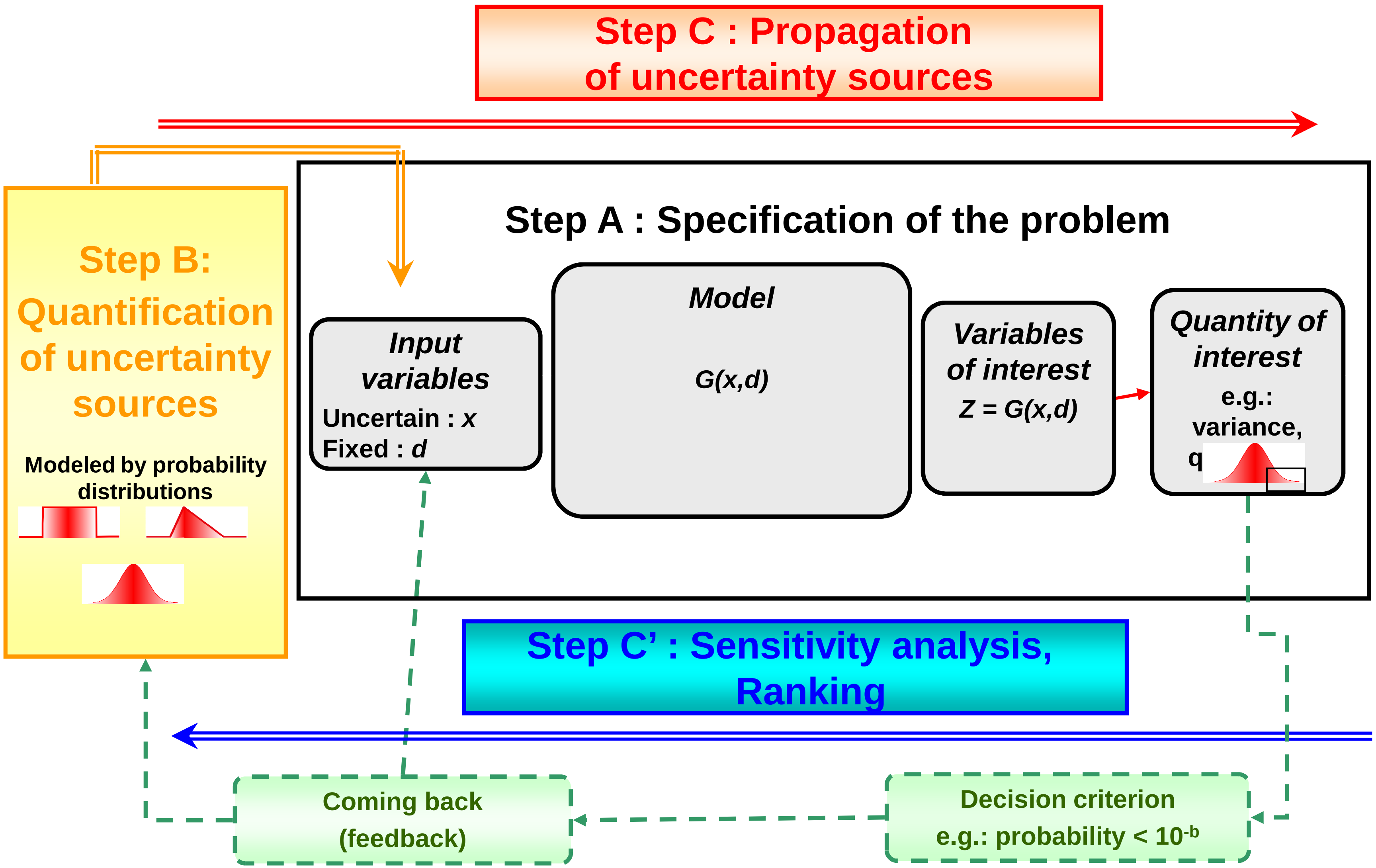
Sensitivity analysis of computer experiments

Fabrice Gamboa
Bertrand Iooss

10/03/2013

The “global methodology” of uncertainty management



Main objectives of sensitivity analysis

• **Reduction of the uncertainty of the model outputs by prioritization of the sources**

- Variables to be fixed in order to obtain the **largest reduction** (or a fixed reduction) **of the output uncertainty**

  A purely mathematical variable ordering

- Most influent variables in a given output domain

→ if reducibles, then R&D prioritization

→ else, modification of the system

• **Simplification of a model**

- **determination of the non-influent variables,** that can be fixed without consequences on the output uncertainty
- building a simplified model, a metamodel

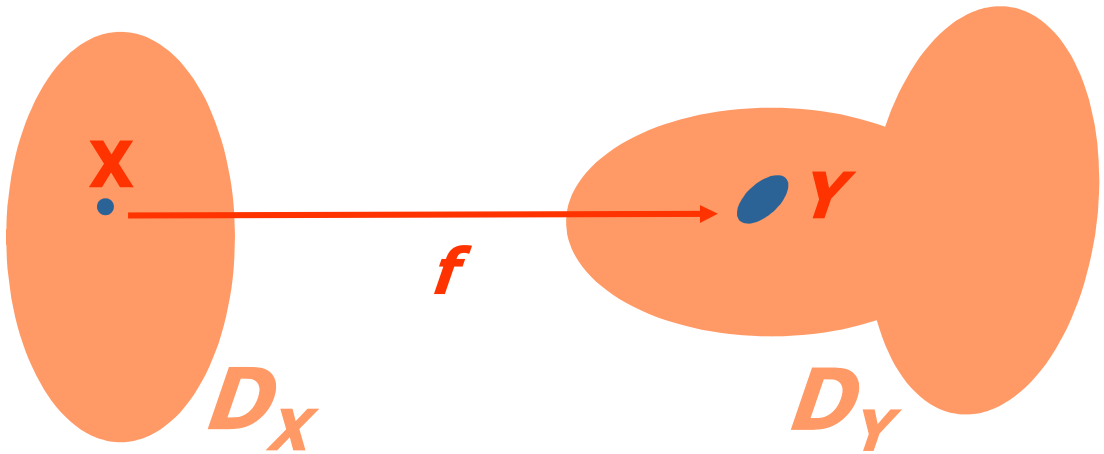
Sensitivity analysis notions

- **Sensitivity**, for example $\partial Y / \partial X_i$

Donne une idée de la manière dont peut répondre la réponse en fonction de **variations potentielles** des facteurs

- **Contribution = sensitivity x importance**, for example $\frac{\partial Y}{\partial X_i}\sigma(X_i)$

Permet de déterminer le **poids** d'une variable d'entrée (ou groupe de variables) sur l'incertitude de la variable d'intérêt (la sortie)

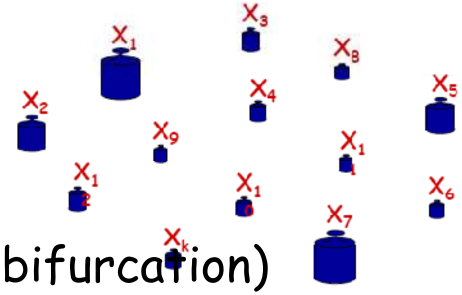


Overall classification of sensitivity analysis methods

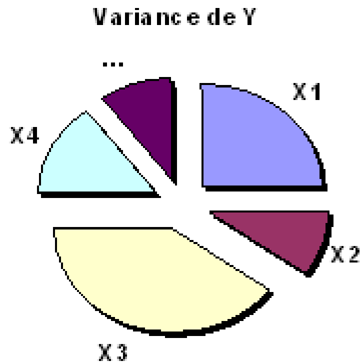
(quantity of interest = variability of the output)

Three types of answers:

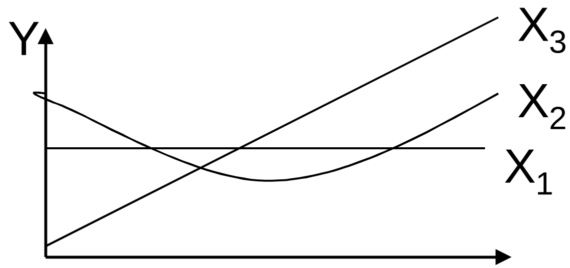
1. Screening :
   - classical design of experiments,
   - numerical design of experiments (Morris, sequential bifurcation)



2. Quantitative measures of global influence :
   - correlation/regression on values/ranks
   - statistical tests,
   - functional variance decomposition (Sobol),
   - other measures : entropy, distribution distances



3. Deep exploration of sensitivities
   - smoothing techniques (param./non parametric)
   - metamodels



Screening with n < p (supersaturated designs)

Many inputs (p >> 10) and cpu time costly computer code

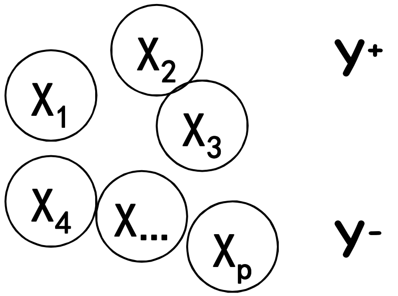
Objective: less computations than number of inputs

Hypotheses:

- Number of influent inputs << total number of inputs
- Monotony of the model, no interaction between inputs
- Knowledge of the direction of the output variation / each input

Example: method of sequential bifurcations

2 runs



Screening with n < p (supersaturated designs)

