Sampled Based Methods for Sensitivity + Uncertainty Analyses of Model Output

Outline

- Setting the stage
- Screening methods
- Variance-based methods
- What else?
- Conclusions / Tools

Michael Flechsig
Scientific Data Management Group
of Department Data & Computation
at Potsdam Institute for Climate Impact Research

PIK Potsdam
Telegrafenberg
14473 Potsdam Germany
http://www.pik-potsdam.de

Model Uncertainties due to

- The real system itself
- Incomplete knowledge about the real system
- Delimitation of the modelled system's part
- Model structure / model equations
- Discretization / numerical algorithms
- Machine dependent representation of numbers
- Model calibration / parameterisation / driving forces
- Model result interpretation / communication

Uncertainty + Sensitivity Analysis

- **Uncertainty Analysis UA** **(Janssen, RIVM, The Netherlands)**

  The study of the uncertain aspects of a model and of their influence on the (uncertainty of the) model output
- **Sensitivity Analysis SA** **(Saltelli, EU JRC, Ispra)**

  The study of how the uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model input
- UA more general than SA, but also more fuzzy
- Often UA applies SA techniques, especially, if we consider a model

  $Y = f(\mathbf{x}) \quad \mathbf{x} = (x_1, x_2, ..., x_k)$

  with f = state transition function / algorithm / computer code

  $\mathbf{x}$ = k input factors: parameters, initial / boundary values, triggers (for submodels, processes, driving forces)

  Y = model output

Ideal SA Method

- Cope with scale and shape of the input factors
  - Range of the factor variation and shape / parameters of the pdf
- Include multi-dimensional averaging
  - Global versus local methods
- Model independent (model free)
  - Cope with non-linear / non-additive, non-monotonic models
- Grouping of factors
  - Treat grouped factors as if they were single factors
- Cost efficient
  - Pay attention to computational costs **C**

- SA types
  - Local or global
  - Qualitative or quantitative

Steps in a Sensitivity Analysis

- Determine a goal
  - do not simply focus on model output but rather on a thesis to prove or disprove
- Determine input factors
- Select a SA method dependent on
  - the goal
  - (in)dependency of factors
  - computational costs C of the model
- Generate a sample
  - determine a marginal distribution function for each of the input factors or
  - scan the factorial space (deterministically)
- Perform the SA method / evaluate the model
- Analyse model output, draw conclusions

SA Objectives Settings (Saltelli et al., 2004)

- Factors prioritisation --> diagnostic modelling
  - Identify factor(s) that is/are most deserving of better experimental measurement to reduce the target output uncertainty the most.
- Factors mapping --> diagnostic modelling
  - Classify realizations of Y and afterwards of factors **x** into ”acceptable” and ”unacceptable” classes (“regions”)
- Factors fixing --> prognostic modelling
  - Identify this factor that can be fixed at any value over the corresponding range of uncertainty without significantly reducing the output variance
- Variance cutting --> prognostic modelling
  - Reduce output variance by simultaneously fixing the smallest number of factors

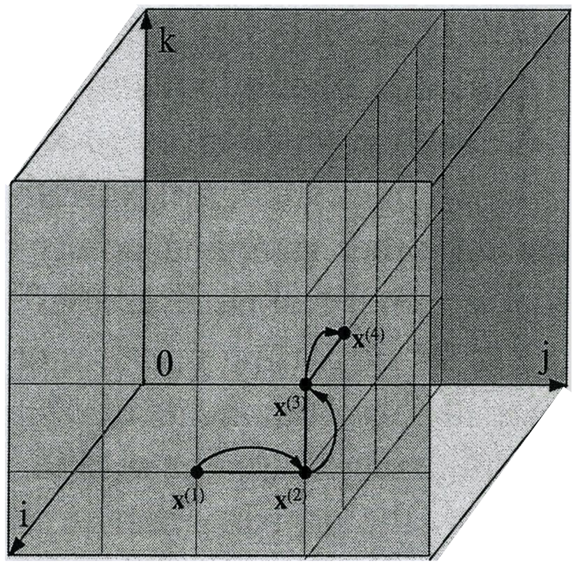
Screening Methods

- To come up with a short list of important factors
  - Principle of parsimony, Occam's razor
- A means of quality assurance
- Often one-at-a-time approaches w.r.t. factors
- Drawbacks
  - Tend to neglect factor interactions
  - Tend to provide qualitative measures: ranks instead of quantities

