbf{y}(\mathbf{x}) = [y_1(\boldsymbol{x}), y_2(\boldsymbol{x}), \ldots, y_{nY}(\boldsymbol{x})]$

How important are the individual elements of ***x*** with respect to the uncertainty in ***y***(***x***)?

- Most of the time the sensitivity analysis is performed on **epistemic uncertainty** as it is the one that can be reduced.

Purpose of sensitivity analysis (2/2)

Sandia National Laboratories

Sensitivity Analysis can be used for

- Ranking parameters in term of their importance relative to the uncertainty in the output
- Verification and validation of the model. It is a powerful tool to check whether the system performs as expected
- Leading further uncertainty quantification towards the parameters that really matters in an iterative process

Sensitivity is usually **not** used for

- Prediction: The purpose is not to construct a meta-model
- Determining the importance of one parameter or one feature of the system to the response. It only looks at the influence of the **uncertainty** in the input on the **uncertainty** of the output

Sensitivity Analysis techniques

- **Deterministic methods** whose basis is to invert the problem. Examples of such methods are adjoint method or gradient-based method. In most of big analysis such methods are impractical due to the complexity of the (often non-linear) equations and potential coupling between the systems.

- **Local methods** can be applied on deterministic or probabilistic problems. Most of them calculate derivative of the function relative to one or multiple parameter via Taylor series expansion. We are usually interested more on the global effect of an input variables than on its effect around a specific point in the hypercube

- **Reliability methods** are a particular case of local method that focus on the likelihood of particular event. They may be interesting when the probability of interest of an event is so low that it would not be practical to use Monte Carlo to capture it. The study of local sensitivity around this particular point of interest has a thus a particular interest

- **Global methods** are the most widely used techniques in Complex System analysis as they use the information obtained for Uncertainty Analysis. Description follows

Global sensitivity analysis

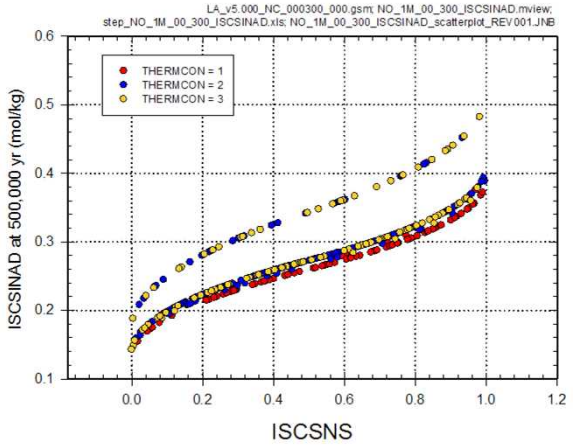
Sandia National Laboratories

- These methods try to capture the influence of the inputs on the whole input hyperspace.
- Working on the overall for both the input and output of interest, they had to consider the whole range of distribution of the variables.
- Several methods cover the whole range of uncertainty in inputs and output of interest. They include response surface, variance based decomposition and sampling-based methods. We will focus for this presentation on tools used with sampling-based methods.
- For sampling-based methods, the same sample set is used to perform uncertainty analysis and sensitivity analysis. While, depending on the problem, the sample size may be large, there is no need to rerun the code a second time.

Traditional sensitivity techniques
1: scatterplots

- Scatterplots consists in plotting the relation between inputs (on the x-axis) and output (on the y-axis)
- They are the simplest way to observe any dependency between input and output without making any assumption. And are usually enough to understand relationship between input and output
- They are usually bidimensional (one output vs. one input) but can be sometime tri-dimensional to show three way interactions.
- But, they may be impractical if hundreds of inputs and outputs are in consideration and they do not **quantify** the importance of the relation between inputs and outputs

Color coding or use of different symbols may help represent conjoint influence from discrete and continuous parameter