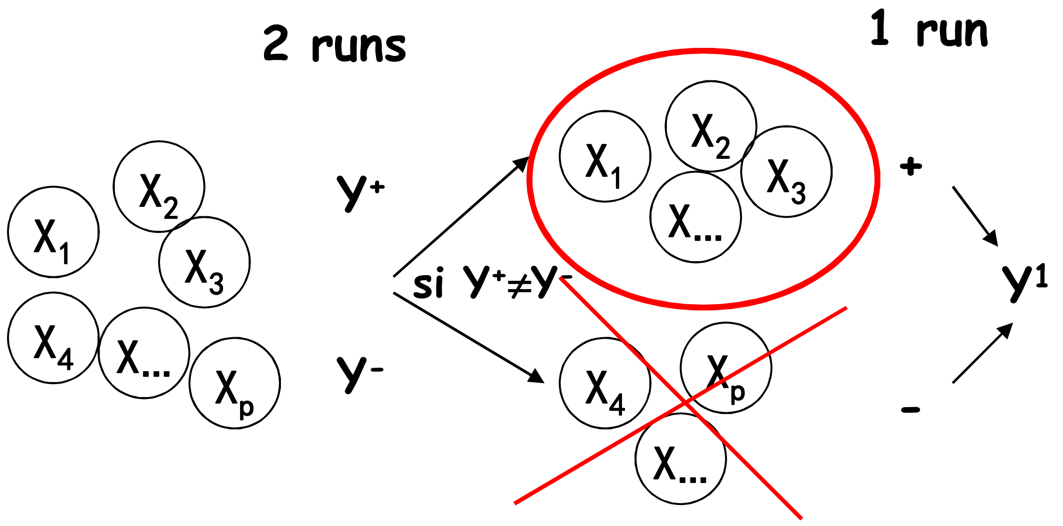
Many inputs (p >> 10) and cpu time costly computer code

Objective: less computations than number of inputs

Hypotheses:

- Number of influent inputs << total number of inputs
- Monotony of the model, no interaction between inputs
- Knowledge of the direction of the output variation / each input

Example: method of sequential bifurcations



Screening with n < p (supersaturated designs)

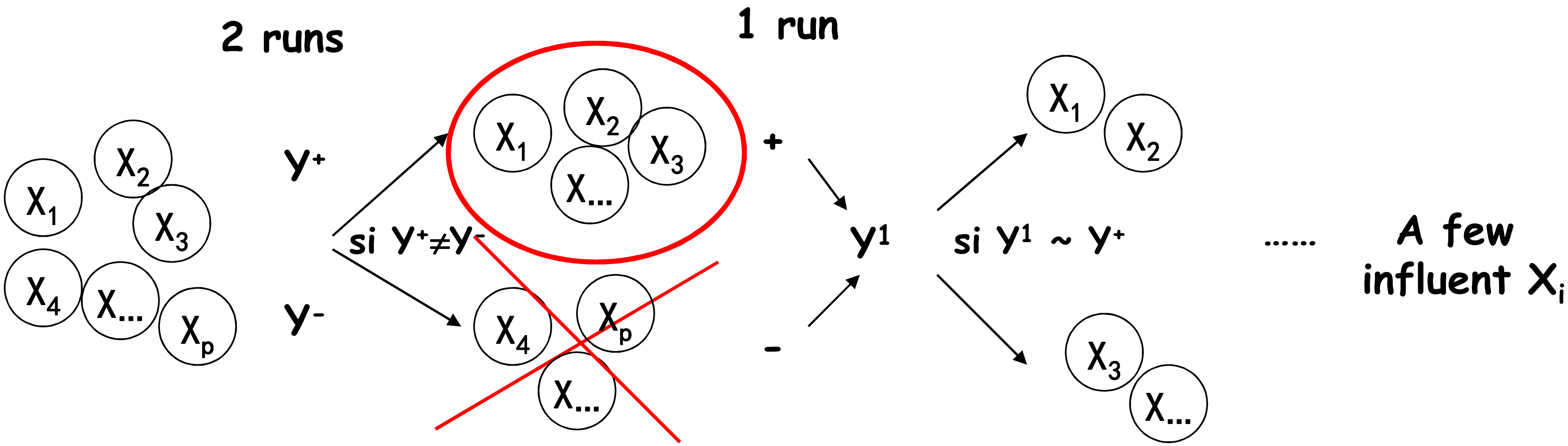
Many inputs (p >> 10) and cpu time costly computer code

<u>Objective</u>: less computations than number of inputs

<u>Hypotheses</u>:

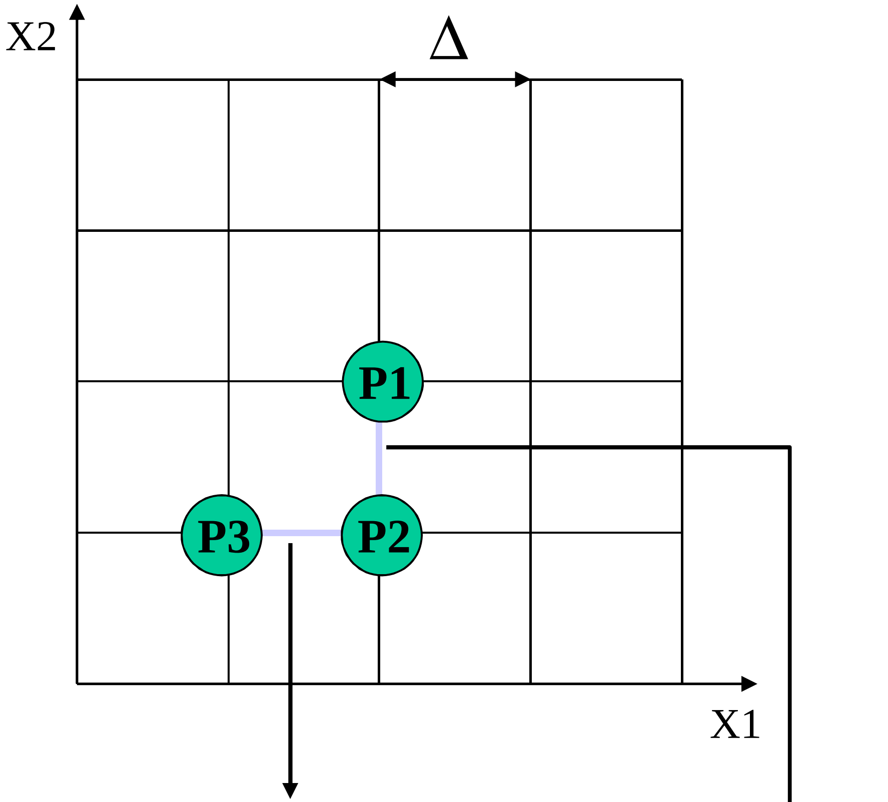
- Number of influent inputs << total number of inputs
- Monotony of the model, no interaction between inputs
- Knowledge of the direction of the output variation / each input

Example: method of sequential bifurcations



Number of runs: *n ~ p / 2*

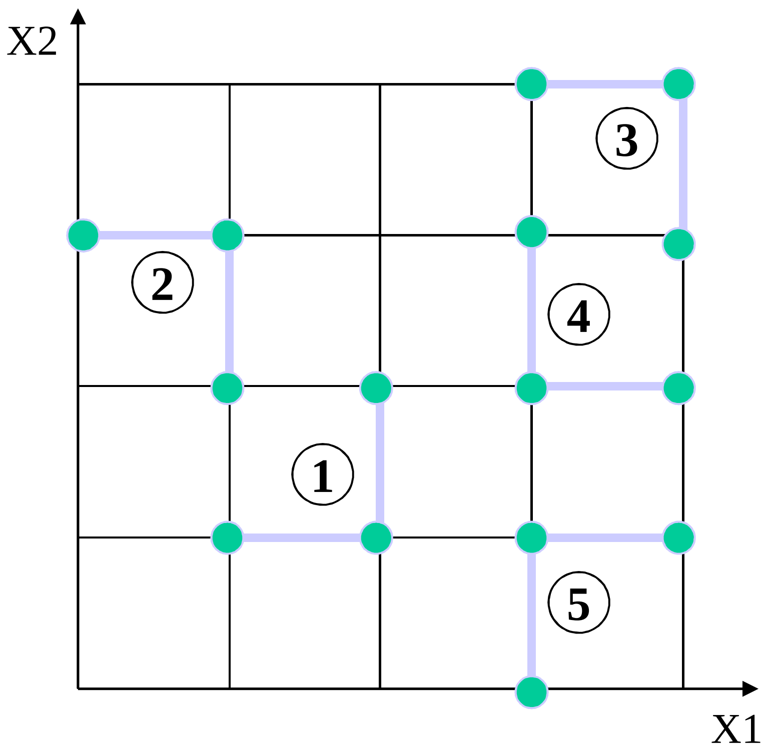
Screening without hypothesis on function: Morris' method



$$d_{X_1} = \frac{f(P3) - f(P2)}{\Delta} \qquad d_{X_2} = \frac{f(P2) - f(P1)}{\Delta}$$

- Discretization of input space
- Needs p+1 experiments
- OAT (One-at-A-Time)
- computation of one elementary effect for each input

Morris' method



- OAT design is repeated R times (total: n = R*(p+1) experiments)