Morris' Design (1991)

model free

- Grid input space **x** with p levels for each factor and constant grid widths Δ_i
- Determine local elementary effect d_i of x_i
  from two grid points in **x** that differ
  only in one factor x_i by Δ_i:
  $d_i = (Y(x+e_i\Delta_i) - Y(x)) / \Delta_i$
- Select randomly T trajectories
  of length k (from k+1 points)
  where one elementary effect d_i
  can be derived
  from two consecutive points
  for each trajectory



Example: k=3, p=5, factors i,,j,k

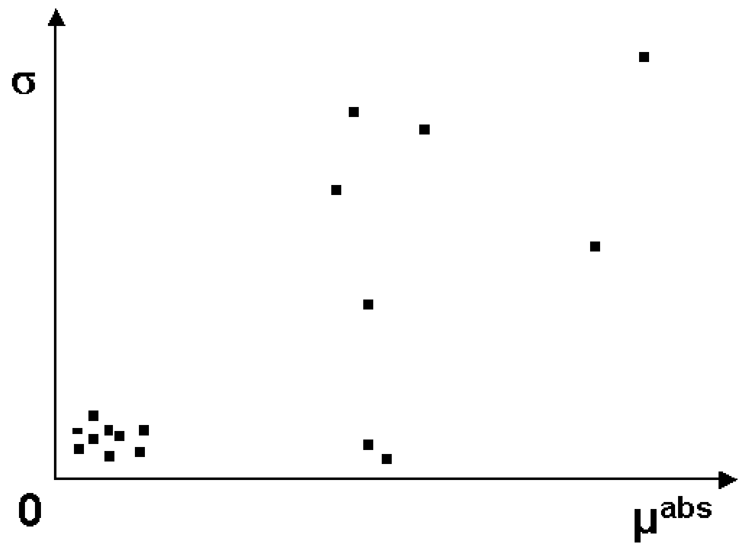
(from Saltelli *et al.*)

- Consider distributions $F_i = \{ d_i \}$ and compute σ_i = std. deviation of F_i
  $F_i^{abs} = \{ |d_i| \}$ and compute μ_i^{abs} = mean of F_i^{abs}

Morris' Design (ctd.)

Interpretation:

- high μ_i^{abs}
  x_i has important overall influence on Y
- high σ_i
  x_i is involved in interactions
  with other factors /
  effect of x_i is non-linear



- Computational costs (usually: p=4-6, T=10) **C = T*(k+1)**

Cotter's Design (1979)

model free

- Consider only low level 1 and upper level p for each factor
- Two runs with all factors for level 1 (X^L) and level p (X^U)
- Each factor x_i in turn at its upper level p while all the other factors remain at their low level 1: $X_i^L = (x_1^L ,..., x_{i-1}^L , x_i^U , x_{i+1}^L ,..., x_k^L)$
- Each factor X_i in turn at its low level p while all the other factors remain at their upper level p: $X_i^U = (x_1^U ,..., x_{i-1}^U , x_i^L , x_{i+1}^U ,..., x_k^U)$

- Consider $M_i = | (Y(X^U) – Y(X_i^U)) – (Y(X_i^L) – Y(X^L)) | +$ even order effects ...
  $| (Y(X^U) – Y(X_i^U)) + (Y(X_i^L) – Y(X^L)) |$ odd order effects ...
  involving factor x_i

  as a measure to estimate importance of factor x_i
- Drawback: important factors may remain undetected
- Computational costs **C= 2*(k+1)**

Group Screening Designs

- Andres' iterated fractional factorial design IFFD (1993)
  - Randomly assign factors to groups
  - Estimation of main, quadratic and two-factor interaction effects
  - Drawback: Model output should be determined by only a few highly influential factors
  - Computational costs (super-saturated design) **C < k**