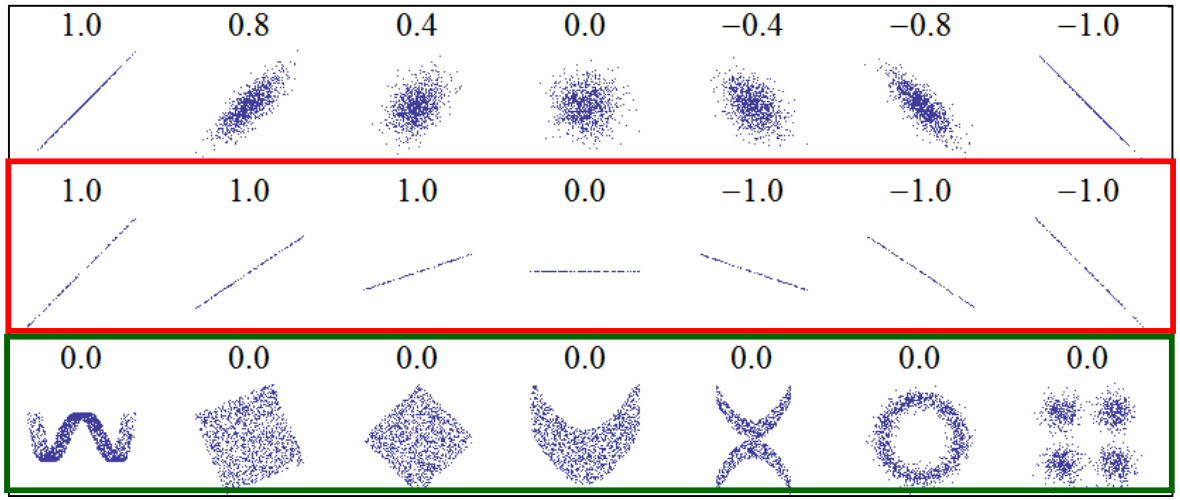
Traditional sensitivity techniques
2: correlation coefficient (CC)

- The (Pearson) Coefficient of Correlation measures the strength and direction of a **linear** relationship between two quantities X_i and Y_j

$$\rho(X_i,Y) = \frac{\mathrm{cov}(X_i,Y)}{\sqrt{\mathrm{var}(X_i).\mathrm{var}(Y)}} \Rightarrow r(X_i,Y) = \frac{n\sum_{i=1}^{n} x_{i,1}.y_i - \sum_{i=1}^{n} x_{i,1}\sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_{i,1}^2 - \left(\sum_{i=1}^{n} x_{i,1}\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

Note 1: as the covariance is normalized by the variance of the two terms in consideration, the slope of the relation does not change the correlation value

Note 2: there may be dependency amongst variables that will not be captured by correlation



Source: http://en.wikipedia.org/wiki/Correlation

Traditional sensitivity techniques
3: partial correlation coefficient (PCC)

Sandia National Laboratories

- Measures the strength and direction of a linear relationship an input X_i and an output Y_j **after** the linear effect of the remaining input parameters has been taken out from **both** X_i and Y_j

Step 1: linear regression models of Y_j *and* X_i

$$\tilde{Y}_{i,j} = a_0 + a_1 X_1 + a_2 X_2 + \cdots a_{i-1} X_{i-1} + a_{i+1} X_{i+1} + \cdots a_n X_n$$

$$\tilde{X}_i = b_0 + b_1 X_1 + b_2 X_2 + \cdots b_{i-1} X_{i-1} + b_{i+1} X_{i+1} + \cdots b_n X_n$$

Step 2: Calculation of residual

$$ry_{i,j} = Y_j - \tilde{Y}_{i,j}$$

$$rx_i = X_i - \tilde{X}_i$$

Step 3: Calculation of correlation between $ry_{i,j}$ and rx_i

$$PCC_{(X_i,Y)} = \frac{\mathrm{cov}\left(rx_i, ry_{i,j}\right)}{\sqrt{\mathrm{var}(rx_i).\mathrm{var}(ry_{i,j})}}$$

Traditional sensitivity techniques
4: Regression Coefficient (RC)