- It gives an R-sample for each elementary effect

$$\left\{d^i_{X1}\right\}_{i=1\ldots R}$$

$$\left\{d^i_{X2}\right\}_{i=1\ldots R}$$

- Sensitivity measures:

$$\mu_i^* = \mathrm{E}(|d_{X_i}|)$$

$$\sigma_i = \sigma(d_{X_i})$$

Morris: sensitivity measures

- $\mu_i^* = \mathrm{E}(|d_{X_i}|)$ is a measure of the **sensitivity**:

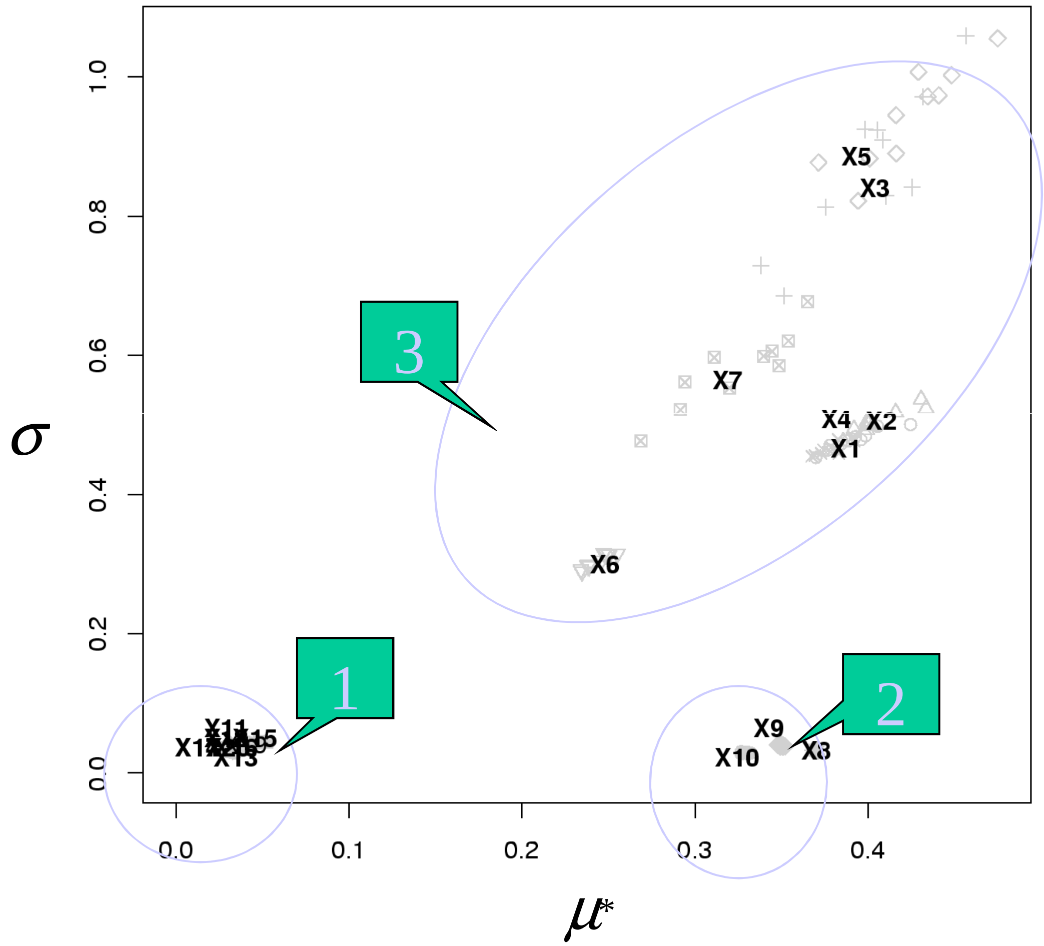
  Important value → important effects (in mean)
  → sensitive model to input variations

- $\sigma_i = \sigma(d_{X_i})$ is a measure of the **interactions** and of the **non linear effects**:

  important value → different effects in the R-sample
  → effects which depend on the value:
  - of the input X_i => non linear effect
  - or of the other inputs => interaction

  (the distinction between the two cases is impossible)

Morris : example



20 factors
210 simulations
→ Graph (mu*, sigma)

Distinction between 3 groups:

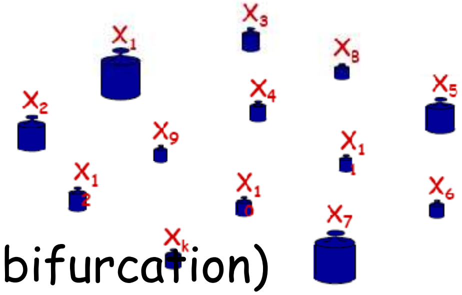
1. Negligible effects
2. Linear effects
3. Non linear effects and/or with interactions

Cas test : non monotonic function of Morris

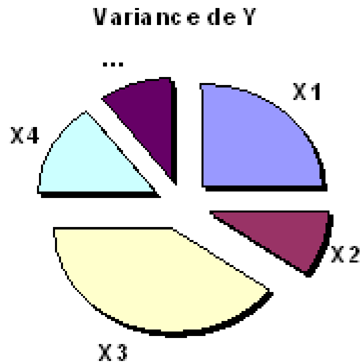
Overall classification of sensitivity analysis methods

(quantity of interest = variability of the output)

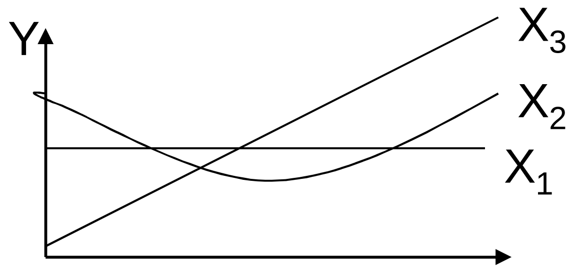
Three types of answers:



1. <u>Screening :</u>
   - classical design of experiments,
   - numerical design of experiments (Morris, sequential bifurcation)



2. <u>Quantitative measures of global influence :</u>
   - correlation/regression on values/ranks
   - statistical tests,
   - functional variance decomposition (Sobol),
   - other measures : entropy, distribution distances



3. <u>Deep exploration of sensitivities</u>
   - smoothing techniques (param./non parametric)
   - metamodels

Sensitivity analysis for one scalar output

Sample ($\mathbf{X} \in \Re^p$, $Y(\mathbf{X}) \in \Re$) of size $N > p$

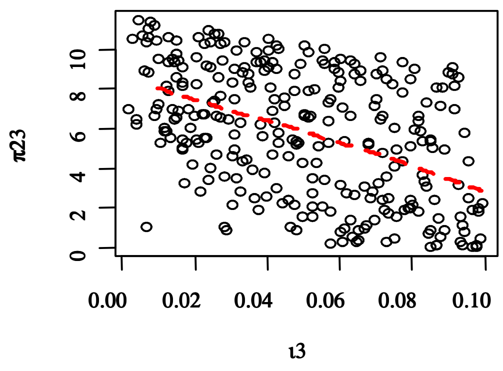
Preliminary step: graphical vizualisation (for ex: scatterplots)

Remark: *it can be a Monte Carlo sample, a quasi-Monte Carlo sample or any other designs*

