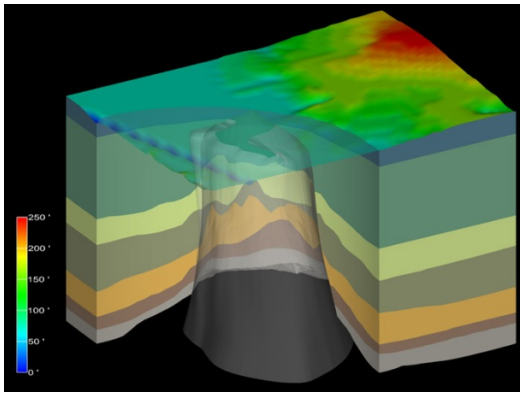
SAND2013-xxx

Exceptional service in the national interest

SAND2013-4900C

Sandia National Laboratories



An introduction to uncertainty and spatial variability

How can they be recognized, classified, measured and managed?

IAEA Workshop on Spatial variability – July 1-4, 2013

Cédric J. Sallaberry

U.S. DEPARTMENT OF ENERGY | NNSA 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.

Outline

- Definition of uncertainty
- Type of uncertainty
- Uncertainty representation
- Spatial Variability
- Uncertainty characterization
- Conclusion

Definition of uncertainty

Sandia National Laboratories

- Uncertainty represents all the unknowns that affect the outcome or result we are looking at.
- It encompass any lack of certainty such as poorly known inputs or randomness in future events, as well as lack of accuracy of models and techniques used.
- For analysis purposes, uncertainty can be classified in different group (aleatory, epistemic, variability). This allow a more rigorous treatment of uncertainty with a better definition of the uncertainty of the outcome.

Distinction between the different types of uncertainty

Sandia National Laboratories

- **Aleatory uncertainty**: (Perceived) randomness in the occurrence of future events
- **Epistemic uncertainty**: Lack of knowledge w.r.t. the appropriate value to use for a quantity that has a fixed, but poorly known, value in the context of a specific analysis.
- **Spatial variability:** inherent variability over space of a quantity, that usually cannot be measured precisely or at the expected scale.
- Probability usually used to characterize both aleatory uncertainty, epistemic uncertainty and spatial variability
- Alternatives to probability to the representation of epistemic uncertainty exist, such as evidence theory, possibility theory, interval analysis and others

Probabilistic framework

- Formal definition of probability involves three components
  - A set S that contains everything that could occur in the particular "universe" under consideration
  - A set $\mathbb{S}$ of subsets of S with the properties that (i) if $\varepsilon \in \mathbb{S}$, then $\varepsilon^c \in \mathbb{S}$ and (ii) if $\{\varepsilon_i\}$is a countable collection of elements of $\mathbb{S}$, then $\bigcup_i \varepsilon_i$and $\bigcap_i \varepsilon_i$are elements of $\mathbb{S}$
  - A function p_s defined for elements of $\mathbb{S}$ such that (i) $p_s(S)=1$, (ii) if $\varepsilon \in \mathbb{S}$, then $0 \le p_s(\varepsilon) \le 1$, and (iii) if $\{\varepsilon_i\}$ is a countable collection of disjoint elements of $\mathbb{S}$, then $p_s(U\varepsilon_i)=\Sigma_i p_s(\varepsilon_i)$
- Triple $(S, \mathbb{S}, p_s)$ is called a probability space
- Terminology
  - S called the sample space or universal set
  - Elements of S are called elementary events
  - Elements of $\mathbb{S}$ are called events
  - P_s called a probability measure

Evidence Theory
Definition of Evidence Space

Sandia National Laboratories

- Formal definition an evidence theory representation of uncertainty involves 3 components
  - A set S that contains everything that could occur in the particular "universe" under consideration
  - A (countable) set $\mathbb{S}$ of subsets of S
  - A function m defined for subsets ε of S such that (i) $m(\varepsilon)>0$ if $\varepsilon \in \mathbb{S}$, (ii) $m(\varepsilon)=0$ if $\varepsilon \notin \mathbb{S}$ and (iii) $\Sigma_{\varepsilon \in \Sigma} m(\varepsilon)=1$
- Triple (S, $\mathbb{S}$, m) is called an evidence space
- Terminology
  - S called the sample space or universal set
  - Elements of S are called elementary events
  - Elements of $\mathbb{S}$ are called focal elements
  - m called a basic probability assignment (BPA)
- Nature of $m(\varepsilon)$: Amount of "likelihood" that is associated with ε but cannot be further partitioned to subsets of ε.

Evidence Theory
Representation of Uncertainty

Sandia National Laboratories

- Representation of uncertainty
  - Belief
  - Plausibility

- Belief:
  - Definition: $Bel(\varepsilon) = \sum_{U \subset \varepsilon} m(U)$
  - Concept: Amount of "likelihood" that must be associated with E.

- Plausibility:
  - Definition: $Pl(\varepsilon) = \sum_{U \cap \varepsilon \neq \emptyset} m(U)$
  - Concept: Amount of "likelihood" that could potentially be associated with E.

Possibility Theory
Definition of Possibility Space

Sandia National Laboratories

- Formal definition a possibility theory representation of uncertainty involves 2 components
  - A set S that contains everything that could occur in the particular "universe" under consideration
  - A function r such that (i) $0 \leq r(x) \leq 1$ for $x \in S$ and (ii) $\sup\{ r(x): x \in S \} = 1$
- Doublet (S, r) is called a possibility space
- Terminology
  - S called the sample space or universal set
  - r is referred to a possibility distribution function
- Nature of r: Amount of "likelihood" or "credence" that can be assigned to each element of S. Analogous to membership value for elements of a fuzzy set.

Possibility Theory
Representation of Uncertainty

Sandia National Laboratories

- Representation of uncertainty
  - Possibility
  - Necessity

- Possibility:
  - Definition: $Pos(\varepsilon) = sup\{r(x): x \in \varepsilon\}$

  - Concept: Measure of amount of information that does not refute the proposition that E contains the "correct" value for *x*.

- Necessity:
  - Definition: $Nec(\varepsilon) = 1 - Pos(\varepsilon^c) = 1 - sup\{r(x): x \in \varepsilon^c\}$

  - Concept: Measure of amount of uncontradicted information that supports the proposition that E contains the "correct" value for *x*.

Interval Analysis

Sandia National Laboratories

- Define range of values for x
- Determine resultant range of values for $y = f(x)$
- No uncertainty structure imposed on x, only range of values
- Different in spirit from probability theory, evidence theory and possibility theory representation of uncertainty
- Corresponds to "degenerate" evidence theory and possibility theory representation of uncertainty
  - Evidence theory: sample space has BPA of 1
  - Possibility theory: Possibility distribution function identically equal to 1

Comparison of different uncertainty representations - setup

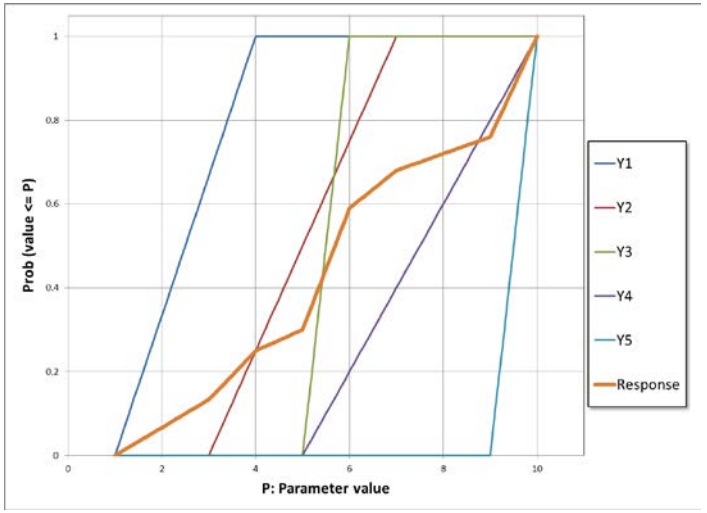
Sandia National Laboratories

- 5 experts are asked to define a parameter of interest. The results are that:
  - Expert 1 considers that it could be between 1 and 4
  - Expert 2 considers that it could be between 3 and 7
  - Expert 3 considers that it could be between 5 and 6
  - Expert 4 considers that it could be between 5 and 10
  - Expert 5 considers that it could be between 9 and 10
- There are no reason to trust more one expert than another.

- The question one tries to answer is how likely the value will be between 2 and 8 (Response=[2,8]).

Comparison of different uncertainty representations – probabilistic theory

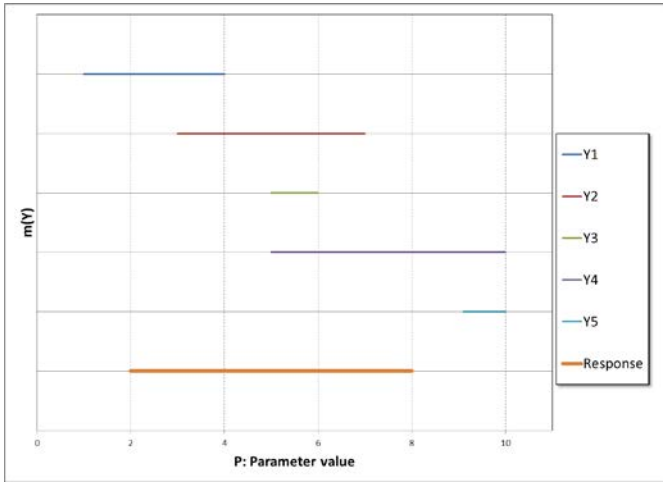
- Each expert is associated with a weight of 1/5 (no reason to favor one vs. another).
- Each range is represented using uniform distribution (only min and max are available).
- $Prob(Response) = Prob(2 \leq P \leq 8) = Prob(P \leq 8) - Prob(P \leq 2) = \frac{49}{75} \sim 0.6533$
- The answer is a single number
- But the framework introduces some subjectivity when the uniform distribution is selected.