- Represent the coefficients of the linear regression model, estimated using Least Square approach

$$X = (X_1, \cdots, X_n) \rightarrow Y_j \quad \Longrightarrow \quad \tilde{Y}_j = \sum_{i=1}^{n} \theta_i X_i$$

model **Linear regression**

•We want to select the θ_i such that they minimize the square difference between the output Y_j and its linear regression. (That's why it's called Least Square)

•The minimum of the function is obtained when its derivative is zero.

$$\min_{\theta} f(\theta) = \min_{\theta} \left\| Y_j - \tilde{Y}_j \right\|^2 = \min_{\theta} \left\| Y_j - \theta X \right\|^2$$
$$\rightarrow f'(\theta) = 0$$
$$\rightarrow 2X^T (Y_j - \theta X) = 0$$
$$\rightarrow (X^T X)\theta = X^T Y$$
$$\rightarrow \theta = (X^T X)^{-1} X^T Y$$

Traditional sensitivity techniques
5: Standardized Regression Coefficient (SRC)

- Standardized Coefficients of the Linear Regression model corresponds to linear coefficients of the Standardized model. The standardization of a variable is performed by subtracting the mean and dividing the result by the standard deviation

$$X = (X_1, \cdots, X_n) \rightarrow Y_j$$

model

Standardization

$$Y_j^* = \frac{Y_j - \mu_{Yj}}{\sigma_{Yj}}; \quad X_i^* = \frac{X_i - \mu_{Xi}}{\sigma_{Xi}}$$

$$X^* = (X_1^*, \cdots, X_n^*) \rightarrow Y_j^*$$

Standardized model

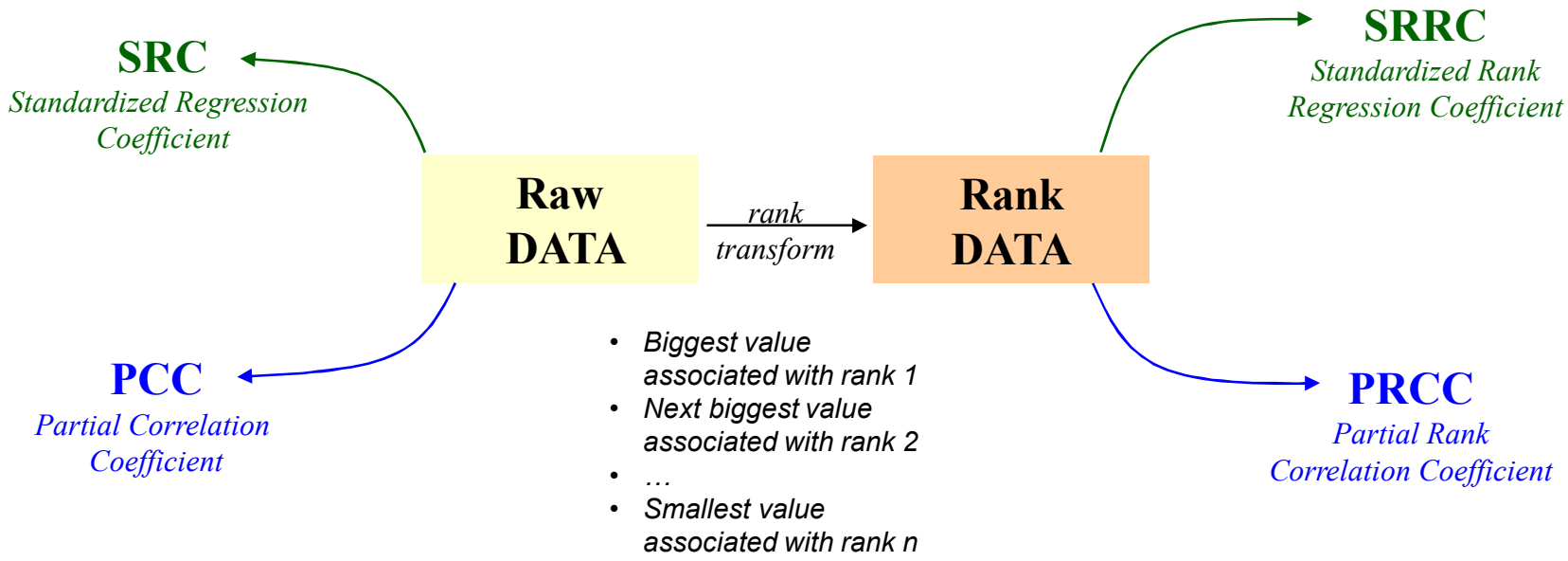
$$\tilde{Y}_j^* = \sum_{i=1}^{n} \theta_i^* X_i^*$$