Graphial representation : scatterplots

Measure the linear characher of the cloud

Example :
N = 300

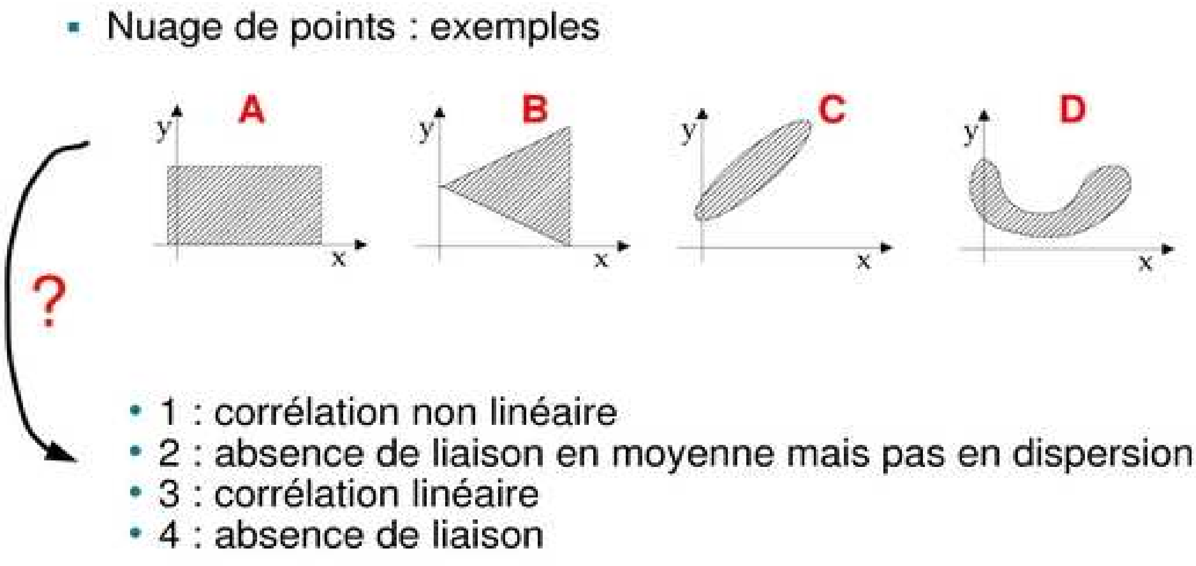


***N* runs**

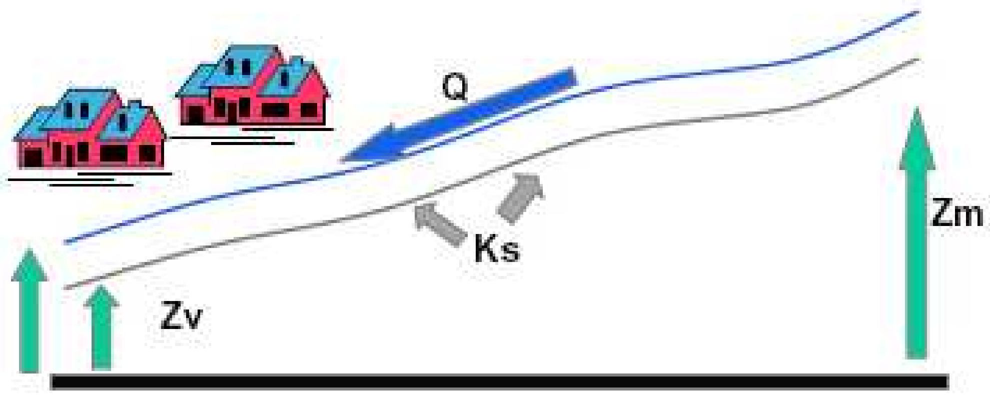
Graphs Output with respect to each input

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

$$\hat{\rho} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$



Flood model - Scatterplots – Output S



Q = river flowrate ~ Gumbel on [500,3000]

Ks = friction coefficient ~ normal on [15,50]

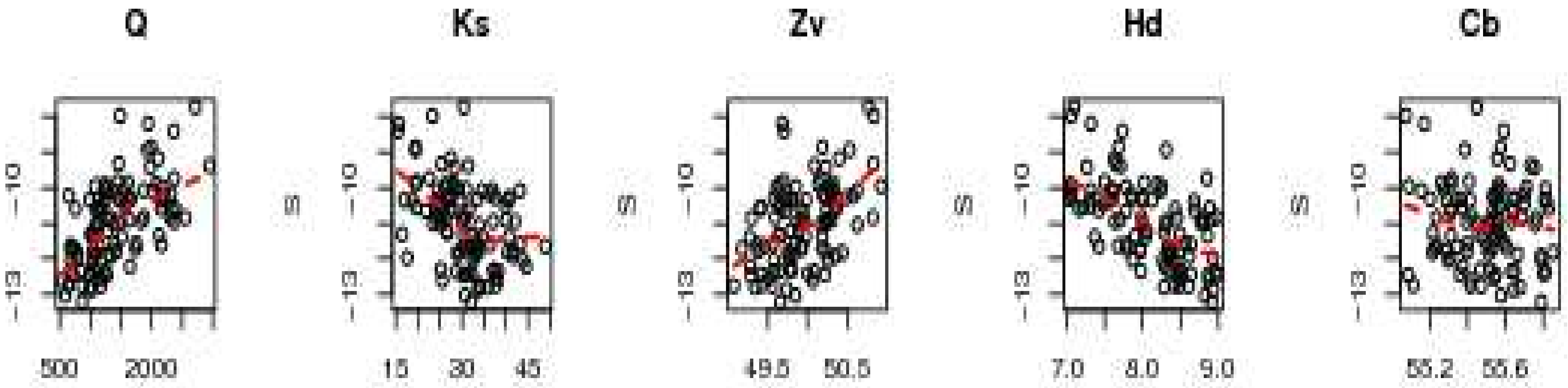
Zv = downstream river bed heigth ~ triangular on [49,51]

Hd = dyke heigth ~ triangular on [7,9]

Cb = bank heigth ~ triangular on [55,56]

$$S = Z_v + H - H_d - C_b \;\text{ avec }\; H = \left(\frac{Q}{BK_s\sqrt{\frac{Z_m - Z_v}{L}}} \right)^{0.6}$$

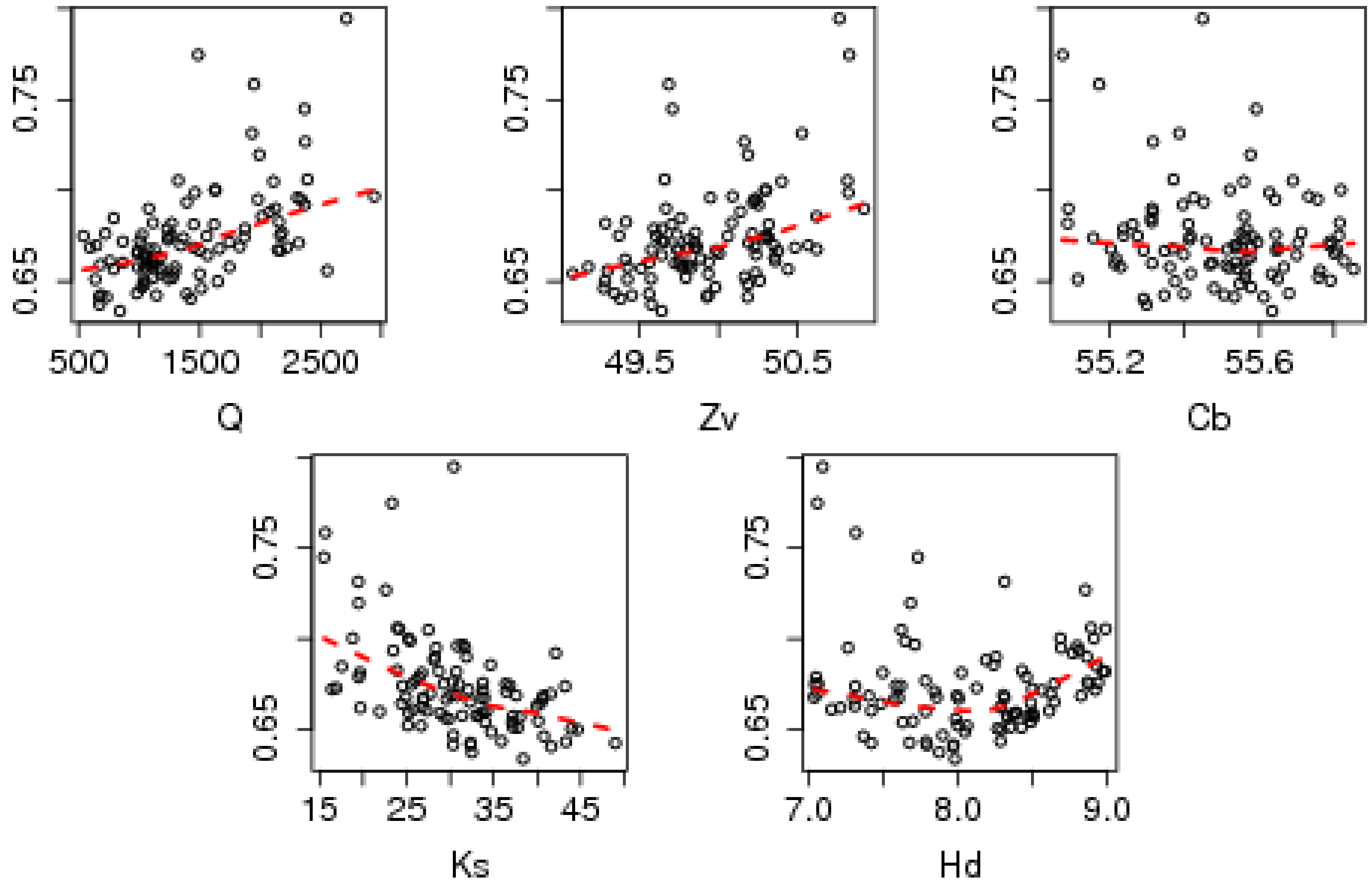
Monte Carlo sample – *N* = 100



Flood model - Scatterplots – Output Cp

$$C_p = 1\!\!1_{S>0} + \left\{0.2 + 0.8\left[1 - \exp\left(-\frac{1000}{S^4}\right)\right]\right\} 1\!\!1_{S\leq 0} + \frac{1}{20}\left(H_d 1\!\!1_{H_d>8} + 8 1\!\!1_{8\leq H_d}\right) ,$$

Monte Carlo sample – N = 100



<u>Major drawback</u>: only first order relations between inputs are analyzed and not their interactions (=> needs of other data anlysis tools)