Comparison of different uncertainty representations – evidence theory

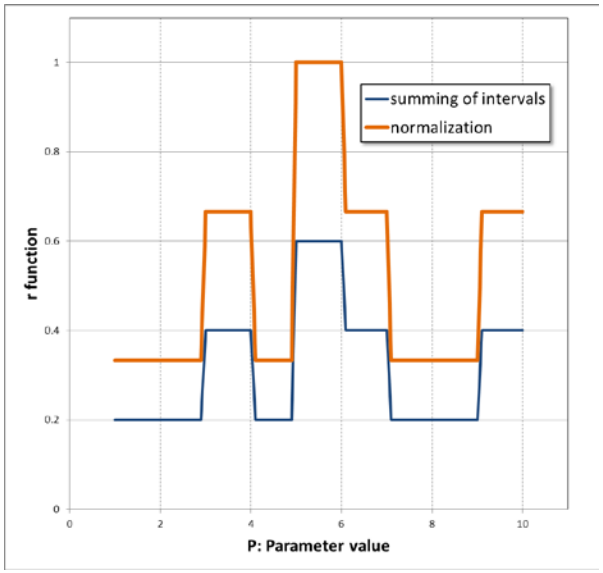
- Each expert is associated with a weight of 1/5 (no reason to favor one vs. another).
- Each range is represented as an interval, whose measure is 1/5
  - $Bel(Response) = \sum_{Y_i \subset R} m(Y_i) = m(Y_2) + m(Y_3) = 2/5$
  - $Pl(Response) = \sum_{Y_i \cap R \neq \emptyset} m(Y_i) = m(Y_1) + m(Y_2) + m(Y_3) + m(Y_4) = 4/5$
  - No subjectivity is introduced via a distribution selection
  - As a result, a range (instead of a value) is given. The solution is between 2/5 and 4/5

Y_1=[1,4], $m(Y_1)$ = 1/5
Y_2=[3,7], $m(Y_2)$ = 1/5
Y_3=[5,6], $m(Y_3)$ = 1/5
Y_4=[5,10], $m(Y_4)$ = 1/5
Y_5=[9,10], $m(Y_5)$ = 1/5



Comparison of different uncertainty representations – possibility theory

- Each expert is associated with a weight of 1/5 (no reason to favor one vs. another).
- The function *r* is represented by summing the weights whose intervals cover the region of interest and normalizing the resulting function so its sup over the region is 1.
- $Pos(Response) = sup\{r(x): x \in [2,8]\} = 1$
- $\mathrm{Nec}(Response) = 1 - sup\{r(x): x \in [0,2[\cup]8,10]\} = 1 - \frac{2}{3} = \frac{1}{3}$
- As a result, a range (instead of a value) is given. The solution is between 1/3 and 1



Comparison of different uncertainty representations – interval analysis

Sandia National Laboratories

- In such situation, interval analysis will look at any possible interval response.
- If Expert 5 is right then the answer is that there is zero chance to be between 2 and 8
- If Experts 2 or 3 at right, then the answer will always be between 2 and 8

- As a result, the interval analysis gives a range from 0 to 1.

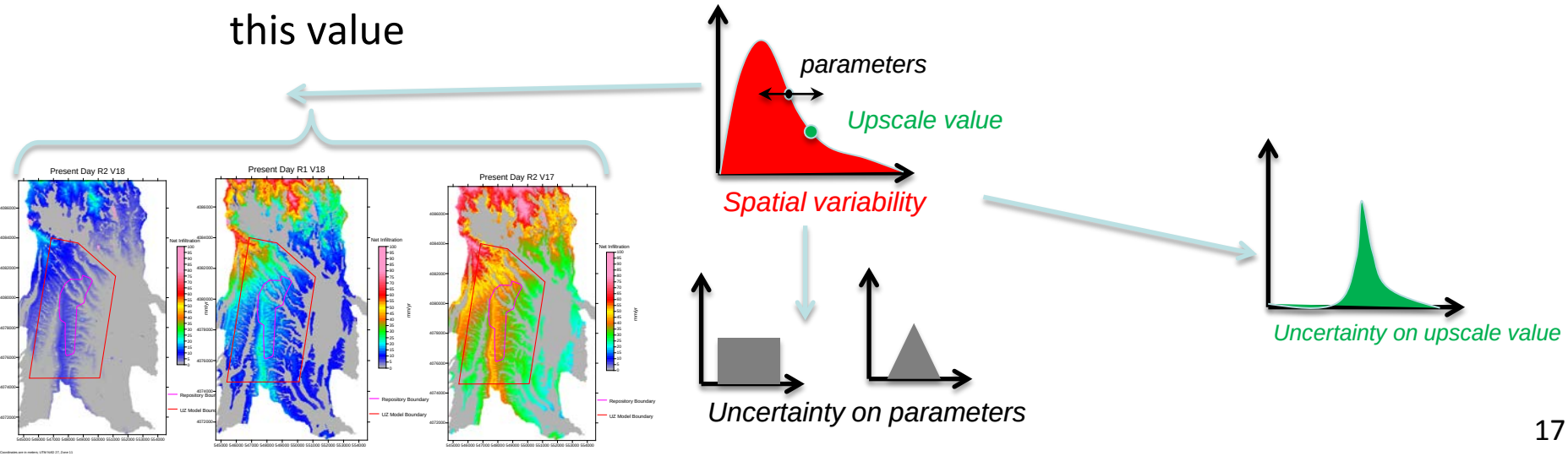
- In such example, the interval analysis is not really informative.

Spatial variability

- Spatial variability is sometimes classified as aleatory uncertainty (as randomness in space) and sometimes as epistemic (as not random but poorly known) and sometimes has its own category
- Historically, spatial variability has been studied separately from uncertainty analysis and has its own branch of study (geostatistics). Next set of lectures will focus on presenting these techniques.

Spatial variability and uncertainty

- In a complex system analysis, uncertainty analysis can capture spatial variability in different way:
  - Generate potential spatial variability maps matching the size and dimension defined for the codes using the spatially variable quantities and sample over these maps selections at each (aleatory and/or epistemic) realization
  - Sample the epistemic parameters defining the spatial variability distribution (representing lack of knowledge over spatial variability)
  - Generate upscale value and sample over (epistemic) uncertainty on this value



Synthetizing information to form uncertainty distribution

Sandia National Laboratories

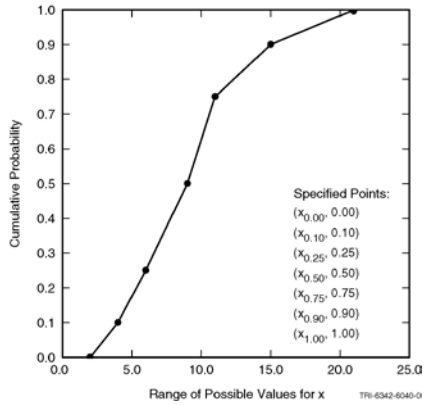
- Can be most important, but most expensive, part of an analysis. Without the correct characterization, the uncertainty reflected in the results will be wrong.
- Uncertainty characterization must be consistent with use of a variable in the model and analysis under consideration
- Many models use spatially or temporally averaged quantities as input
- Do not confuse spatial or temporal variability with the uncertainty in the appropriate value to use for an average over space and time.

Uncertainty Characterization: Expert Review (1/4)

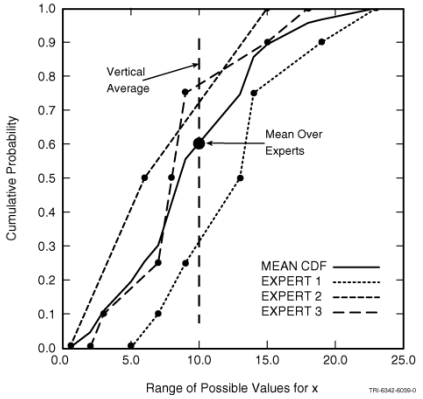
- **Used when no data is available.**
- **e**=[e_1,e_2,…,e_{nE}] vector of epistemically uncertain analysis inputs
- Derive distribution D_i for each element e_i of **e**
- Correlations and other restrictions between elements of **e** possible
- Usually obtained through expert review process that draws on
  - Observational and experimental data of varying levels of relevance
  - Knowledge of relevant processes and physics
  - Requires care, thought and thorough documentation
  - Not an individual expounding with feet on desk !
- Definition of distribution D_i for element e_i of **e**
  - Recommended: Specify minimum, maximum and selected quantiles
  - Not recommended: Specify named distribution and associated parameters
  - Specify any appropriate correlations and restrictions
  - Document !

Epistemic Uncertainty characterization Expert Review (2/4)

- Possibilities
  - Single "expert" for each element e_i of **e**
  - Multiple "experts" for each element e_i of **e**



Single expert



Multiple experts (Equal weight to each expert)

- Scope of review
  - Relatively small study in which single analyst both develops the uncertainty characterizations and carries out the analysis
  - Large analysis on which important decisions will be based and for which uncertainty characterizations are carried out by teams of outside experts who support the analysts actually performing analysis.

Epistemic Uncertainty characterization Expert Review (3/4)

- Example expert review strategy (used in NRC's NUREG-1150 analyses)
  - Seek broad input in selection of experts and diversity in selected experts
  - Structured review process with sequence of meetings
  - Meeting 1: (i) educate reviewers on the use of probability to characterize epistemic uncertainty, (ii) Describe analysis and uncertain variables to be characterized, (iii) Supply known information on uncertain variables, and (iv) send experts home to think about problem.
  - Meeting 2: (i) Have each expert present views on requested uncertainty characterization and any additional relevant information that has been found but no uncertainty characterization, and (ii) send experts home to think about problem.
  - Meeting 3: Elicitation team (elicitation expert, project analyst, scribe) meets separately with each expert to obtain probability distributions characterizing epistemic uncertainty

- Details: Hora SC, Iman RL. Expert Opinion in Risk Analysis: The NUREG-1150 Methodology. *Nuclear Science and Engineering* 1989; 102(4):323-331.

Epistemic Uncertainty characterization Expert Review (4/4)

- Possible analysis strategy
  - Perform initial exploratory analysis with "crude" characterization of the distributions $D_1, D_2, \ldots D_{nE}$ characterizing epistemic uncertainty
  - Use sensitivity analysis to determine the elements of **e** that dominate the uncertainty in analysis outcomes of interest
  - Perform detailed uncertainty assessments for the important variables identified in the sensitivity analysis
  - Carry out final decision-supporting analysis with new distributions

- Desiderata in epistemic uncertainty assessment
  - Avoid being either deliberatively optimistic (i.e., non-conservative) or deliberatively pessimistic (i.e., conservative) in uncertainty assessments
  - Be honest w.r.t. the uncertainty that is present

Uncertainty Characterization: Bayesian updating

- **Used when data becomes available, to update expert elicitation**
- Start with initial uncertainty characterization (Prior distribution)
- Obtain new information (experiment, observation)
- Update uncertainty characterization with Bayes theorem (Posterior distribution)

Posterior probability For x_i

Prior probability For x_i

Likelihood: probability of observing information I if x_i is correct value for x

Possible value for x out of i possible values: $x_1, x_2, \ldots, x_n$

$$p_{\text{pos}}\left(x_i \middle| I\right) = \frac{p_{\text{pr}}\left(x_i\right) p_{\text{L}}\left(I \middle| x_i\right)}{\sum_{j=1}^{\text{n}} p_{\text{pr}}\left(x_j\right) p_{\text{L}}\left(I \middle| x_j\right)}$$

New information from experiment or observation

Note: Bayes theorem with continuous probability distributions involves density functions and integrals rather than probabilities and sums

Uncertainty Characterization: Maximum Entropy

- **Used when some data is available**
- Entropy $H(x)$ measure of uncertainty inherent in a distribution for x

$$H(x) = \begin{cases} -\sum_{i=1}^{n} p_i \log p_i & , \text{ discrete}: p_i \text{ probability of } x_i \\ -\int_X d(x) \log[d(x)] dx & , \text{ continuous}: d(x) = \text{density for } x \text{ on } X \end{cases}$$

(convention: 0 log 0 = 0)

- $H_1(x) < H_2(x)$ => more inherent uncertainty in distribution producing $H_2(x)$ than in distribution producing $H_1(x)$
- Maximum entropy can be used to complete the definition of incompletely defined probability distributions with the least incorporation of additional information
- Examples of maximum entropy distributions
  - Only range [a,b] known => uniform [a,b]
  - Only quantiles known => piecewise uniform between quantiles
  - Only mean and variance known => normal
  - Only range [a,b] , mean and variance known => beta on [a,b]
  - Only range [0,∞) and mean known => exponential

Uncertainty Characterization: Bootstrap

- **Used when some data is available**
- Start with initial data set of size n
- Generate R new bootstrap samples of size n by sampling with replacement from original data set
- Estimate quantity x of interest for each of the R bootstrap samples (i.e., obtain x_i for i=1,2,…,R)
- Use the results x_1, x_2,… , x_R to estimate uncertainty in x

Uncertainty Characterization: data fitting

Sandia National Laboratories

- **Used when a lot of data is available.**
- Fit several distributions to the data and select the one that fit "best" the data.
- Definition of best fitting depend on the analysis and will guide into which fitting technique is the most appropriate:
  - Moment fitting if mean and stdev have to be preserved for instance
  - Distribution fitting to minimize the difference between distributions (Anderson-Darling or Kolmogorov-Smirnov test)
  - Fitting of one particular part of the distribution (tail …)
  - Several software are available that can do fitting to a large range of distribution types.