Linear regression

The standardized values will **not** be affected by unit change

Traditional sensitivity techniques
6: rank transform data

- All the previous techniques suppose that the relation between input and output is linear
- It is generally NOT the case
- A simple way to relax part of this constraint is to work with **RANK**



Traditional sensitivity techniques
7: coefficient of Determination (R^2)

- The coefficient of Determination, noted R2 measures the part of the variance of the output that is explained by the regression model.
- It is a good indicator whether the traditional sensitivity analysis is sufficient or not

$R^2 \sim 1$: Most of the variance of Y is explained

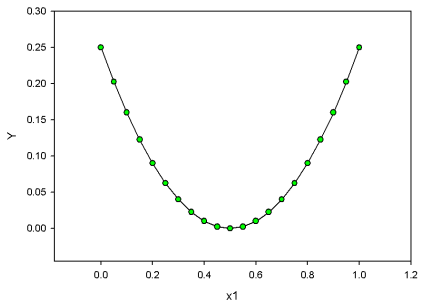
No other analysis is required to determine the most influent parameters

$R^2 \sim 0$: Most of the variance of Y is NOT explained

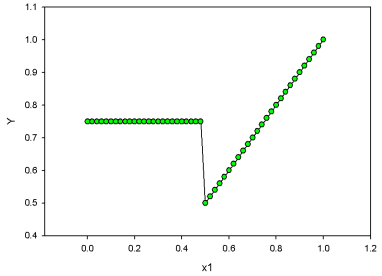
Some influent parameters may be missing OR the influence of the parameters selected is misrepresented. It may be thus necessary to look at scatterplots or/and apply different regression techniques

Traditional sensitivity techniques
8: limit of traditional methods

- Traditional methods will fail to capture this kind of relationship between x and y

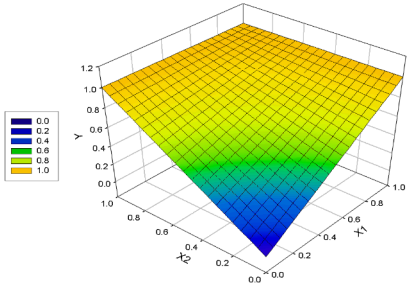


$$y = (x_1 - 0.5)^2$$

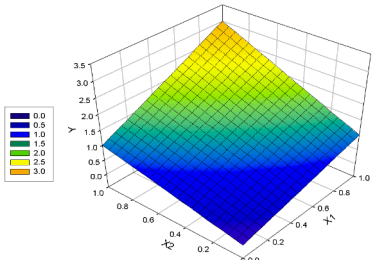


$$y = \begin{cases} 0.75 & \text{if } x < 0.5 \\ x & \text{else} \end{cases}$$

- Conjoint influence will not be capture by traditional additive techniques



$$y = x_1 + x_2 - x_1.x_2$$



$$y = x_1 + x_2 + x_1.x_2$$

Variance based method

1: complete variance decomposition

Sandia National Laboratories

- We would like to find a method that :
  - capture any kind of relationship between input and output
  - capture conjoint influence
- Main Idea: Decompose the function into functions depending on any possible combinations of inputs

$$y = f(\mathbf{x}) = f_0 + \sum_{i=1}^{nX} f_i(x_i) + \sum_{i} \sum_{j>i} f_{ij}(x_i, x_j) + \cdots + f_{1,2,\ldots,nX}(x_1, x_2, \cdots, x_{nX})$$