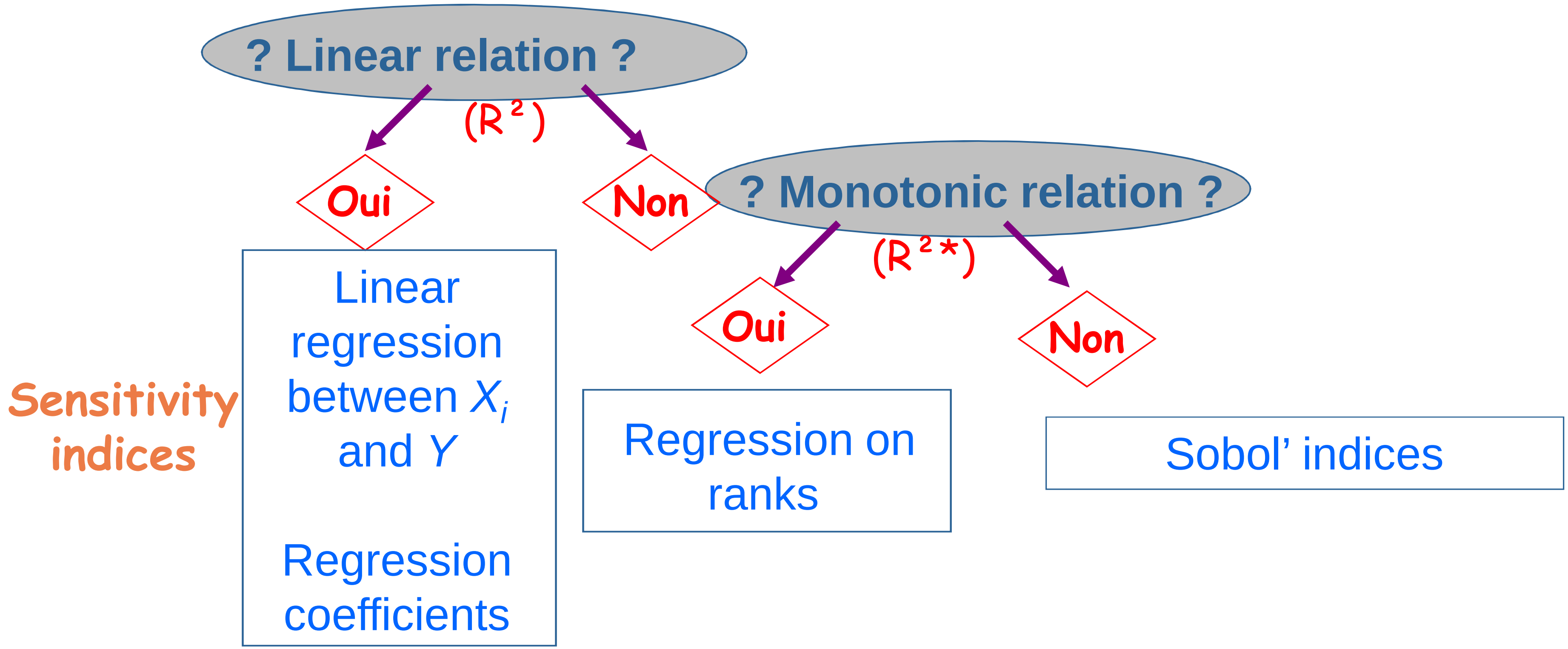
Sensitivity analysis for one scalar output

Sample ($\mathbf{X} \in \Re^p$, $Y(\mathbf{X}) \in \Re$) of size $N > p$

Preliminary step: graphical vizualisation (for ex: scatterplots)

Quantitative sensitivity analysis methodology

[Saltelli et al. 00, Helton et al. 06]



Sensitivity indices in case of linear inputs/output relation

Independent input variables $\mathbf{X}=(X_1,\ldots,X_p)$

Sample : n realizations of $(\mathbf{X},Y)$

$$Y=\beta_0+\sum_{i=1}^{p}\beta_i X_i$$

- SRC index: $$SRC(X_i):=\beta_i\sqrt{\frac{\mathrm{Var}(X_i)}{\mathrm{Var}(Y)}}$$

Sign of β_i gives the direction of variation of Y in fct of X_i

- SRC is similar to the linear correlation coefficient (Pearson)

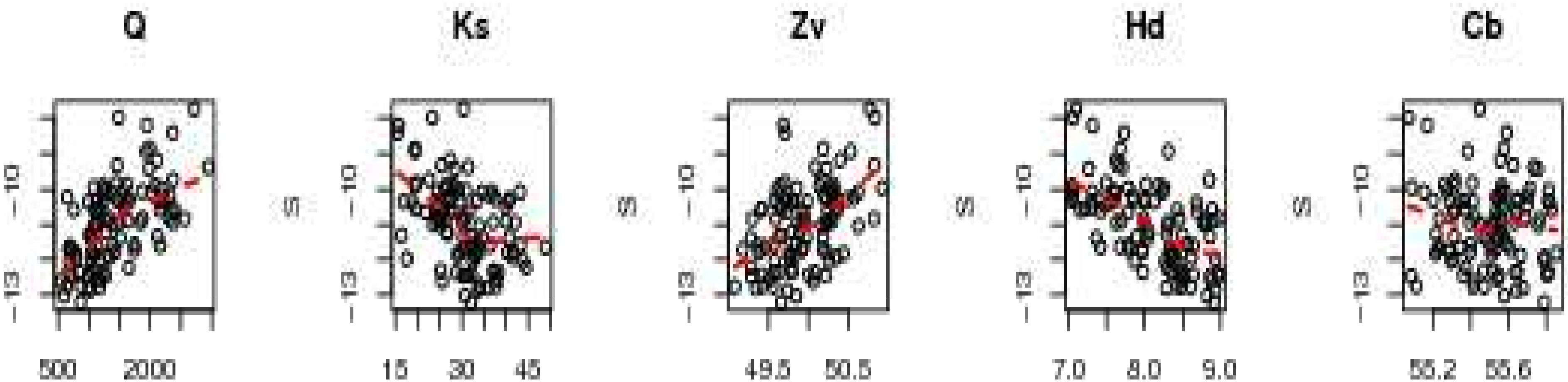
- Validity of the linear model via

The residuals diagnostics and R^2 : $$R^2=1-\frac{\sum_{i=1}^{n}(\hat{Y}_i-Y_i)^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}$$

- We have $R^2=\sum_{i=1}^{p}\mathrm{SRC}^2(X_i)$ => nice interpretation of SRC

Flood model - Output S

Monte Carlo sample – N = 100



31% 15% 18% 29% 6%

Sensitivity indices (SRC^2)

The model is linear (R^2=0.99)

SRC coefficients are sufficient for the quantitative sensivity analysis

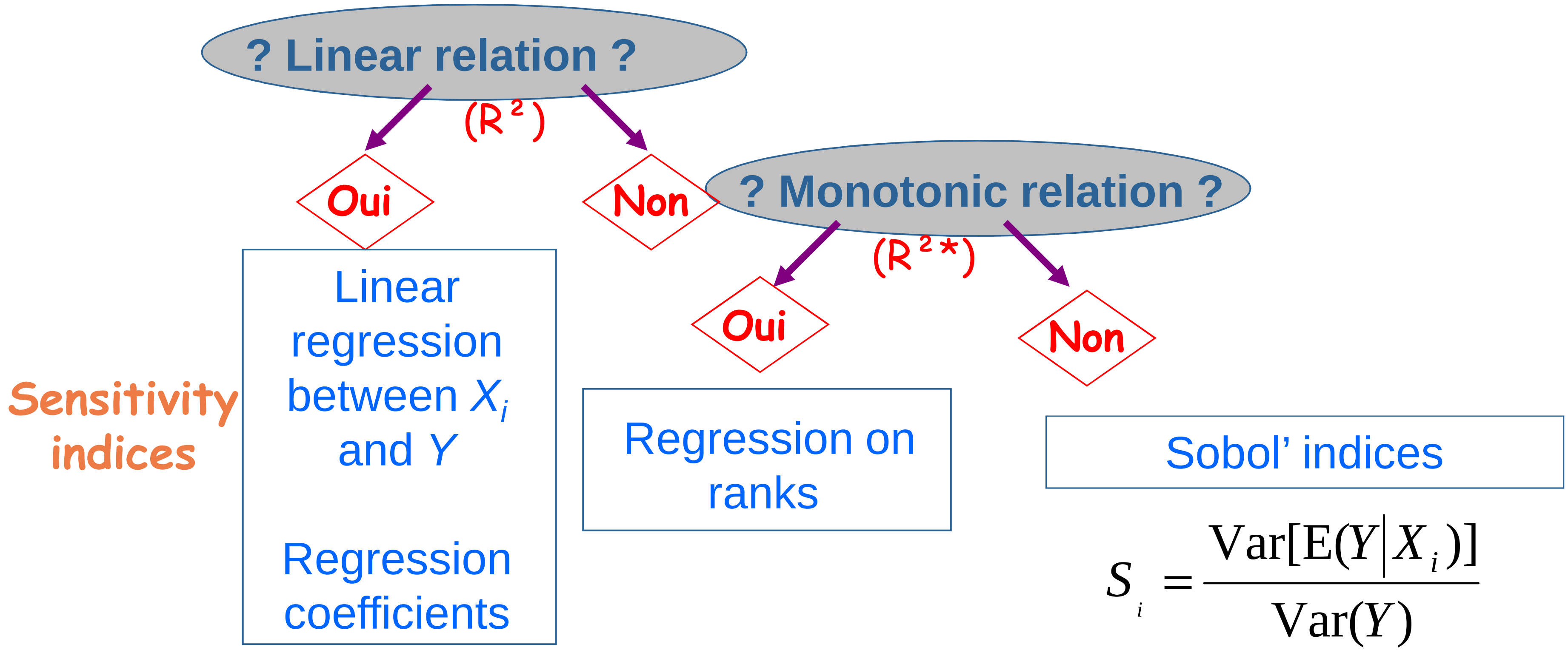
Sensitivity analysis for one scalar output