- Bettonvil's sequential bifurcation (1990)
  - Assign factors to groups
  - Groups that do not contain important factors are eliminated while bifurcating (splitting) the others
  - Drawback: factors must have knows signs
  - Computational costs **C unpredictable**
    but factors with largest main effects are identified first

Variance Decomposition Methods

- Quantitative measures
- Assumption:
  - all what we want to know about the model is captured by its variance
- Relation to
  - design of experiments DOE (correlation ratios) and
  - ANOVA studies (variance decomposition)

Basic Approach

model free

- $V_X(Y)$ unconditional variance of Y
- $V_{-Xi}(Y|X_i=x_i^*)$ conditional variance of Y when fixing X_i to x_i^*
- $E_{Xi*}(V_{-Xi}(Y|X_i=x_i^*))$ averag. resid. variance of Y that can be obtained when fixing X_i
  abbr.:= $E(V(Y|X_i))$ the smaller this value the more influential the factor X_i
- $E(V(Y|X_i)) \leq V_X(Y)$
  $V_X(Y) = V(E(Y|X_i)) + E(V(Y|X_i))$
  = main effect + residual of X_i on Y
- $S_i = 1 - E(V(Y|X_i)) / V_X(Y) = V(E(Y|X_i)) / V_X(Y) \in [0,1]$ importance measure
  sensitivity index
- Solving the multi-dimensional integral correlation ratio
  by Monte Carlo method with sample size N: 1st order effect
  Computational costs for all k S_i **C= N*(k+1)**
- $\Sigma S_i \leq 1$ equality for additive models and orthogonal factors
  (without interaction effects)

Method of Sobol' (1990)

- Variance decomposition approach **for orthogonal factors**
- Decompose

  $Y(x_1, \ldots, x_k) = Y_o + \Sigma\, Y_i(x_i) + \Sigma\,\Sigma\, Y_{ij}(x_i,x_j) + \Sigma\,\Sigma\,\Sigma\, Y_{ijl}(x_i,x_j,x_l) + \ldots + Y_{123\ldots k}(x_1,\ldots,x_k)$

  with $1 \leq i < j < l < \ldots \leq k$ (in total 2^k terms)

  Y_0 constant

  $\{ Y_{i\ldots s} \}$ orthogonal system

- $Y_0 = \int Y(x)dx$

  $Y_{i\ldots s}(x_i,\ldots,x_s) = \int Y(x)dx_{-(i\ldots s)}$ - sum of all $Y_{i\ldots s}(x_i,\ldots,x_s)$ of lower order

- $V_X(Y) = \int Y^2(x)dx - Y_0^2$ total variance

  $V_{i\ldots s} = \int Y_{i\ldots s}^2(x_i,\ldots,x_s)dx_i\ldots dx_s$ partial variance

  and

  $V_X(Y) = \Sigma\, V_i + \Sigma\,\Sigma\, V_{ij} + \Sigma\,\Sigma\,\Sigma\, V_{ijl} + \ldots + V_{123\ldots k}$

  $1 = \Sigma\, S_i + \Sigma\,\Sigma\, S_{ij} + \Sigma\,\Sigma\,\Sigma\, S_{ijl} + \ldots + S_{123\ldots k}$

  with $1 \leq i < j < l < \ldots \leq k$

Method of Sobol' (ctd.)