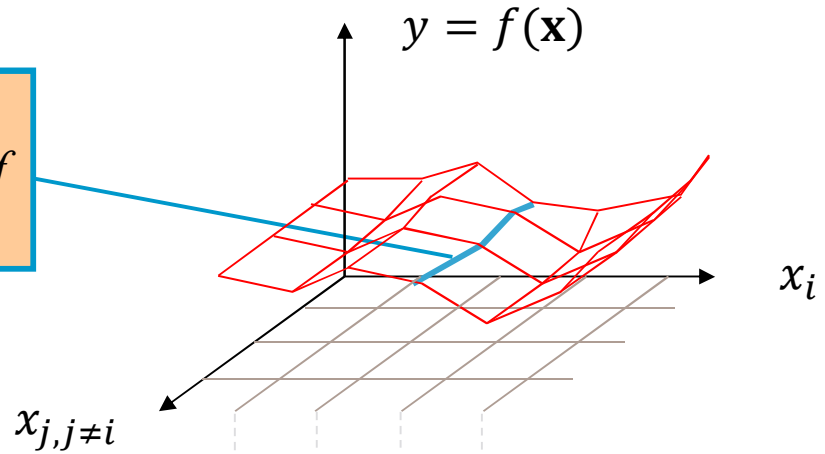
However, this decomposition is **NOT** unique

- If all the functions are **orthogonal** and if their expected value (except f_0) is zero then f_0=E[y] and the decomposition is unique

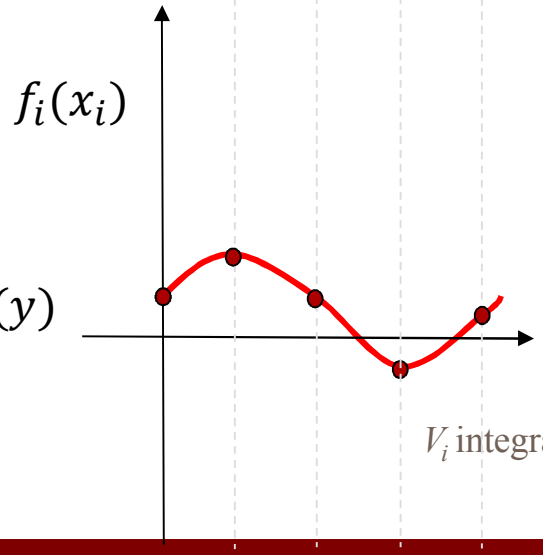
Variance based method

2: Sobol variance decomposition

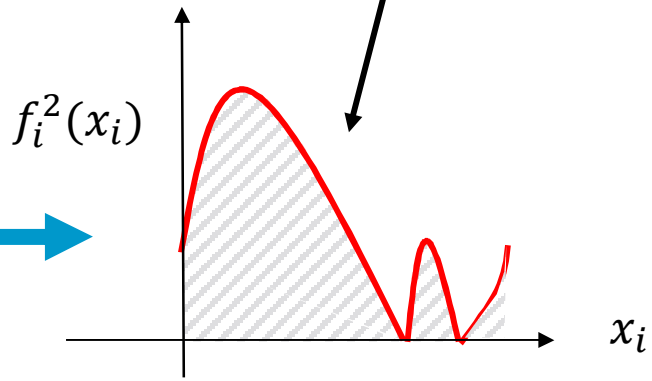
We calculate the average of the function f for a given value of x_i



$$f_i(x_i) = \mathrm{E}(y|x_i) - \mathrm{E}(y)$$



$$V_i = \int f_i^2(x_i)dx_i$$



V_i integration of the square of f_i on the whole range of x_i.

Variance based method

3: FAST

- **Basic Idea:** In the moments calculation, converting the output from a function of *nX* variables (i.e., the elements x_i of x) to a function of one variable (i.e., s) lead to convert the multi-dimensional integral to a mono-dimensional integral.