Sample ($\mathbf{X} \in \Re^p$, $Y(\mathbf{X}) \in \Re$) of size $N > p$

Preliminary step: graphical vizualisation (for ex: scatterplots)

Quantitative sensitivity analysis methodology

[Saltelli et al. 00, Helton et al. 06]



$$S_i = \frac{\mathrm{Var}[\mathrm{E}(Y|X_i)]}{\mathrm{Var}(Y)}$$

Functional decomposition

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

with $f(\mathbf{x}) \in L^2(\mathbf{x}) \quad \mathbf{x} \in [0;1]^p$

Infinity of possible decompositions

BUT, unicity of decomposition if: $\int f_{i_1 \ldots i_s}(x_{i_1}, \ldots, x_{i_s}) dx_j = 0 \;\; \forall j = i_1, \ldots, i_s$

<u>Properties</u> ($x_i \sim U[0,1]$ for i=1,…,*p* , the x_i s are independent)

$$f_0 = \int f(\mathbf{x})\, d\mathbf{x} = \mathrm{E}(y)$$

$$f_i(x_i) = \int f(\mathbf{x}) dx_{-i} - f_0 = \mathrm{E}(y \mid x_i) - f_0$$

$$f_{ij}(x_i, x_j) = \mathrm{E}(y \mid x_i, x_j) - \mathrm{E}(y \mid x_i) - \mathrm{E}(y \mid x_j) + f_0$$

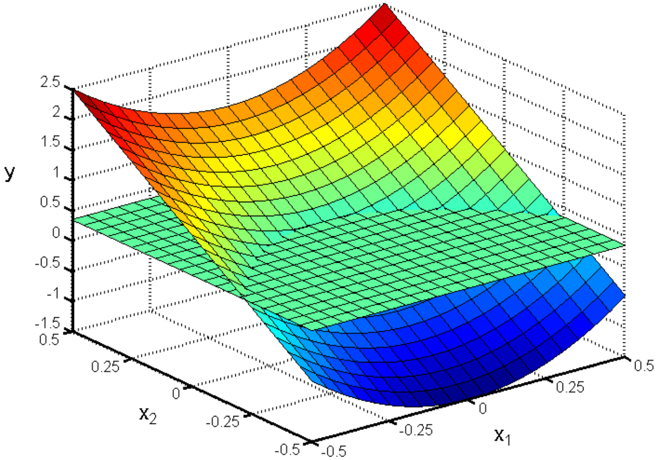
<u>Example :</u> $f(x_1, x_2) = x_1 + x_2 \;;\; x_1 \sim U[0;1] \;;\; x_2 \sim U[0;1]$

$$f_0 = 1 \;;\; f_1(x_1) = x_1 - \frac{1}{2};\;\; f_2(x_2) = x_2 - \frac{1}{2} \;;\; f_{12}(x_1, x_2) = 0$$

Another example

$$f(x_1, x_2) = 4x_1^2 + 3x_2$$

$$x_1, x_2 \in U[-1/2; 1/2]$$



$$f_0 = 0$$

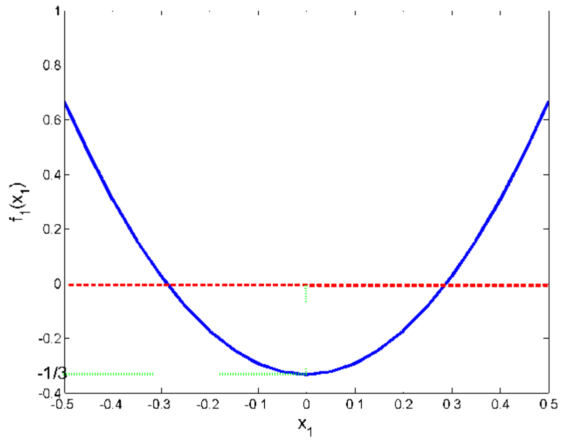
$$f_1(x_1) = 4x_1^2$$

$$f_2(x_2) = 3x_2$$

$$f_{12}(x_1, x_2) = 0$$

$$f_0 = E(y) = \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} (4x_1^2 + 3x_2)\,dx_1 dx_2 = \frac{1}{3}$$

$$f_1(x_1) = E(y \mid x_1) - f_0 = \int_{-1/2}^{1/2} (4x_1^2 + 3x_2)\,dx_2 = 4x_1^2 - \frac{1}{3}$$



$$f_2(x_2) = E(y \mid x_2) - f_0 = 3x_2$$

$$f_{12}(x_1, x_2) = 0$$

[JRC 2010]

Sensitivity indices without model hypotheses

Functional ANOVA *[Efron & Stein 81]* (hyp. of independent X_i s) :

$$\mathrm{Var}(Y) = \sum_{i=1}^{p} V_i(Y) + \sum_{i<j}^{p} V_{ij}(Y) + \cdots + V_{12\cdots p}(Y)$$

where $V_i(Y) = \mathrm{Var}[\mathrm{E}(Y|X_i)]$

$$V_{ij} = \mathrm{Var}\left[\mathrm{E}(Y|X_i X_j)\right] - V_i - V_j, \ldots$$

Sobol indices definition:

- First order sensitivity indices: $S_i = \dfrac{V_i}{\mathrm{Var}(Y)}$
- Second order sensitivity indices: $S_{ij} = \dfrac{V_{ij}}{\mathrm{Var}(Y)}$
- ...

Another example

$$y = f(x_1, x_2) = 4x_1^2 + 3x_2 \qquad x_1, x_2 \in U[-1/2, 1/2]$$

On a vu :

$$f_0 = E(y) = \frac{1}{3}$$

$$f_1(x_1) = E(y \mid x_1) - f_0 = 4x_1^2 - \frac{1}{3}$$

$$f_2(x_2) = E(y \mid x_2) - f_0 = 3x_2$$

$$f_{12}(x_1, x_2) = 0$$

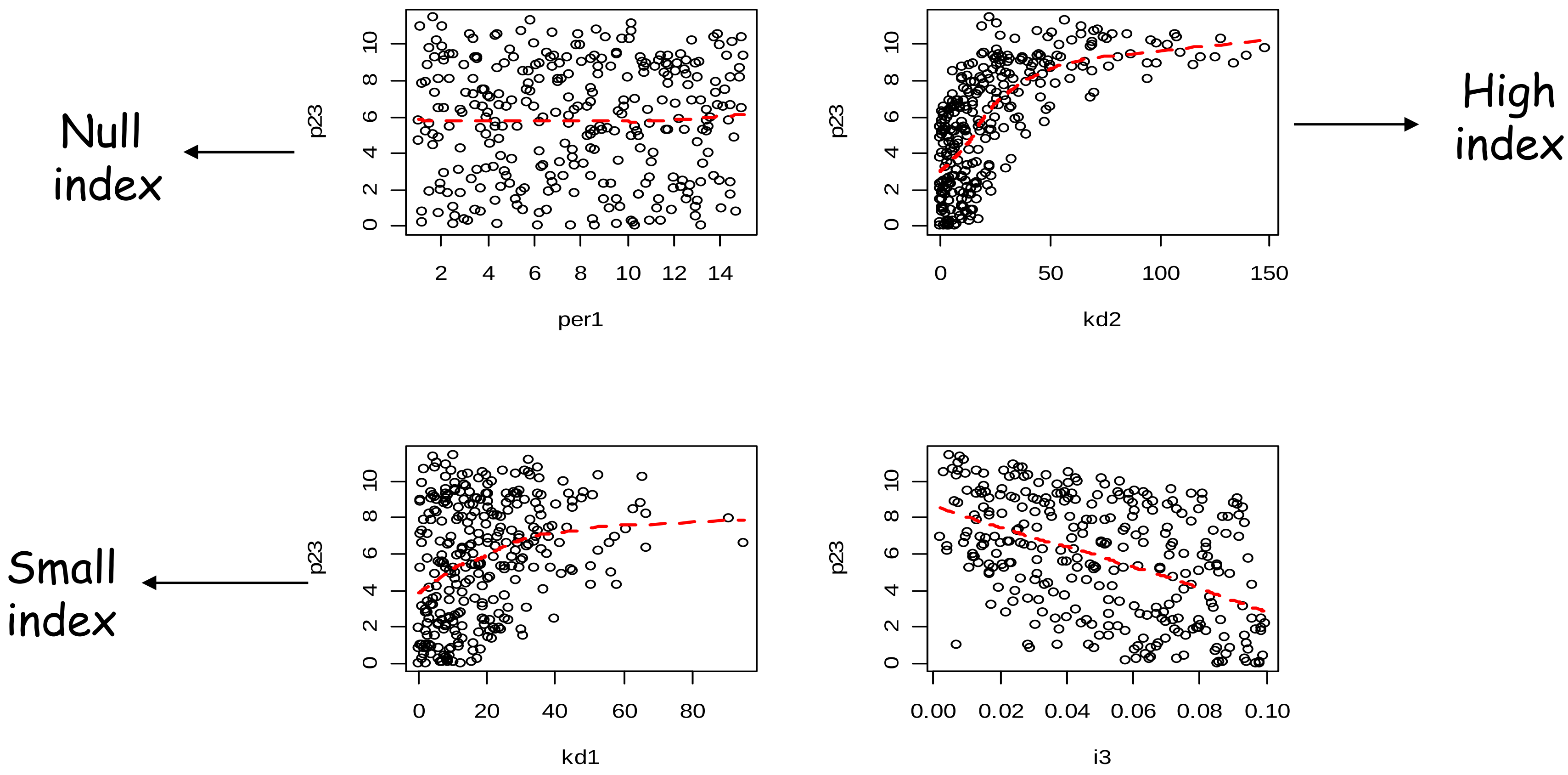
$$\boxed{S_1 = \frac{Var[f_1(x_1)]}{V} = \frac{0.0\overline{8}}{0.83\overline{8}} = 0.106}$$

$$\boxed{S_2 = \frac{Var[f_2(x_2)]}{V} = \frac{0.75}{0.83\overline{8}} = 0.894}$$