Conclusions

- Uncertainty is an inherent part of any complex system analysis and needs to be addressed.
- Uncertainty can be classified into aleatory and epistemic. Probabilistic approach is the traditional approach to deal with uncertainty but other methods (evidence theory, interval analysis, possibilistic theory…) can be used when appropriate
- Spatial variability is different than uncertainty but is closely linked to it and uncertainty has to be considered to represent spatial variability
- Several techniques and tools are available for uncertainty characterization, depending on the level of information available.

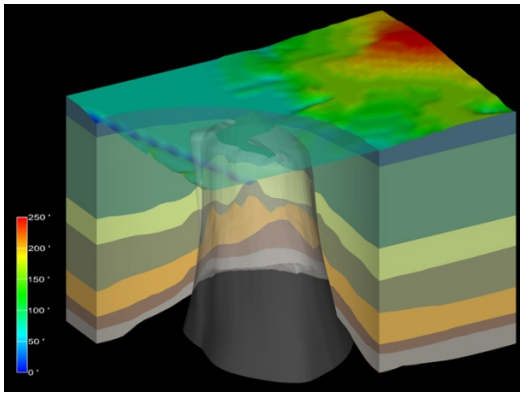
References

- **A. Saltelli, K. Chan, E.M. Scott** (Ed.) *Sensitivity Analysis* – Wiley series in probability and statistics (2001)
- **Shafer G.** *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton Univ. Press 1976.
- **Klir GJ.** *Uncertainty and Information: Foundations of Generalized Information Theory*. New York, NY: Wiley-Interscience, 2006.
- **Dubois D, Prade H.** *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. New York, NY: Plenum 1988.
- **Helton JC, Johnson JD, Oberkampf WL.** An Exploration of Alternative Approaches to the Representation of Uncertainty in Model Predictions. *Reliability Engineering and System Safety* 2004; 85(1-3):39-71.
- **Cooke RM**. *Experts in Uncertainty: Opinion and Subjective Probability in Science. Oxford; New York: Oxford University Press, 1991.*
- **Shannon CE, Weaver W.** *The Mathematical Theory of Communication. Urbana, IL: University of Illinois Press, 1949*
- **Press SJ.** *Bayesian Statistics: Principles, Models and Applications. New York, NY: Wiley, 1989.*
- **Efron B, Tibshirani RJ.** *An Introduction to the Bootstrap. Boca Raton, FL: Chapman and Hall, 1993.*

SAND2013-xxx

Exceptional service in the national interest

Sandia National Laboratories



Uncertainty treatment - exercice

IAEA Workshop on Spatial variability – July 1-4, 2013

Cédric J. Sallaberry

U.S. DEPARTMENT OF ENERGY | NNSA 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.

Example used: set up

inputs

- X_1, X_2, X_3 and X_4 uniformly distributed between 0 and 1
- Corr(X_1,X_4)= 0.8 . All other correlations are set to 0
- ε is an input representing random noise which varies between -0.2 and 0.2

output

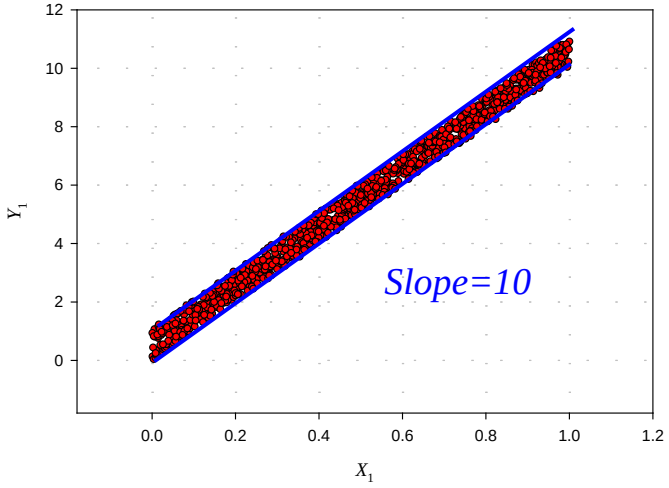
$$Y_1 = 10X_1 + X_2 + 0.1X_3 + 0.01X_4$$

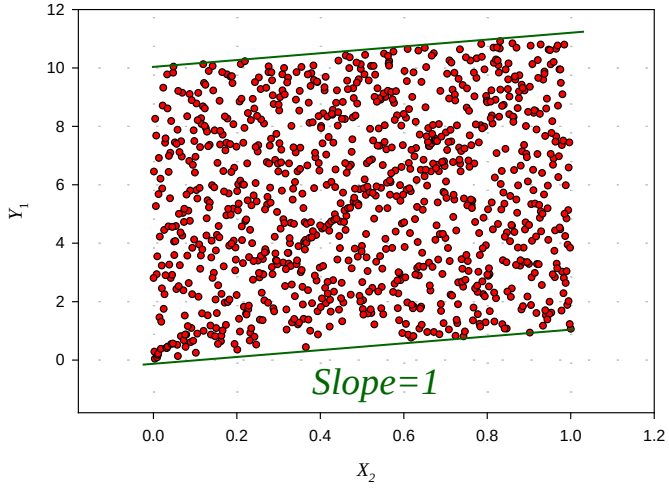
$$Y_2 = 10X_1 + X_2 + 0.1X_3 + 0.01X_4 + \varepsilon$$

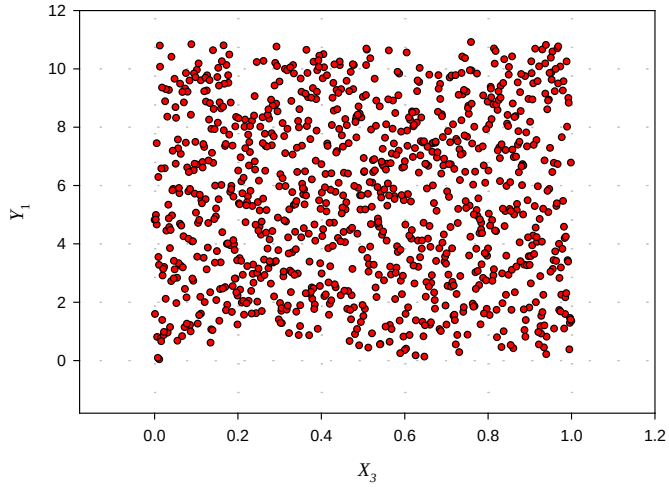
sampling

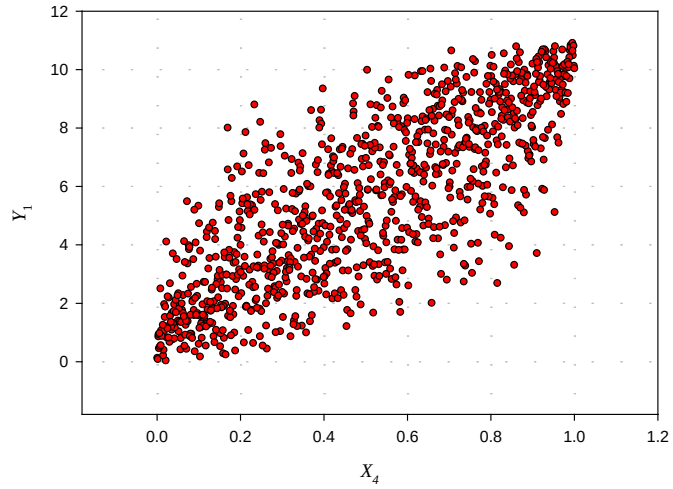
- Latin Hypercube Sampling (LHS)
- Iman-Conover correlation technique
- Sample size = 1,000