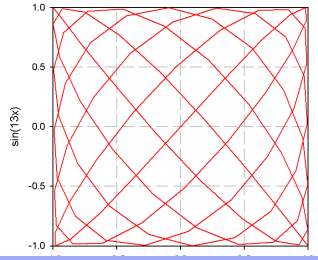
$$\int_{\Omega} f^r(\mathbf{x})\,p(\mathbf{x})\mathrm{d}\mathbf{x} = \int_{\Omega} f^r(x_1, x_2, \cdots, x_{nX})\,p(\mathbf{x})\mathrm{d}x_1\mathrm{d}x_2\ldots\mathrm{d}x_{nX}$$

$$\cong \frac{1}{2\pi}\int_{-\pi}^{\pi} f^r\left[G_1(\sin\omega_1 s), G_2(\sin\omega_2 s), \cdots, G_{nX}(\sin\omega_{nX} s)\right]$$

Each input x_i is associated with a unique frequency ω_i
The functions G_i are used to provide a better coverage of the domain

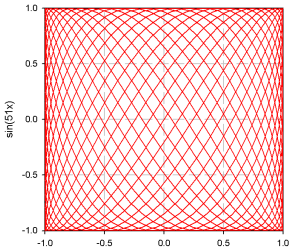
Small frequencies

Good approximation of search curve with a small number of points



But …
Bad coverage of the domain

Good coverage of the domain



Large frequencies

But …
Need large number of points for approximating the search function

Variance based method

3: First order and total sensitivity indices

$$V(y) = \sum_{i=1}^{nX} V_i + \sum_{j>i} V_{i,j} + \cdots + V_{1,2,\ldots,nX}$$

Dividing by $V(y)$

$$1 = \sum_{i=1}^{nX} S_i + \sum_{j>i} S_{i,j} + \cdots + S_{1,2,\ldots,nX}$$

Sensitivity indices

with

- S_i contribution of x_i to the variance of y
- S_{ij} contribution of the interaction of x_i and x_j to the variance of y
- …
- $S_{1,2,\ldots,nX}$ contribution of the interaction of all parameters to the variance of y

Total Order

- By fixing the value of all variables but x_i one can calculate the influence of all inputs with their interactions, except with x_i (S_{-i}).
- The difference $S_{Ti} = 1 - S_{-i}$ represents the influence of x_i solely and all its interaction with the others inputs.
- This index is called total sensitivity index of x_i.

Variance based method

4: conclusion on these methods

Sandia National Laboratories

Strong Points of Variance Decomposition Methods

- Capture nonlinear and nonmonotonic relationship between input and output
- Allows calculation of conjoint influence of two or more inputs

Weak Points of Variance Decomposition Methods

- Non negligible cost in number of simulations required
- Suppose input parameters are independent to each other

Enhanced sensitivity techniques

1. Methodology

- We would like to use a method that is not limited to linear/monotonic additive relations but does not cost as much as variance based methods
- The following strategy is proposed to capture the benefit of both approaches:
  - 1: create a surrogate model (response surface) using non linear technique, based on existing sample
  - 2: evaluate the appropriateness of the surrogate model by calculating coefficient of determination (R^2)
  - 3: if model appropriate, use it to generate complete variance decomposition (analytical model can be run a million times in a few seconds to a few minutes)
  - 4: estimate first order and total sensitivity indices on the surrogate model
- Examples of three regressions techniques follow.