[JRC 2010]

Graphical interpretation

First order Sobol' indices measure the variability of conditional expectations (mean trend curves in the scatterplots)



Sobol' indices properties

$$1 = \sum_{i=1}^{p} S_i + \sum_i \sum_j S_{ij} + \sum_i \sum_j \sum_k S_{ijk} \ldots + S_{1,2,\ldots,k}$$

$\sum_i S_i \leq 1$ Always

$\sum_i S_i = 1$ Additive model

$1 - \sum_i S_i$ Measure the degree of interactions between variables

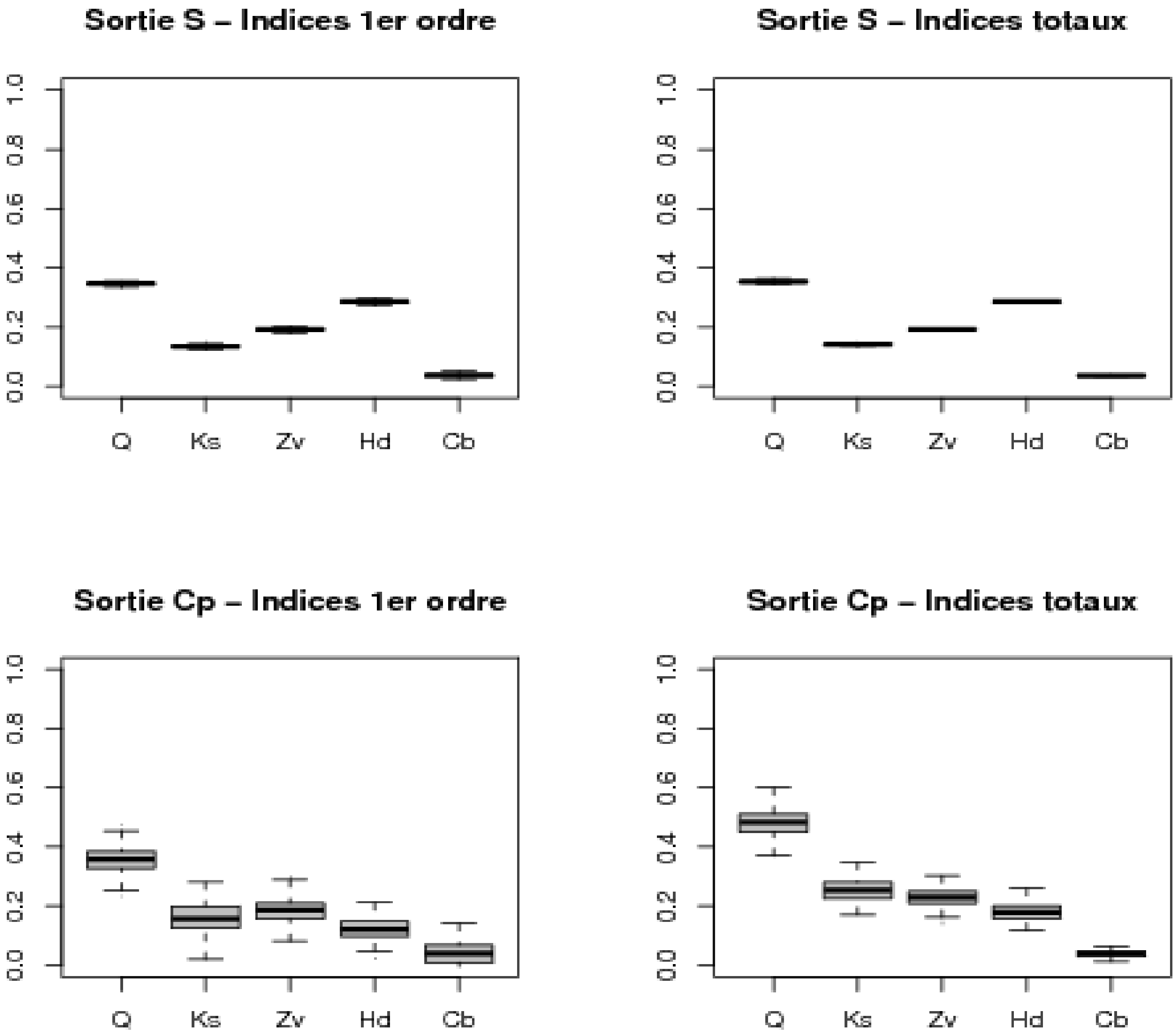
<u>Examples :</u> p =4 gives 4 indices S_i, 6 indices S_{ij} , 4 indices S_{ijk}, 1 indice S_{ijkl}

General case : 2^p-1 indices to be estimated

Total sensitivity index: $S_{Ti} = S_i + \sum_j S_{ij} + \sum_{j,k} S_{ijk} + \ldots = 1 - S_{\sim i}$

[Homma & Saltelli 1996]

Flood model


Sortie S – Indices 1er ordre
Sortie S – Indices totaux
Sortie Cp – Indices 1er ordre
Sortie Cp – Indices totaux
0.0
0.2
0.4
0.6
0.8
1.0
Q
Ks
Zv
Hd
Cb

Sobol indices computation

- Indices for X_i (1st order and total) :

$$S_i = \frac{V_i}{\mathrm{Var}(Y)} \text{ and } S_{T_i} = 1 - \frac{V_{\sim i}}{\mathrm{Var}(Y)}$$

- Formulations of the conditional variances:

Let $\mathbf{X} = (X_{i,} X_{\sim i})$ and $\mathbf{X}'$ an independent copy of $\mathbf{X}$

$$V_i(Y) = \mathrm{Var}[\mathrm{E}(Y|X_i)] = \int \mathrm{E}^2(Y|X_i)dX_i - \left(\int \mathrm{E}(Y|X_i)dX_i\right)^2 = \mathrm{Cov}\left[f(X_{i,} X_{\sim i}),\, f(X_{i,} X'_{\sim i})\right]$$

$$V_{\sim i}(Y) = \mathrm{Var}[\mathrm{E}(Y|X_{\sim i})] = \mathrm{Cov}\left[f(X_{i,} X_{\sim i}),\, f(X'_i, X_{\sim i})\right]$$

Direct estimation via Monte Carlo

2 i.i.d. samples : $\left(X_i^{(j)}\right)_{i=1,..,p;\, j=1,..,n}$ and $\left(X_i'^{(j)}\right)_{i=1,..,p;\, j=1,..,n}$

Variance (classical estimator) : $\hat{V}(Y) = \frac{1}{n}\sum_{k=1}^{n} f\left(\mathbf{X}^{(k)}\right)^2 - \hat{f}_0^2$ avec $\hat{f}_0 = \frac{1}{n}\sum_{k=1}^{n} f\left(\mathbf{X}^{(k)}\right)$

Conditional variances estimation:

$$\hat{V}_i(Y) = \frac{1}{n}\sum_{k=1}^{n} f\left(X_1^{(k)},\ldots,X_{i-1}^{(k)},X_i^{(k)},X_{i+1}^{(k)},\ldots,X_p^{(k)}\right) f\left(X_1'^{(k)},\ldots,X_{i-1}'^{(k)},X_i^{(k)},X_{i+1}'^{(k)},\ldots,X_p'^{(k)}\right) - f_0^2$$

Indices 1st order : cost = n (p +1)

$$\hat{V}_{\sim i}(Y) = \frac{1}{n}\sum_{k=1}^{n} f\left(X_1^{(k)},\ldots,X_{i-1}^{(k)},X_i^{(k)},X_{i+1}^{(k)},\ldots,X_p^{(k)}\right) f\left(X_1^{(k)},\ldots,X_{i-1}^{(k)},X_i'^{(k)},X_{i+1}^{(k)},\ldots,X_p^{(k)}\right) - f_0^2$$

Indices 1st order + total indices : cost = n (p +2), by inverting $\left(X_i^{(j)}\right)_{i=1,..,p;\, j=1,..,n}$ and $\left(X_i'^{(j)}\right)_{i=1,..,p;\, j=1,..,n}$ in $\hat{V}_{\sim i}(Y)$