Example used: result for Y_1

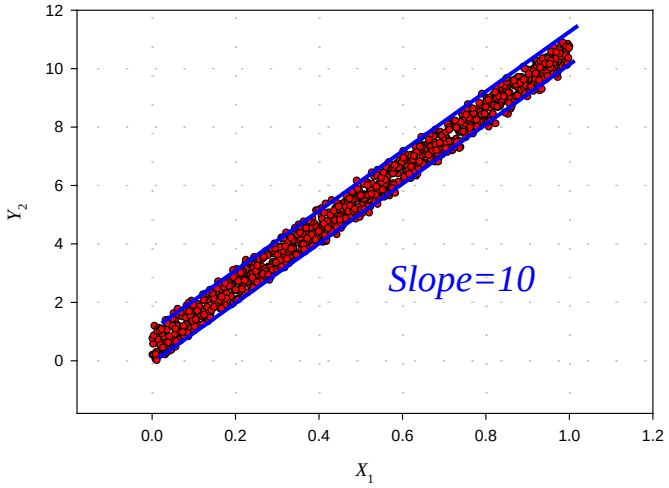


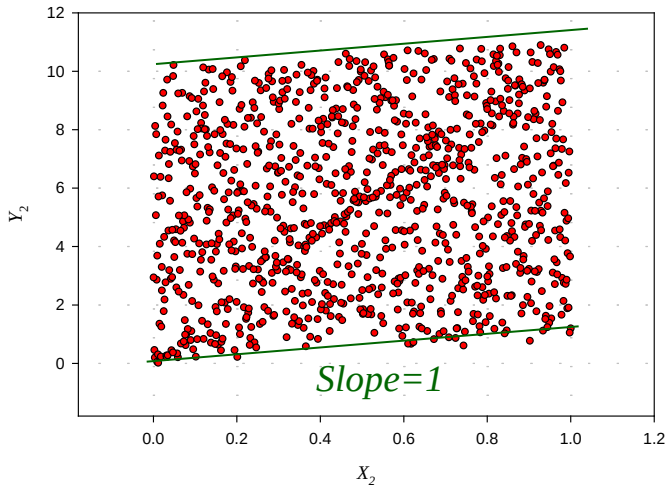


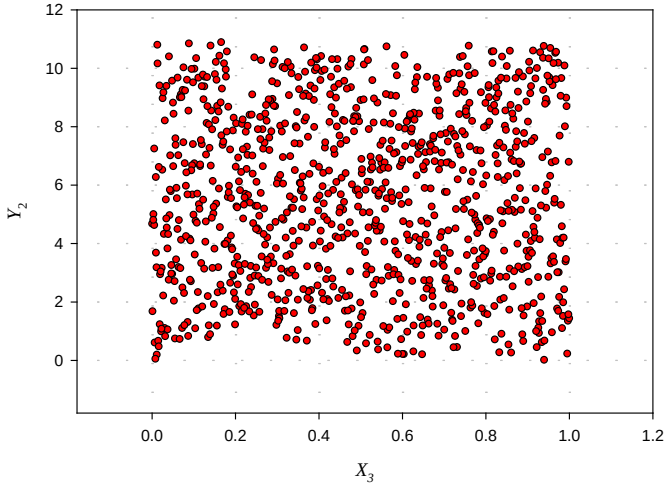


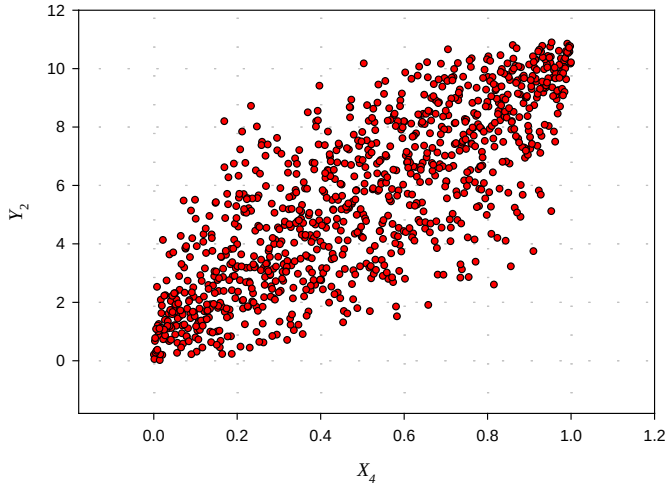


Example used: result for Y_2









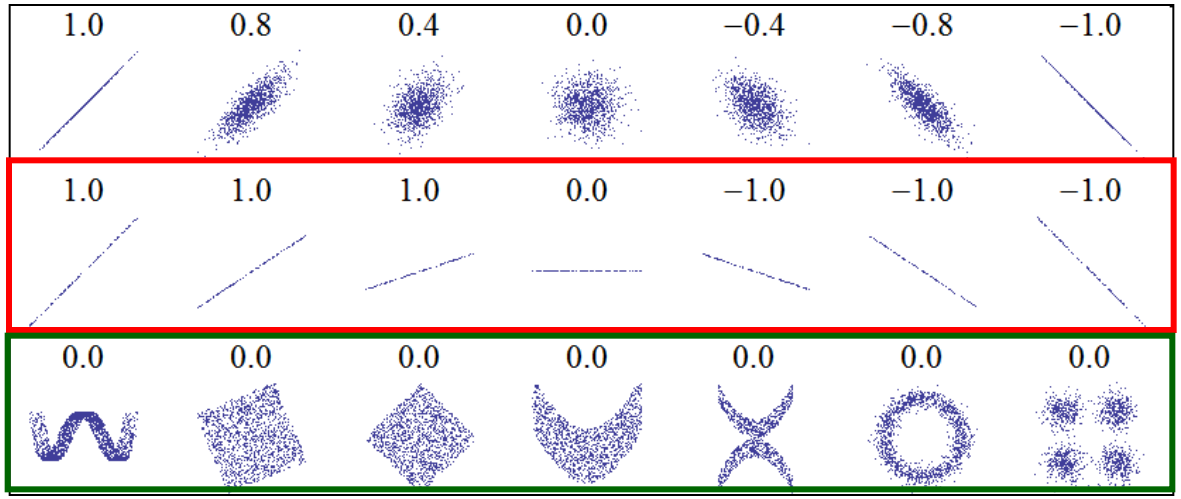
Random noise does not seem to influence the scatterplots

Correlation Coefficient (CC)

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



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

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

Correlation Coefficient (CC) in the example

Results for Y_1

	X_1	X_2	X_3	X_4	Y_1
X_1	1.000	0.040	-0.001	**0.800**	**0.995**
X_2	0.040	1.000	0.021	0.112	0.139
X_3	-0.001	0.021	1.000	0.045	0.011
X_4	**0.800**	0.112	0.045	1.000	**0.804**
Y_1	**0.995**	0.139	0.011	**0.804**	1.000

Even if X_4 has negligible effect in the equation of Y, its correlation coefficient is strong due to its correlation with X_1

Results for Y_2

	X_1	X_2	X_3	X_4	Y_2
X_1	1.000	0.040	-0.001	**0.800**	**0.995**
X_2	0.040	1.000	0.021	0.112	0.139
X_3	-0.001	0.021	1.000	0.045	0.011
X_4	**0.800**	0.112	0.045	1.000	**0.804**
Y_2	**0.995**	0.139	0.011	**0.804**	1.000

Random noise does not seem to influence the correlation

Partial Correlation Coefficient (PCC)

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

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

Partial Correlation Coefficient (PCC) in the example 1/2

In absence of noise, as the model is perfectly linear, each PCC is equal to 1.00 this reflects that each parameter has a perfect linear influence on the output Y_1

PCC values	X_1	X_2	X_3	X_4
Y_1	1.00	1.00	1.00	1.00
Y_2	1.00	0.93	0.25	0.02

When random noise is added, the linear influence of the parameter is partially or totally hidden. Depending of it's importance in the equation, a parameter may have a PCC varying from 1.00 to 0.02 in the analysis of Y_2

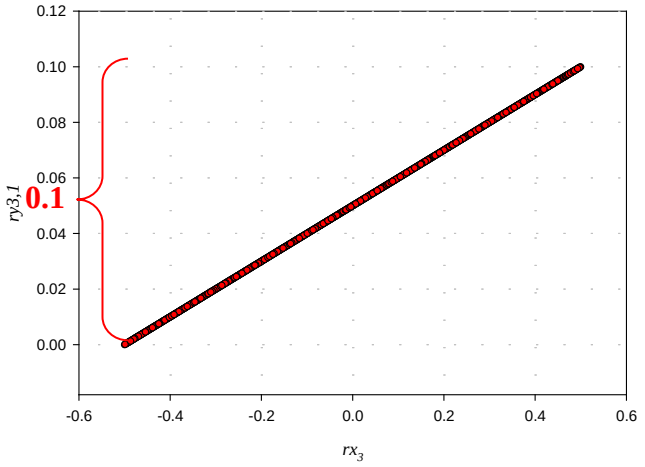
Note that ε would be in the group of input parameters, all PCC would be equal to 1.00 for Y_2

Partial Correlation Coefficient (PCC) in the example 2/2

$$ry_{3,1} = Y_1 - \tilde{Y}_{3,1} = 0.1X_3$$
$$rx_3 = X_3 - 0.5$$

$X_3 \in [0,1]$

$Y_1 \in [0,0.1]$

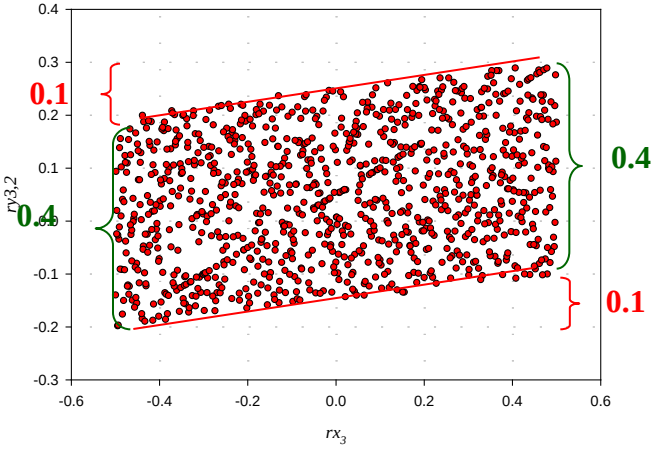


$$\tilde{Y}_{3,1} = \tilde{Y}_{3,2} = 10X_1 + X_2 + 0.01X_4$$
$$\tilde{X}_3 = 0.5$$

$$ry_{3,2} = Y_2 - \tilde{Y}_{3,2} = 0.1X_3 + \varepsilon$$
$$rx_3 = X_3 - 0.5$$

$X_3 \in [0,1]$; $\varepsilon \in [-0.2, 0.2]$

$Y_2 \in [-0.2,0.3]$



Regression Coefficient (RC)

Coefficients of the Linear Regression model, estimated using Least Square

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

model

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

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.**

$$\min_{\theta} f(\theta) = \min_{\theta} \left\| Y_j - \tilde{Y}_j \right\|^2 = \min_{\theta} \left\| Y_j - \theta X \right\|^2$$

$$\rightarrow f'(\theta) = 0$$

$$\rightarrow 2X^T (Y_j - \theta X) = 0$$

$$\rightarrow (X^T X)\theta = X^T Y$$

$$\rightarrow \theta = (X^T X)^{-1} X^T Y$$

Regression Coefficient (RC) in the example 1/2

$$\theta = \left(X^T X\right)^{-1} X^T Y_1 \rightarrow \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix} = \begin{pmatrix} -2.6\times10^{-8} \\ 10 \\ 1 \\ 0.1 \\ 0.01 \end{pmatrix}$$

$$Y_1 = 10X_1 + X_2 + 0.1X_3 + 0.01X_4$$

$$\tilde{Y}_1 = 10X_1 + X_2 + 0.1X_3 + 0.01X_4 - 2.6\times10^{-8}$$

$$\theta = \left(X^T X\right)^{-1} X^T Y_1 \rightarrow \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix} = \begin{pmatrix} -0.03 \\ 10.03 \\ 1.02 \\ 0.101 \\ 0.014 \end{pmatrix}$$

$$Y_2 = 10X_1 + X_2 + 0.1X_3 + 0.01X_4 + \varepsilon$$

$$\tilde{Y}_2 = 10.03X_1 + 1.02X_2 + 0.101X_3 + 0.014X_4 - 0.03$$

Regression Coefficient (RC) in the example 2/2

Note : values are not normalized. Changing the unit will change the result

Let change X_2 as uniformly distributed between 0 and 0.1
Y_1 will then be defined with:

$$Y_1 = 10X_1 + 10X_2 + 0.1X_3 + 0.01X_4$$

$$\theta = \left(X^T X\right)^{-1} X^T Y_1 \rightarrow \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix} = \begin{pmatrix} -2.6\times10^{-8} \\ 10 \\ 10 \\ 0.1 \\ 0.01 \end{pmatrix}$$

The change in unit changes the value of the regression coefficient associated with X_3 while its influence does not change

