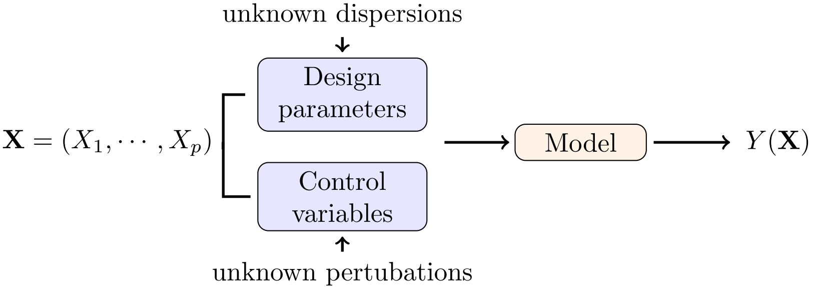
The design and sensitivity analysis of experiments

Thibault Delage
EDF R&D

02 May 2017

The design of experiments

The general context



unknown dispersions

$\mathbf{X} = (X_1, \cdots, X_p)$

Design parameters

Control variables

Model

$Y(\mathbf{X})$

unknown pertubations

Hypotheses

- The components of $\mathbf{X}$ are independent
- Each $\mathbf{X}_i$ is distributed into $[0, 1]$

The design of experiments

What is the design of experiments ?

- Set the experimentation/simulation points in the inputs space
- Select the combinations of input values that will provide the most informative inputs-output relationship

What is the design of experiments for ?

- Explore the model with a limited number of inputs
- Identify area of interest with respect to Y
- Provide an optimized design for sensitivity analysis

Designs families

1. Factorial designs
   - Settings
     - Turn quantitative inputs $\mathbf{X}$ into factors with levels $\{0, 1\}$
     - To a simple form of model $Y(\mathbf{X})$, there corresponds an optimized design (with the smallest number of points)
   - Example
     - The One at A Time (OAT) design
     - The full factorial design
     - The fractional factorial design
2. Numerical designs of experiments
   - Settings
     - More general than factorial designs
     - Space filling designs for computer experiments
   - Example
     - The Latin Hypercube sampling (LHS)
     - The Quasi Monte Carlo

The One at A Time (OAT) design

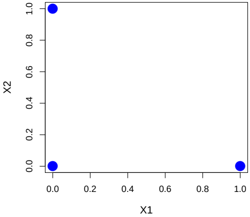
The linear model

$$Y(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

The optimal design for this model is the OAT design, $X_i \in \{0, 1\}$

Example with $p = 2$: $Y(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

X_1	X_2
0	0
1	0
0	1



$N = p + 1$ simulations

Drawbacks

- Do not detect interactions, discontinuities
- In high dimension, the inputs space is partially covered

The full factorial design

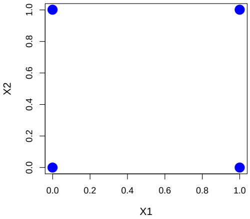
The linear model with interactions

$$Y(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_p X_p + \sum_{i<j} \beta_{i,j} X_i X_j + \cdots + \beta_{1,\dots,p} X_1 \cdots X_p + \varepsilon$$

The optimal design is the full factorial design

Example with $p = 2$: $Y(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{1,2} X_1 X_2 + \varepsilon$

X_1	X_2	$X_1 X_2$
0	0	1
1	0	0
0	1	0
1	1	1



$N = 2^p$ simulations

Drawbacks

- If p large, very expensive : $p = 10$, $N = 1024$ simulations !

The fractional factorial design

Linear model with partial interactions

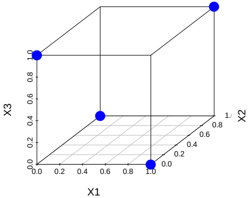
$$Y(X) = \beta_0 + \beta_1 X_1 + \beta_p X_p + \sum_{i<j} \beta_{i,j} X_i X_j + \cdots + \varepsilon$$

- Choice of aliasing effects ⇒ define the degree q of fraction
- Reduce the number of experiments to 2^{p-q}

Example with $p = 3$ and $q = 1$:

$$Y(\mathbf{X}) = \beta_0 + \sum_{i=1}^{3} \beta_i X_i + \sum_{i<j=1}^{3} \beta_{i,j} X_i X_j + \beta_{1,2,3} X_1 X_2 X_3 + \varepsilon$$

X_1	X_2	$X_3 \equiv X_1 X_2$
0	0	1
1	0	0
0	1	0
1	1	1



$N = 2^{3-1}$ simulations

$$\Rightarrow Y(\mathbf{X}) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + \varepsilon$$

The fractional factorial design

Resolution

- Resolution III : Main effects may be aliased with two interaction effects
- Resolution IV : Two interaction effects can be aliased
- Resolution V : Main effects are aliased with n-interaction effects, $n \geq 4$

	p							
	3	4	5	6	7	8	9	10
$N = 2^{p-q}$ / $N = 2^p$	8	16	32	64	128	256	512	1024
4	III							
8		IV	III	III	III			
16			V	IV	IV	IV	III	III
32				V	IV	IV	IV	IV
64					VII	V	IV	IV
128						VIII	VI	V
256							IX	VI
512								X