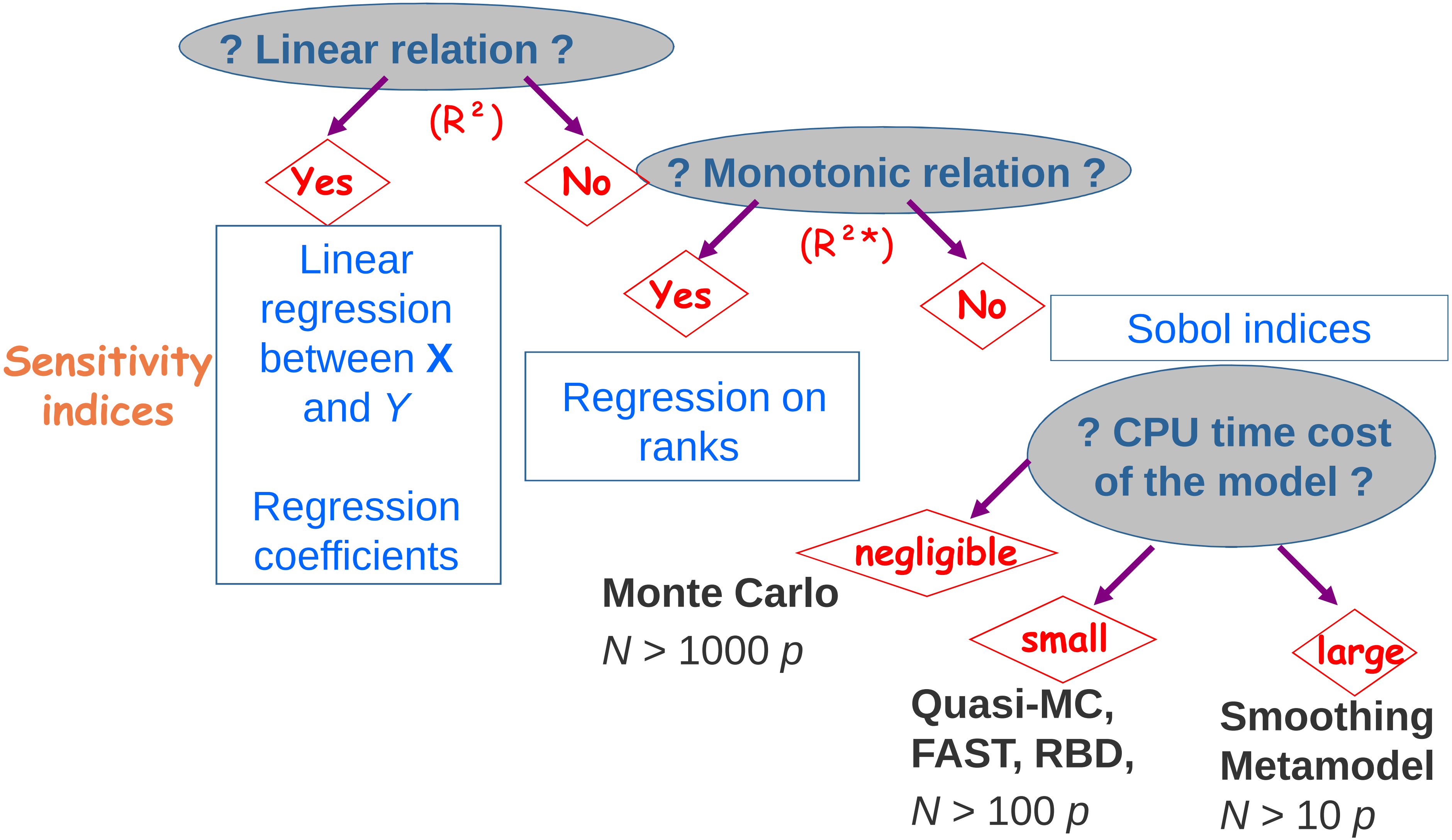
In practice, n ~ 1e4 => problem of the cost in terms of required model runs

Other formula (Jansen-Sobol estimator):

$$\hat{V}_i = \frac{1}{n}\sum_{k=1}^{n} f(X_{k,1}^{(2)},\ldots,X_{k,p}^{(2)})\left[f(X_{k,1}^{(1)},\ldots,X_{k,i-1}^{(1)},X_{k,i}^{(2)},X_{k,i+1}^{(1)},\ldots,X_{k,p}^{(1)}) - f(X_{k,1}^{(1)},\ldots,X_{k,p}^{(1)})\right]$$

The sampling-based approaches

Sample ($\mathbf{X} \in \Re^p$, $Y(\mathbf{X}) \in \Re$) of size $N > p$



Flood model – Output Cp

From the 100-size Monte Carlo sample, a **Gaussian process metamodel** is fitted

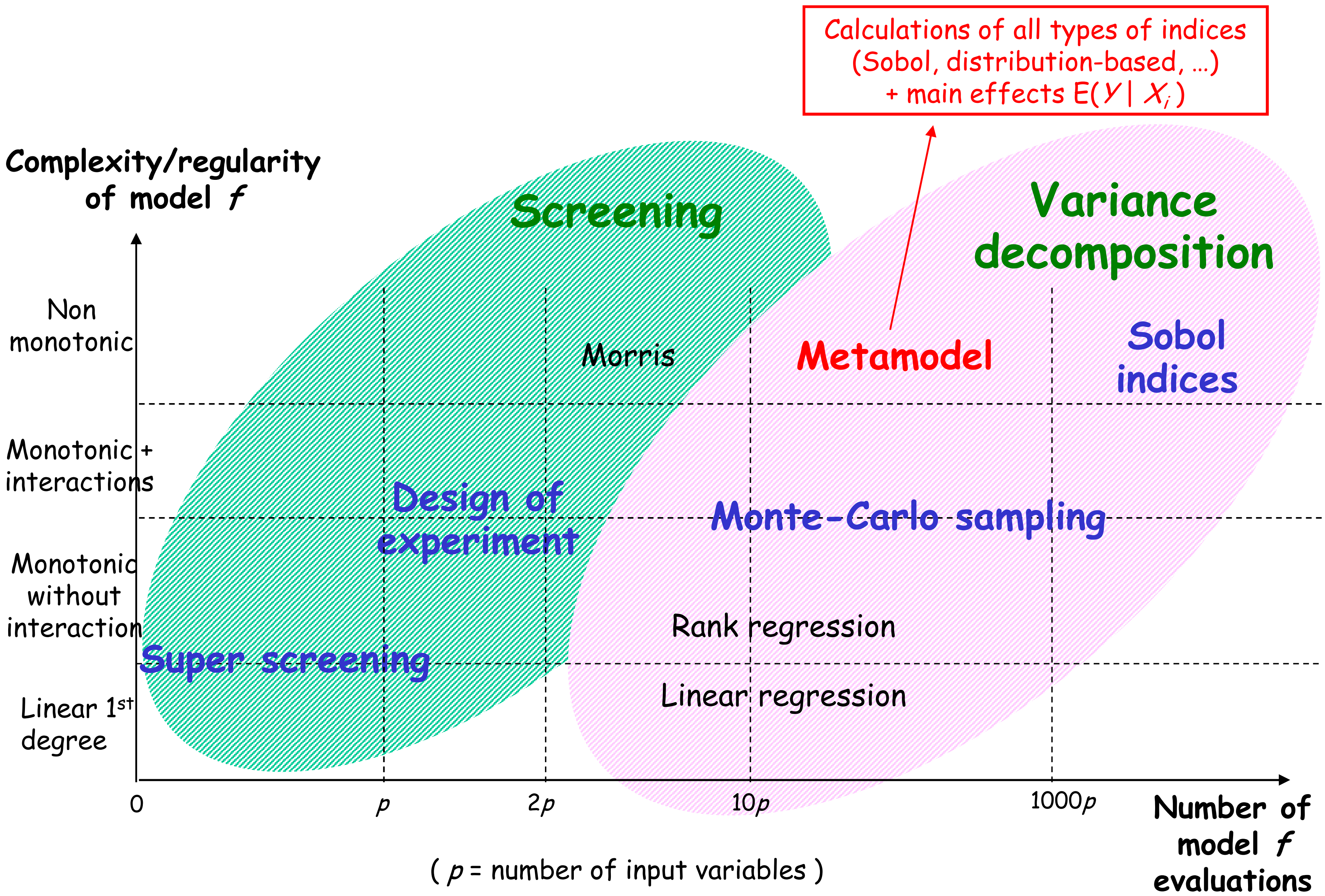
Predictivity of the Gp metamodel : Q_2 = 99%

N=1e5
100 replicates

N x (p+2) x 100 = 7e7 evaluations

Indices (en %)	Q	K_s	Z_v	H_d	C_b
S_i modèle	35.5	15.9	18.3	12.5	3.8
S_i métamodèle	38.9	16.8	18.8	13.9	3.7
S_{T_i} modèle	48.2	25.3	22.9	18.1	3.8
S_{T_i} métamodèle	45.5	21.0	21.3	16.8	4.3

Classification of sensitivity analysis methods



(p = number of input variables)

Bibliography

– Fang et al., *Design and modeling for computer experiments*, Chapman & Hall, 2006

– J.C. Helton, J.D. Johnson, C.J. Salaberryet C.B. Storlie: Survey of sampling-based methods for uncertainty and sensitivity analysis. Reliability Engineering and System Safety, 91:1175–1209, 2006.

– Kleijnen, *The design and analysis of simulation experiments*, Springer, 2008

– Koehler & Owen, Computer experiments, 1996

– A. Saltelli, K. Chan & E.M. Scott, *Sensitivity analysis,* Wiley, 2000

– A. Saltelli et al., Global sensitivity analysis - The primer. Wiley, 2008.