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

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

Linear regression

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

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

Standardized Regression Coefficient (SRC) in the example

SRC values	X_1	X_2	X_3	X_4
Y_1	0.99	0.1	0.01	0.00
Y_2	0.99	0.1	0.01	0.00

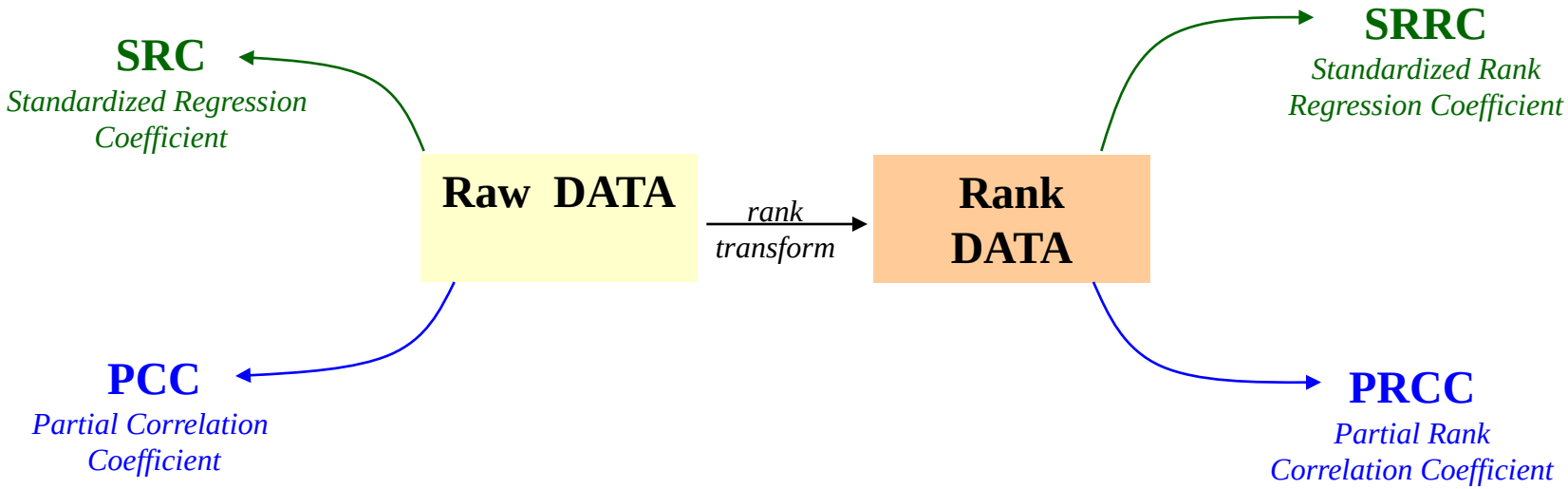
Results are the same whether there is random noise or not.

Working with Rank 1/2

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**



Working with Rank 2/2

Advantages

- It extends the relation from linear to monotonic patterns
- It does not involve any additional calculation than ranking
- It does not create over fitting

Drawbacks

- As it is a non-parametric method, it cannot be used for prediction
- It Still does not capture non-monotonic pattern or conjoint influence of multiple input parameters

Coefficient of Determination (R^2)

The coefficient of Determination, noted R^2 measures the part of the variance of the output that is explained by the regression model.

$R^2 \sim 1$: 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

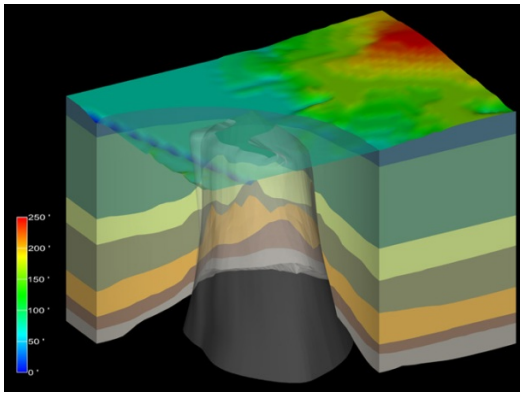
Conclusive Remarks

- The rank regression is a powerful tool for performing sensitivity analysis.
  - It captures monotonic relations which are the most common in the type of problem we deal with
  - It is fast and can thus be applied on large set of data
- This technique is pretty robust and does not usually over-fit (i.e. over-estimate the value of R^2)
- The R^2 is an indicator of the quality of the regression, informing when more analyses are necessary to understand the system.
- Experience shows that most of the time, even when the R^2 is low, the most influent parameters are selected (their influence is however under-estimated)

SAND2013-xxx

Exceptional service in the national interest

Sandia National Laboratories



Uncertainty analysis in site characterization and safety assessment

IAEA Workshop on Spatial variability – July 1-4, 2013

Cédric J. Sallaberry

U.S. DEPARTMENT OF ENERGY | NNSA 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.

Outline

- Introduction
- Purpose of uncertainty analysis
- Classical tools
- Example of two different analyses
- Conclusions

Introduction

Sandia National Laboratories

- Indicated importance now widely recognized
- Novelty of simply doing a complex calculation now past
- Greater use of computational analyses to support decisions
- Questions that are asked:
  - What is the uncertainty in the calculated result ?
  - How does this uncertainty affect the decision making process?

- Acknowledging uncertainty and presenting a distribution of results instead of a single number is a **more honest and more insightful** representation of the current state of knowledge.

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.

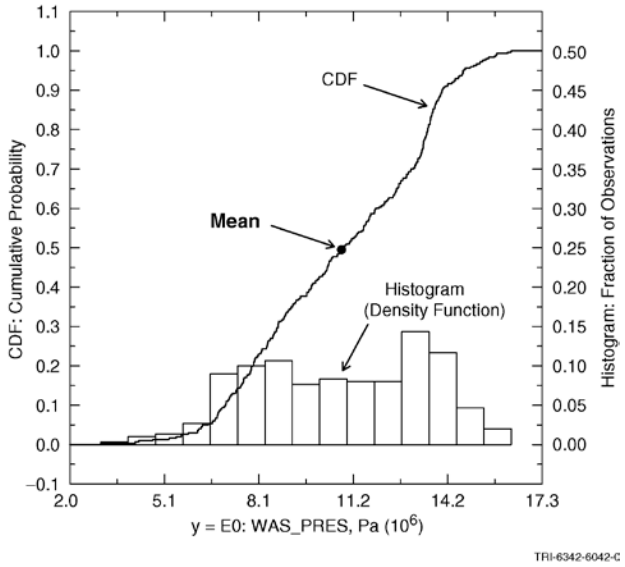
Classical statistics

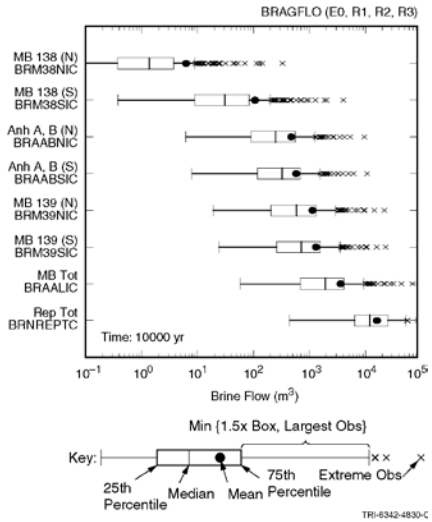
- Mean (over aleatory uncertainty, epistemic uncertainty or both) is always presented as a central tendency. It is often a good summary that included both the consequence of an event and the likelihood
- While the mean is usually a required value it is **Not** sufficient as it groups all the information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

WIPP and Yucca Mountain Two different PAs.

Sandia National Laboratories

- ***The Yucca Mountain Project (YMP) differs from the Waste Isolation Pilot Plant (WIPP) in several important ways:***
  - *<u>Difference in waste type :</u> Yucca Mountain is intended for the disposal of commercial spent nuclear fuel and defense high-level waste. WIPP is intended for the disposal of defense transuranic waste*
  - *<u>Difference in geology:</u> The repository for Yucca Mountain was planned to be in the unsaturated zone in volcanic tuff. The WIPP repository y is located in the saturated zone in bedded salt*
  - *<u>Difference in regulatory:</u> NRC defined a regulatory limit on **Expected Dose** (over aleatory) for Yucca Mountain, while EPA defined containment limit on normalized released for WIPP.*
  - *This is a difference not only in the value chosen but also the **mathematical framework** in which the output is defined.*

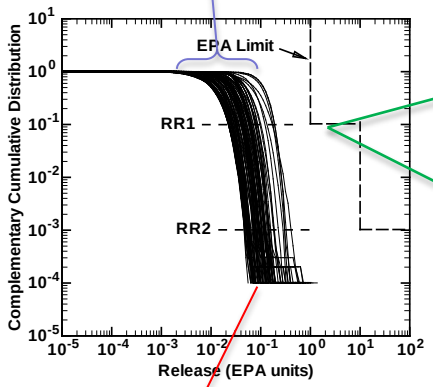
Examples of uncertainty analyses 1:Waste Isolation Power Plant (WIPP)

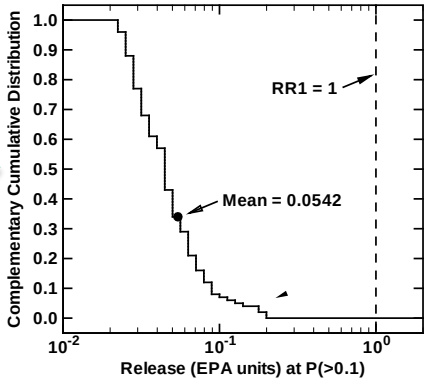
Sandia National Laboratories

- Regulation on actual Releases

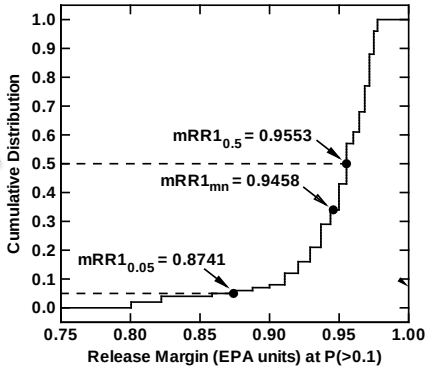
The spread of curve represents the effect of epistemic uncertainty (i.e. lack of knowledge)

Quantiles over **aleatory uncertainty** can be read from the Y-axis





At a given probability level epistemic uncertainty can be represented as a CDF or CCDF

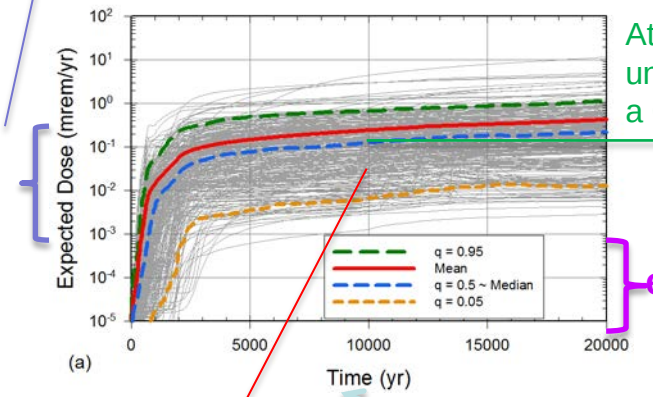


Each curve represents the effect of aleatory uncertainty (i.e. randomness)

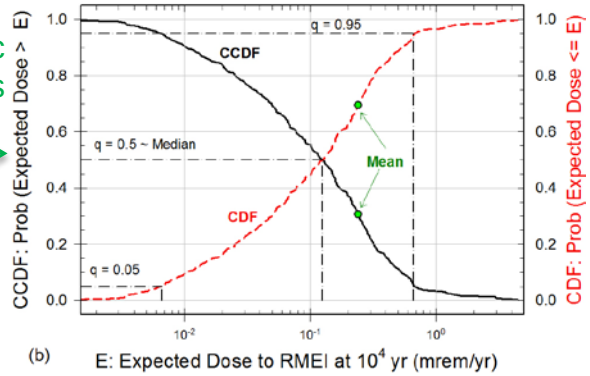
Examples of uncertainty analyses 2:Yucca Mountain Project (YMP)

- Regulation over expected (over aleatory) dose.