- 1 $= \Sigma S_i + \Sigma\Sigma S_{ij} + \Sigma\Sigma\Sigma S_{ijl} + \ldots + S_{123\ldots k}$ **(*)**

- S_{ij} $= V(E(Y|X_iX_j) / V_X(Y) - V(E(Y|X_i)) / V_X(Y) - V(E(Y|X_j)) / V_X(Y)$
  $= S_{ij}^c - S_i - S_j$
  2nd order effect = closed effect − 1st order effects
  Computational costs for **(*)** **$C = N*2^k$**

- Saltelli, 2002, Saltelli *et al.*, 2004:
  S_{Ti} = sum of all terms in **(*)** that include factor X_i — total effect of factor X_i
  $= 1 - V(E(Y|X_{-i})) / V_X(Y) = E(V(Y|X_{-i})) / V_X(Y)$
  Computational costs for all k S_{Ti} **$C = N*(k+1)$**

- Computational costs **$C = N*(k+2)$** for

all k	first order effects	S_i
all k	total effects	S_{Ti}
all k*(k-1)/2	closed effects of order k-2	S_{-ij}^c

by applying re-sampling ("do not re-sample the factor that is to be estimated")

Fourier Amplitude Sensitivity Test FAST from Cukier *et al.* (1973, 1975, 1978)

- Model free
- Goal: determination of $E(Y)$, $V_X(Y)$ and S_i
- Idea: mapping k-dimensional integrals in **x** into 1-dimensional integrals in s
  by transformations $x_i = g_i(\sin\omega_i s)$ — $s \in (-\pi, \pi)$ scalar, ω_i fixed integer angular frequency

  $Y(X_1, \ldots, X_k) = f(X_1, \ldots, X_k) = f(g_1(\sin\omega_1 s), \ldots, g_k(\sin\omega_k s))$
  $= Y(s)$
- Then

  $E(Y) = \int f(s)\, ds / 2\pi$

  $V_x(Y) = \int f^2(s)\, ds / 2\pi - E^2(Y)$

  $\approx 2 \Sigma (a_j^2 + b_j^2)$ $(j=1, \ldots, \infty)$

  $S_i \approx 2 \Sigma (a_{p\omega i}^2 + b_{p\omega i}^2)$ $(p=1, \ldots, M)$ — $p\omega_i$ = higher harmonics of ω_i

  $a_j = \int f(s) \cos(js)\, ds / 2\pi$ — usually: M = 4 or 6

  $b_j = \int f(s) \sin(js)\, ds / 2\pi$ — a_i, b_i: Fourier coefficients
- Computational costs for all k S_i — **$C = 2*k*M*(\max(\omega_i) + 1)$**
- Additional tasks: select g_i and ω_i
  sample the original factor space **X** “sufficiently” for s

What else?

- Other SA techniques
  - Approach by linear regression model and S(R)RC's, P(R)CC's
  - Bayesian uncertainty estimation (e.g., GLUE)
  - Pure local SA measures based on $\partial Y/\partial X_i$ or its approximation
  - Monte Carlo filtering
  - Adjoint models - automated differentiation of model code
  - ...

- Sampling strategies / sample size N
  - Random sampling
  - Latin hypercube sampling (McKay)
  - Importance sampling (Helton)
  - LP_τ quasi random sequences (Sobol')

Conclusions / Tools

- Model-free techniques are essential for complex / non-linear models
- Handle the plethora of factors by
  qualitative methods followed by quantitative methods
- Techniques for diagnostic and prognostic modelling
- Many techniques scale for C linearly with the number of factors

- Related tools

SimLab	EU JRC, Ispra
Unicorn	Delft University of Techn.
UncSam	RIVM, The Netherlands
SimEnv	PIK Potsdam
CrystalBall	Decisioneering Inc.
@Risk	Palisade Corp.

Selected References

- Cukier, R.I., Fortuin, C.M., Shuler, K.E., Petschek, A.G., Schaibly, J.H.: Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory. Journal Chem. Phys., 59, 3873-3878, 1973
- Morris, M.D.: Factorial plans for preliminary computational experiments. Technometrics, 33(2), 161-174, 1991
- Saltelli, A., Chan, K., Scott, E.M. (eds.): Sensitivity analysis. John Wiley & Sons, 2000
- Saltelli, A., Tarantola, S., Camplongo, F., Ratto, M.: Sensitivity analysis in practice: A guide to assessing scientific models. John Wiley & Sons, 2004
- Sobol', I.M.:Sensitivity estimates for non-linear mathematical models. Math. Model. Comput. Exp., 1, 407-414, 1993