Enhanced sensitivity techniques
2: regressions techniques examples

Sandia National Laboratories

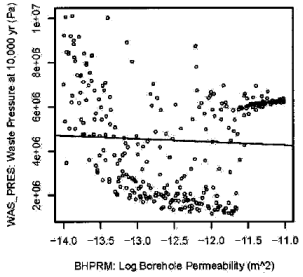
- Quadratic regression
  - A stepwise regression with all parameters (x_i), their square values (x_i^2) and any first order interaction ($x_i.x_j$)
  - **Advantages:** Captures some non-monotonic influence as long as they are close to quadratic, and also allows simple interaction for linear conjoint influence
  - **Limitations:** Method has difficulties finding more complex relation or conjoint influence, and this method is parametric and results can be affected by outliers
- Recursive partitioning
  - A decision tree is used to split the input data into area of influence. The method allow multiple splits and 2 parameters interactions. Order 0 polynomial response (i.e., constant) is generated in each defined region
  - **Advantage:** As for any decision tree analysis, the strong point of the method is to capture change in the output due to trigger points (if one variable is higher than a threshold and/or another variable is between two values) which could not be captured easily with other techniques
  - **Limitation:** The method may have a tendency to find relations where none exists, especially when the number of input variables is large compared to the sample size
- Multivariate adaptive regression splines (MARS)
  - This method is a combination of (linear) spline regression, stepwise model fitting and recursive partitioning
  - **Advantage:** The method leans towards the same flexibility as recursive partitioning with the robustness of rank regression in order to avoid over fitting
  - **Limitation:** Because of the use of spline, its efficiency is limited when used over discrete variables, especially if the number of discrete states is small (2 or 3 values). In such cases, it may completely miss the parameter's influence or underestimate it

Enhanced sensitivity techniques

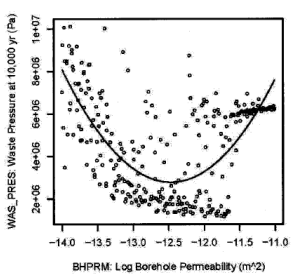
3: Comparison of techniques

- A: stepwise rank regression will capture monotonic influence and may not be sufficient in case of more complex relationship between input and output. However, this method is usually enough 75% to 80% of the cases. The fact is that even in complex analyses, most of the input influences are monotonic.
- B: quadratic regression allows the capture of (simple) non monotonic influence and **conjoint influence** (in the sense of X1 × X2).
- C: Recursive partitioning (aka Tree regression) will split the input space according to the value of the output (high, medium, low …)
- D: splines (such as used in the MARS technique) will consider the influence of input parameters as piecewise, considering smoothness in the local area.

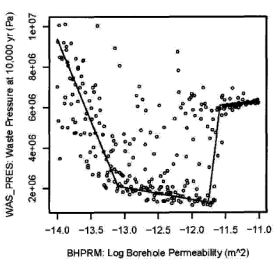
A



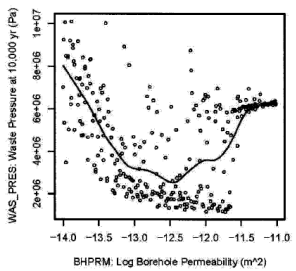
B



C



D



Development of strategy vs. reporting single number

Sandia National Laboratories

- The use of a family of regression techniques makes the analysis a little more complex than when a single stepwise regression technique is used however it seems a more appropriate strategy considering that:
  - Stepwise Linear regression has been shown to be insufficient for some analysis in which the input/output relation was not monotonic and additive
  - The other techniques look only at one aspect of non-monotonicity
  - The more sophisticated the technique is, the more likely it will overfit the data and find non physical relation due to the sample size. It is therefore risky to rely on a single technique
- For the same reason the use of replicated analyses (with different random seed) strengthened the confidence the analyst had on both the output distributions and resulting sensitivity analysis.

Reporting a single number does NOT help the decision maker as many assumptions are inherent of this number and lack of understanding may lead to taking the wrong decision. Even if it requires more work, it is more important to give full information and educate the decision maker

Conclusions on sensitivity Analysis

- Global sensitivity analysis, in conjunction with sampling based methods is the traditional approach used for complex system analysis
- This approach has been successfully used by Sandia for multiple Projects
- Sensitivity analysis is beneficial to any complex analysis even at early stage.
- While the original techniques were looking at linear and monotonic relations between uncertainty inputs and outputs, more sophisticated methods have been developed and offer a more complete understanding of the system behavior