The spread of curve represents the effect of epistemic uncertainty (i.e. lack of knowledge)

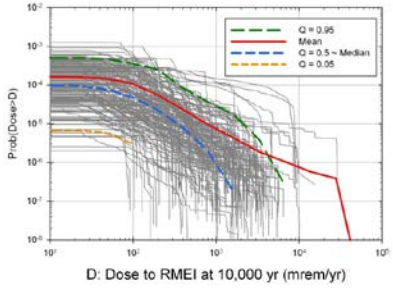


At each time-step epistemic uncertainty can be represented as a CDF or CCDF



Quantiles over **epistemic uncertainty** are estimated at each timestep

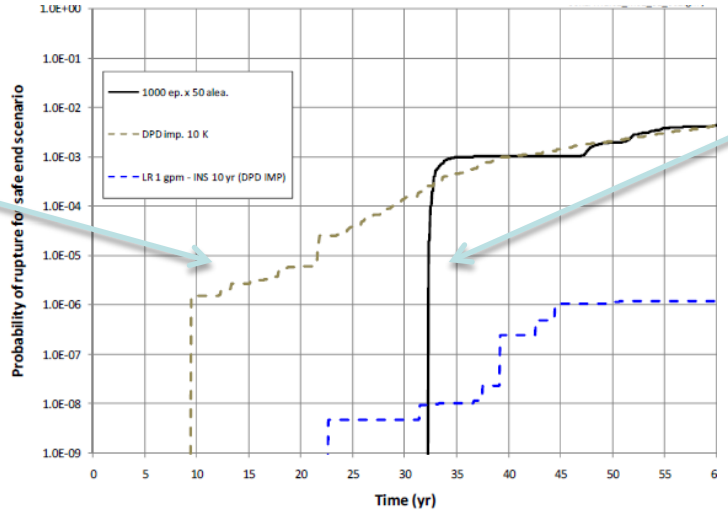
At each time-step, aleatory uncertainty is reduced to a single point (expected value) BUT all the information is available if needed.



Beyond traditional sampling-based methods

- In term of sampling, LHS is still considered as the sampling of choice for many problem as it informs where most of the uncertainty has an effect
- **Sometimes, LHS may not be the method of choice**: when estimated probabilities are pretty low, when a particular area of interest needs to be oversampled …
- Other techniques such as **importance sampling and optimizations** are then more appropriate and efficient
- Such methods have been implemented and can be used by downloading the freeware package DAKOTA developed by Sandia.

Importance sampling with same sample size is a lot more accurate for low probabilities events



Classical LHS missed low probability unless sample size is increased

Results from xLPR project for US-NRC

Conclusions

- 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.

References

1. **Helton JC, Breeding RJ.** Calculation of Reactor Accident Safety Goals. *Reliability Engineering and System Safety* 1993;39(2):129-158.

2. **Helton JC, Anderson DR, Jow H-N, Marietta MG, Basabilvazo G.** Performance Assessment in Support of the 1996 Compliance Certification Application for the Waste Isolation Pilot Plant. *Risk Analysis* 1999;19(5):959 - 986.

3. **Helton JC, Hansen CW, Sallaberry CJ**. Yucca Mountain 2008 Performance Assessment: Conceptual Structure and Computational Implementation. In. *Proceedings of the International High-Level Radioactive Waste Management Conference, September 7-11, 2008*: American Nuclear Society, 2008:524-532.

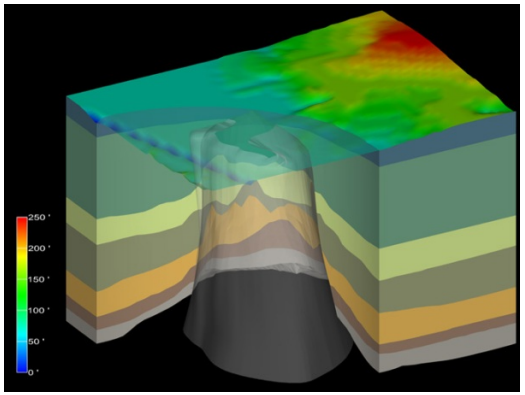
4. **Sallaberry CJ, Aragon A, Bier A, Chen Y, Groves JW, Hansen CW, Helton JC, Mehta S, Miller SP, Min J, Vo P.** Yucca Mountain 2008 Performance Assessment: Uncertainty and Sensitivity Analysis for Physical Processes. In. *Proceedings of the 2008 International High-Level Radioactive Waste Management Conference, September 7-11, 2008*: American Nuclear Society, 2008:559-566.

5. **Hansen CW, Brooks K, Groves JW, Helton JC, Lee PL, Sallaberry CJ, Stathum W, Thom C.** Yucca Mountain 2008 Performance Assessment: Uncertainty and Sensitivity Analysis for Expected Dose. In. *Proceedings of the 2008 International High-Level Radioactive Waste Management Conference, September 7-11, 2008*: American Nuclear Society, 2008:567-574.

SAND2013-xxx

Exceptional service in the national interest

Sandia National Laboratories



Propagating heterogeneity and uncertainty evaluations

from site investigation measurement to safety assessment

IAEA Workshop on Spatial variability – July 1-4, 2013

Cédric J. Sallaberry

U.S. DEPARTMENT OF ENERGY | NNSA 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.

Outline

- Introduction
- Sampling methods
- Spatial variability and sampling
- Reduction of computing cost.
- Conclusions

Introduction

- **Propagation of uncertainty refers to the technique used to transfer the uncertainty defined on the system (uncertainty in input data, parameters uncertainty, model uncertainty) onto the output of interest.**

- **While the part is usually pretty simple in concept, it is the most computationally demanding as it requires to run the code or set of codes a large number of times.**

- **The technique and associated parameters have to be chosen so that it minimizes the cost of such procedure without reducing the quantitative quality of the results.**

Techniques of uncertainty propagation

Sandia National Laboratories

- The technique used to propagate uncertainty depends on the method chosen to treat uncertainty.
- With deterministic adjoint (gradient based) methods, the derivative is calculated at the same time than the solution.
- With local sensitivities, a reference point in the input hyperspace needs to be selected. Once done, the derivatives in the direction of interest are calculated by varying parameters values (usually one at a time) by a pre-decided amount.
- Sampling-based methods used to propagate uncertainty represented as probabilities.

Sampling based methods - concept

- Generate sample: $\boldsymbol{x}_k$, $k = 1,2,\ldots,nS$
- Evaluate $\boldsymbol{y}$: $\boldsymbol{y}(\boldsymbol{x}_k)$, $k = 1,2,\ldots,$nS
- Resultant mapping: $[\boldsymbol{x}_k, \boldsymbol{y}_k(\boldsymbol{x}_k)]$, $k = 1,2,\ldots nS$
- Mapping forms basis for
  - Uncertainty analysis (distribution functions, box plots…)
  - Sensitivity analysis (scatterplots, regression analysis,…)

Components of Sampling-Based Approach

- Characterization of uncertainty in $\boldsymbol{x}$ (i.e., definition of $D_1, D_2, \ldots, D_{nX}$)
- Generation of sample from $\boldsymbol{x}$ (i.e., generation of $\boldsymbol{x}_k$, $k = 1,2,\ldots,nS$, in consistency with $D_1, D_2, \ldots, D_{nX}$)

Focus of this presentation

- Propagation of sample through analysis (i.e., generation of mapping $[\boldsymbol{x}_k, \boldsymbol{y}(\boldsymbol{x}_k)]$, $k = 1,2,\ldots,nS$)
- Presentation of uncertainty analysis results (i.e., approximations to the distributions of the elements of $\boldsymbol{y}$ obtained from $\boldsymbol{y}(\boldsymbol{x}_k)$, $k = 1,2,\ldots,nS$)
- Determination of sensitivity analysis results (i.e., exploration of the mapping $[\boldsymbol{x}_k, \boldsymbol{y}(\boldsymbol{x}_k)]$, $k = 1,2,\ldots,nS$)

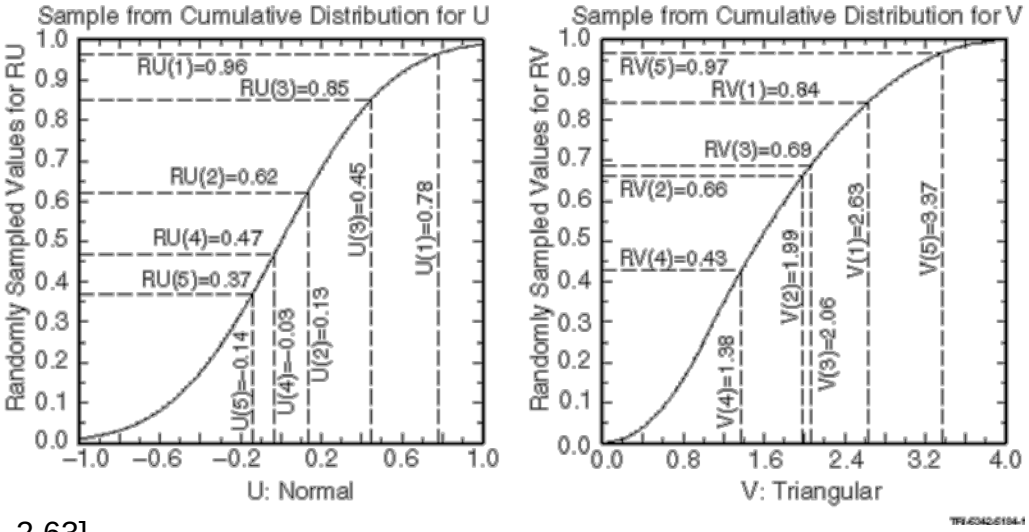
Different sampling techniques

Sandia National Laboratories

- Random sampling
- Importance (stratified) sampling
- Latin hypercube sampling
- Quasi-MC sampling

Simple Random Sampling

- Random sample: $\boldsymbol{x}_k = [x_{1k}, x_{2k}, \ldots, x_{nX,k}]$, $k = 1,2,\ldots,nR$
- Sample elements (i.e., $\boldsymbol{x}_k$'s) from different regions of sample space occur in direct relationship to the probability of these regions
- Each sample element selected independently of all other sample elements



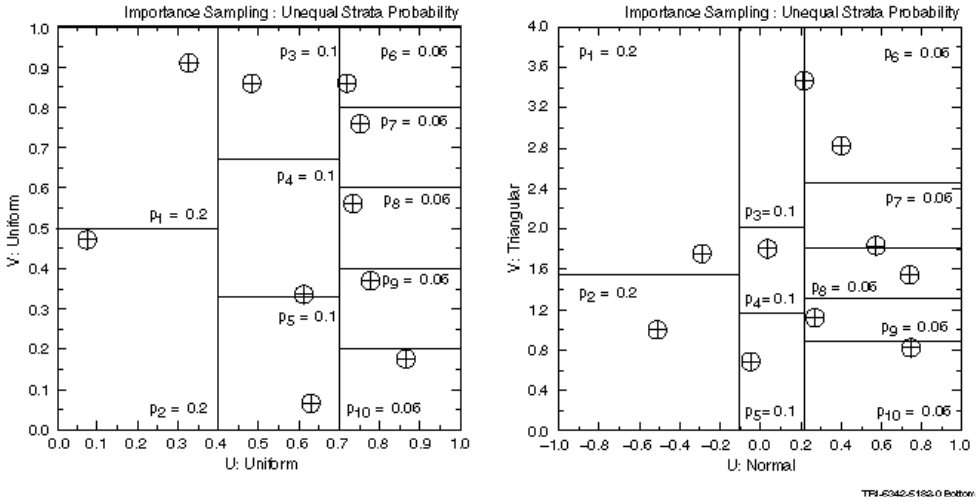
RU(1),*RV*(1): $\mathbf{x}_1 = [0.7, 2.63]$
RU(2),*RV*(2): $\mathbf{x}_2 = [0.13, 1.99]$
...

Note: *See W.H. Press et al. Numerical Recipes, for generation of uniform random samples from [0,1]*

Importance (Stratified) Sampling

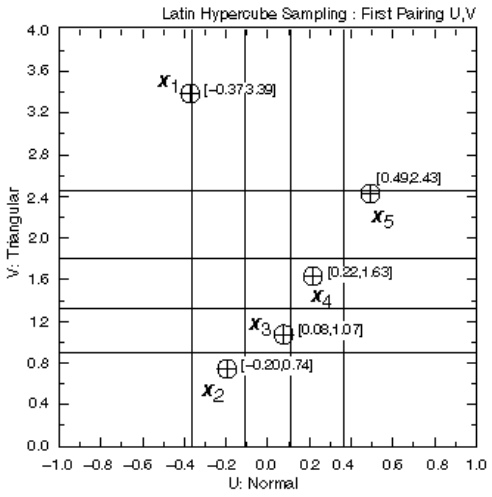
- Importance sample: $\boldsymbol{x}_k = [x_{1k}, x_{2k}, \ldots, x_{nX,k}]$, $k = 1,2,\ldots,nS$

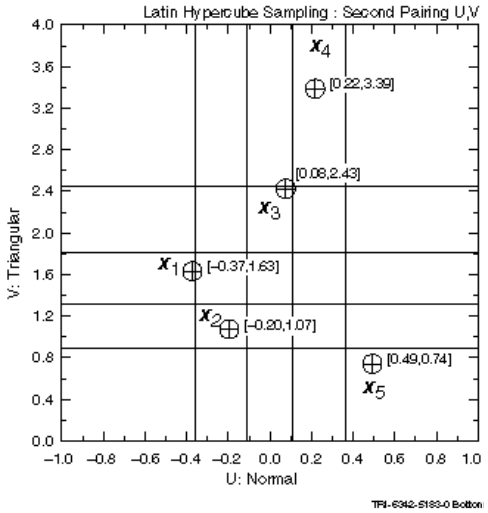
- Sample space divided into strata $S_1, S_2, \ldots, S_{nS}$ which
  - Typically have unequal probabilities
  - Assure inclusion of specific regions of sample space in analysis
- Sample element $\boldsymbol{x}_k$ randomly sampled from strata S_k



Latin Hypercube Sampling

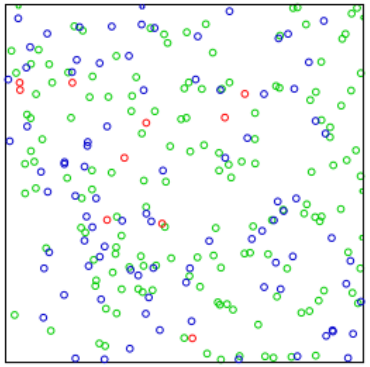
- Latin hypercube sample (LHS): $\boldsymbol{x}_k = [x_{1k}, x_{2k}, \ldots, x_{nX,k}]$, $k = 1,2,\ldots,nLHS$
- Generation of sample
  - Range of each x divided in *nLHS* intervals of equal probability
  - Value for x_j (i.e., x_{jk}) randomly selected from each interval
  - Values for x_1 randomly paired without replacement with values for x_2 to produce *nLHS* pairs
  - Preceding pairs randomly combined without replacement with values for x_3 to produce *nLHS* triples
  - Process continues through all variables to produce *nLHS* sample elements



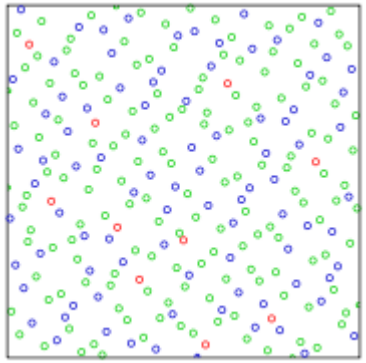


Quasi-Monte Carlo sequences

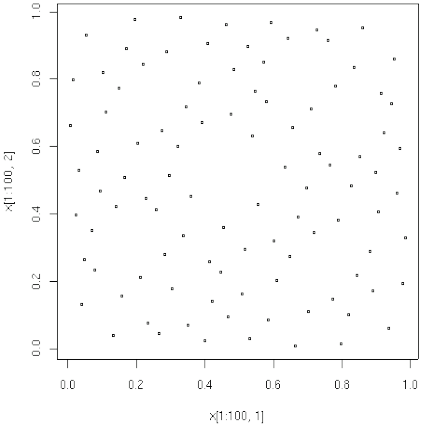
- Quasi random methods do not sample randomly the values using pseudo-random generator but always divide the hyperspace according to a deterministic rule.
- These methods allow a better multidimensional coverage of the input hyperspace but usually introduce a bias



Random sampling



Halton sequences



Sobol sequences

Picture source: Wikipedia

Sample Size

Sandia National Laboratories

- Always consider uncertainty within the context of the analysis
- A complex system is as good as its weakest part.
- The sample size will reflect the accuracy of the Monte Carlo technique. It **does not** reduce the uncertainty, only gives better estimate for the uncertainty and sensitivity analysis.
- The sample size has to be big enough so that the accuracy of the solution is not limited by the sample size (accuracy is as good as or close to the numerical accuracy of the model).
- However it should not be so large that a simulation becomes too costly without any gain (accuracy of Monte Carlo technique greater than numerical accuracy).
- In our experience, a sample size of a few hundreds to a few thousands gives satisfying results.

Spatial variability sampling

- Spatial variability can also be sampled using the sampling approach.
- However, it is **NOT** appropriate to sample an upscale quantity value from the spatial variability distribution and use this value in the whole domain.
- For each realization, it is however appropriate to sample a map with different values. The collection of maps represents then the uncertainty on the pattern, given a spatial variability distribution.
- An uncorrelated sampling of the values is usually not appropriate as there may be pattern (instead of pure heterogeneity) that may be captured using spatial correlation.
- This approach is different than Kriging as multiple maps are created and less assumptions are made.

Epistemic uncertainty on spatial variability

Sandia National Laboratories

- It is also possible (and likely) that there is uncertainty in the definition of the spatial variability distribution. The parameters defining this distribution may be uncertain
- As a result, for each realization, the parameters have to be sampled to define the distribution from which spatial values will be sampled.
- In a nested loop, it is possible to sample the distribution parameters in the outer loop and generate different spatially variable maps using the inner loop. However, spatial variability is NOT aleatory uncertainty, as aleatory uncertainty represent randomness in the future.

Reducing computer cost while handling spatial variability

Sandia National Laboratories

- One of the major problem of handling spatial variability and uncertainty is the computer cost in term of machine time and data size. Several techniques have been used in order to manage the cost of including spatial variability in complex systems, including (but not limited to):
  - Parallelism
  - Use of upscale quantity
  - Dimension reduction
  - Representation of main features.

Parallel computing – large scale models

Sandia National Laboratories

- While at early stage of development it is possible to run simplified PA on a single machine, their evolution require the addition of more complex and coupled physics and most always end up being too large to be run that way. Moreover, the sampling based method require to run hundreds to thousands of realizations in order to correctly capture the effect of uncertainty.

- Two types of parallelism have to be considered:
  - Parallelism of realizations when using sampling-based procedure. This is the simplest approach as each realization can be run independently and sent to a different processor. It is efficient as long as the number of processors used is lower than the sample size.
  - Parallelism within a realization. It is usually more complex as it requires domain decomposition. However, many finite element codes are developed nowadays with parallelism in mind and can be run on supercomputer.

Propagation of upscale quantity vs. 2D or 3D fine representation

Sandia National Laboratories

- The primary reasons to use upscale values in simplified models are:
  - **The reduction of computational cost**. As in most complex systems, radioactive waste repository modelisation involves a lot of uncertainty requiring to run multiple realizations (even in deterministic mode when “representative scenarios” are defined).
  - **The simplification of the problem.** Each single realization is composed of many models with different scale and capturing different feature of the problem. Multi-physics coupling is often necessary and can sometimes only be achieved with reduced size models

- However, it does NOT mean that more complex representation is not necessary. These simplified models need to be calibrated using more complex 3D representations.

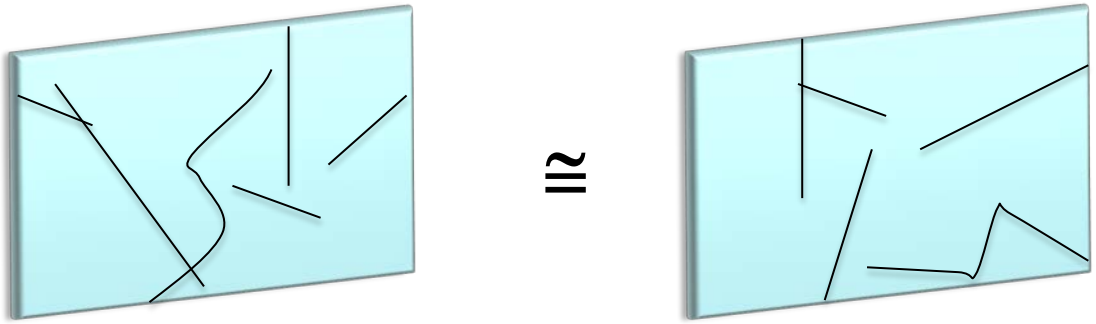
Model simplification via dimension reduction

Sandia National Laboratories

- Reducing the dimensionality of the problem is a way to reduce efficiently the cost of each computer simulation.
- When main pathways can be defined and represented, it is often possible to reduce high spatial dimension (3D) problems into lower dimensions (multi-2D, 2D or multi 1D)
- 3D calculations are however required to validate the lower dimensions models before using them intensively in the sampling problem.

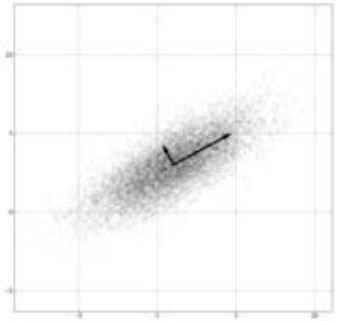
Main features vs. complete representation

- The exact representation of spatial variability is almost always impossible.
- However, this representation may not be necessary as long as the Main features are preserved
- For instance, in a fracture media, if the fracture distribution size and the density of fracture is represented correctly, it may not matters where the fractures actually are, as the transport solution will end up being pretty close.



Principal component

- It may be sometimes hard to define the main features of a system. In such case, **principal component analysis** can be used to help.
- Principal component analysis consists in an orthogonal transformation of a set of variable into independent variables, ordered by importance of variance (i.e. the first term having the highest variance and so on).



Picture source: Wikipedia

Conclusions

- **Sampling-based methods has been the method chosen by Sandia for most of his complex system analyses, including the two radioactive waste repository Performance Assessments (WIPP and YMP).**
- **The complexity and sophistication of the models developed increase with the computer power available. As a result, it is and will still be necessary to consider abstraction and high performance (parallelism) development in order to develop the best PA.**
- **There is NO perfect technique as it depends on the problem of interest. However, these complex codes need often to be run again to take care of bugs or update some features, so the initial choice can be revisited during a P.A. if necessary.**

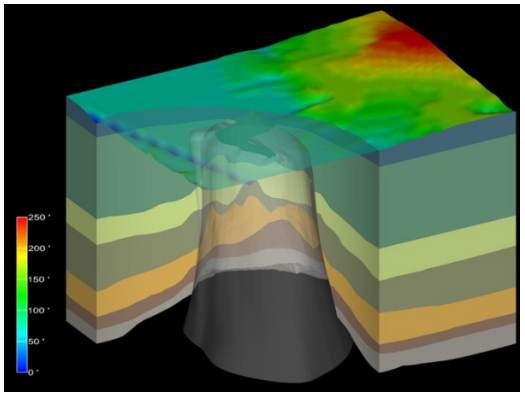
References

- J. C. Helton, F. Davis *Latin Hypercube Sampling and the propagation of uncertainty*
- A. Saltelli, S. Chan (Ed.) Sensitivity analysis
- Jackson, J.E. (1991). *A User's Guide to Principal Components* (Wiley).

SAND2013-xxx

Exceptional service in the national interest

Sandia National Laboratories



Spatial upscaling for modeling

IAEA Workshop on Spatial variability – July 1-4, 2013

Cédric J. Sallaberry

U.S. DEPARTMENT OF ENERGY | NNSA 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.

Outline

- Upscaling purpose
- Upscaling within the context of complex systems
- Difference between spatial variability and uncertainty over upscale quantity
- Upscaling of flow parameters
- Upscaling of transport parameters
- Conclusions

Purpose of upscaling

Sandia National Laboratories

- Upscaling tranforms a solution defined on a fine grid onto a coarser grid because of:
  - The scale difference of models in complex systems. In a repository modelization, a finer grid has to be used at the source, while coarser grid is necessary due to the large area covered to estimate the consequences.
  - The observation or information collected to characterize the site is at micro-scale level, while the model is used at a macro-scale level.
  - Upscaling is also used to reduce the models size when the complex system becomes too big to be run

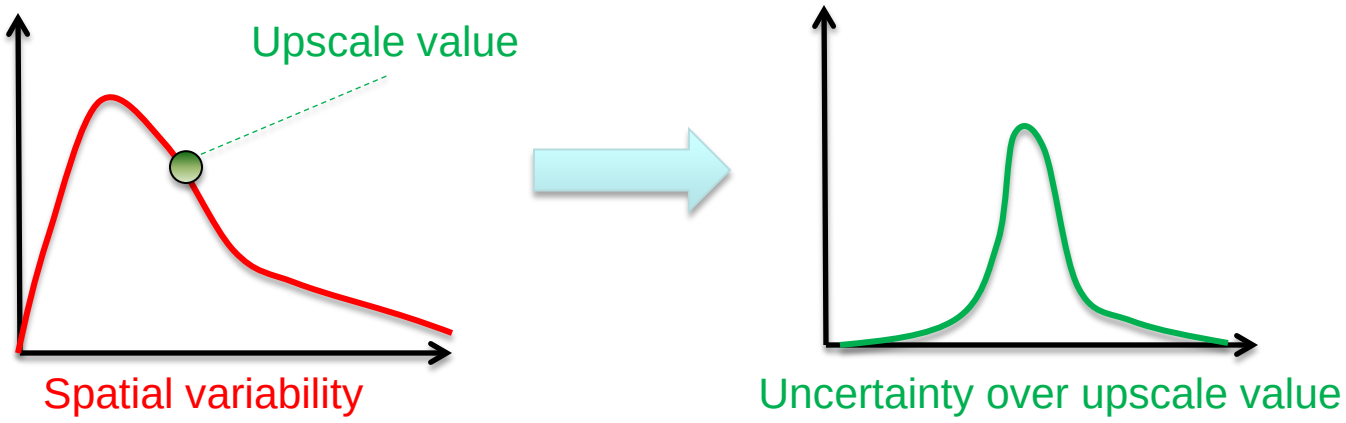
Homogeneity of accuracy in complex systems

Sandia National Laboratories

- When developing models for a complex system, it is important to consider the influence of each upscaling with respect to the overall characterization
- Small scale models are important when microscopic (relative to the overall model) processes need to be modelized.
- The flow of information passed from one model to the another

Difference between spatial variability and uncertainty over an upscale value

- When spatial variability is represented (at a certain scale) with a probability distribution, the upscale quantity represents a statistics (arithmetic mean, geometric mean, median …) of this distribution.
- The uncertainty over this upscale quantity is the **standard error** of the related statistic and not the **standard deviation** of the distribution. Usually it is a lot smaller than the standard deviation.



Upscaling of flow parameters

- Hydraulic conductivity is not an additive property. The value of conductivity on a coarser grid is between the harmonic and arithmetic means.
- As a result, the concept of block conductivity is defined. This does not represent the average of conductivities but the integral that would lead to the average behavior.
- Several techniques

Upscaling conductivity local techniques

Sandia National Laboratories

- Local techniques works only with the intrinsic information within a block (as for material property). They include:
  - Average
  - Power average (where a parameter *p*) is used to vary between harmonic and arithmetic mean
  - Renormalization (fast step procedure that may introduce errors)
  - Stream-tube (appropriate for sand-shale formation – create unidirectional flow)
  - Flow anisotropy (assuming that the flow direction is not known a priori)

Upscaling conductivity non local techniques

Sandia National Laboratories

- Non-local techniques consider that the conductivity in a block is not only dependent on intrinsic information but also on the flow condition (so the boundaries of the block). They include:
  - **Simple laplacian** (looking at first order derivative in each direction)
  - **Simple laplacian extended** (solving flow problem at each block)
  - **Laplacian using flow solution at the measurement scale over the entire aquifer** (better definition of boundary conditions)
  - **Laplacian with skin** (reduces the calculation of boundary conditions to a neighborhood around the block)
  - **Laplacian with periodic boundary conditions** (force the tensor to be symmetric –positive definite)
  - **non-parallel flow** (allowing to consider more complex flow geometry)
  - **Analytical solution** (developing analytical solution to use in each block)
  - **Method of moments** (matching spatial moments instead of block conductivity)
  - **Energy dissipation** (used to define the block conductivity tensors)

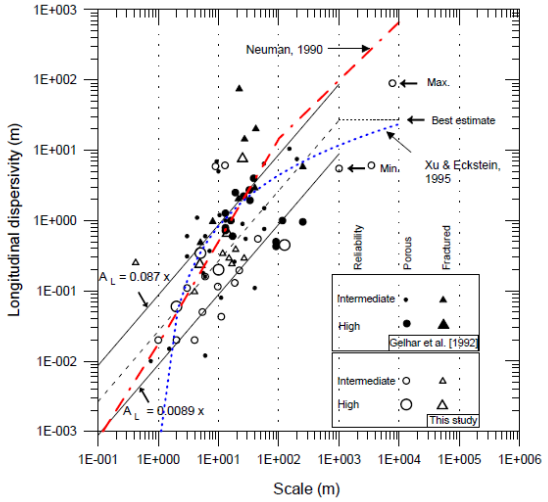
Block geometry

- While previously presented techniques have been developed for classical rectangular grids, other geometry can be more efficient:
  - Elastic grid starts as uniform grid but changes its shape to minimize the heterogeneity of cell conductivity
  - Potential stream-function space discretization
  - Direct block conductivity generation (with no original grid defined) such as scalar blocks, training images or regularization.

Upscaling transport parameters Dispersivity

- Some parameters may also have a different mean at different scale.
- For instance, dispersivity captures essentially two mechanisms, one that is microscopic (molecular dispersion) and the other macroscopic (mechanical dispersion).

- As a result, at larger scale the microscopic effect plays a smaller role. The change in the mean value reflects the transition from a scale where both effects are important to a scale when only one is predominant.



Source: www.sckcen.be/en/content/download/11912/163702/.../Dispersion.pdf

Upscaling of transport parameters techniques

Sandia National Laboratories

- The stochastic method is a popular method easy to implement but limited to cases with low permeability variability and requires a large amount of data
- The fractal method does not require the assumption of low permeability but still needs the same information than the stochastic method
- The self-consistent approach is appropriate in case of highly heterogeneous media, but can lead to large error bias in case of low permeability.

Conclusions

- In complex system analysis, upscaling will almost always be necessary.
- spatial variability is not equivalent to uncertainty over an upscale quantity
- Depending of the nature of the parameter of interest, upscaling procedure (or selected statistic) will differ. Litterature search is necessary to find the most appropriate approach
- The upscaling procedure and value depends also of the scale of origin and final scale of interest.

References

- J- Rodrigo-Ilarri, J. Jaime Gomez-Hernandez, B. Iooss, E. Plischke and K-J Rohlig. *Evaluation and testing of approaches to treat spatial variability in P.A.* – PAMINA deliverable #:D.2.2.D.1 (2008) http://www.ip-pamina.eu/downloads/pamina2.2.d.1.pdf

- Mallants, D. Marivoet, J. and Volckaert, G. 1998. *Review of recent literature on the dispersivity parameter for saturated and fractured porous media.* Technical note 44, Dept. W&D, SCK.CEN. Mol, Belgium