Drawbacks

- Determine which effects are aliasing

Factorial designs

To sum up

- Factorial designs are optimal for analytical polynomial models
- Fractional designs imply aliasing and degree of resolution
- Many other designs exit for polynomial/surface of response

Advantages

- Once the model is well defined, an optimal design can be easily built
- The number of experiments is minimized with respect to the complexity of the model
- The inputs-output relationship can be easily interpreted

Drawbacks

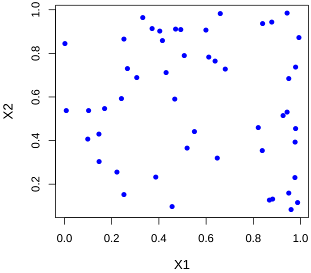
- All experiments have to be made
- Very oriented by the choice of the model
- Not suited to complex models

Numerical designs

Why a numerical design ?

- No assumption on the form of the model $Y(x)$
- General designs well suited to a large number of models
- Fill in "regularly and correctly" the inputs space by a number N of points

The Monte Carlo design



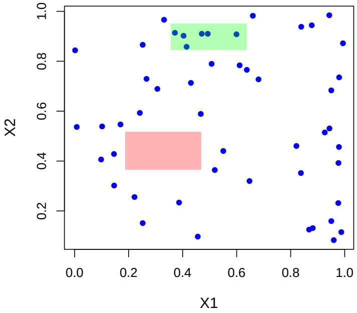
Does this "regurlarly and correctly" fill in the space ?

Numerical designs

What does ”regularly and correctly” mean ?

- **Correctly :** The inputs space is entirely covered
- **Regularly :** No subspaces are over/under covered : No points to be too closed together

The Monte Carlo design



Does this ”regurlarly and correctly” fill in the space ?

🟥 Not correctly

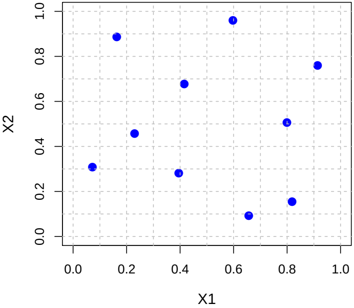
🟩 Not regularly

The Latin Hypercube Sampling (LHS)

The principle

- Divide each axis $[0,1]$ into N equally spaced intervals $[0,1/N),\cdots,[(N-1)/N,1]$
- Select randomly N points s.t. each appears exactly once in each row and each column of this grid

The LHS design



Optimization criteria

- Measure of closeness of the points in the N-points set $\mathcal{D}$. For example,

$$\max_{\mathcal{D}} \min_{\mathbf{x}^1,\mathbf{x}^2\in\mathcal{D}} d(\mathbf{x}^1,\mathbf{x}^2),$$

  Usually, $d(x,y)=\sqrt{\sum_j (x_j-y_j)^2}$

  - Low cost
  - Very efficient for small p only

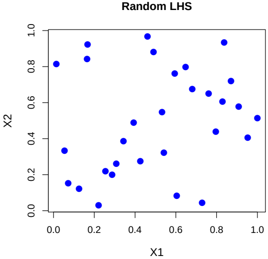
- Low discrepancy sequence : proportion of points falling into an arbitrary set $\mathcal{B}$ close to proportional to the measure of $\mathcal{B}$

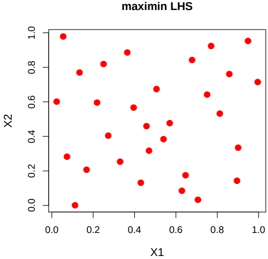
$$D_N(\mathcal{P}) = \sup_{\mathcal{B}} \left| \frac{A(\mathcal{B};\mathcal{P})}{N} - \lambda(\mathcal{B}) \right|$$

- Sobol, Halton, Faure, . . . , sequences
- Fast convergence of the mean estimate
- Very complex to build

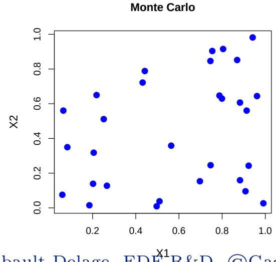
Optimization criteria

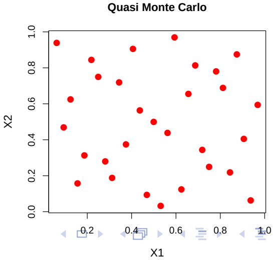
Maximin standard





Low discrepancy sequence





Numerical designs

To sum up

- Numerical designs are more adapted to complex models/computer experiments
- Are space filling designs

Advantages

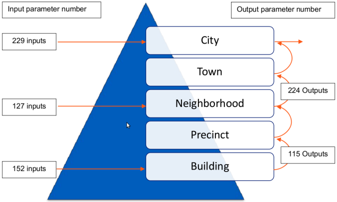
- Well suited for non linear models
- Cover a large variation domain
- The experiments are deterministic and the number can be increased if necessary

Drawbacks

- Define the best criteria to be optimized
- Can fail in high dimension

Sensitivity analysis methods

Industrial case study : The simulation of sustainable cities

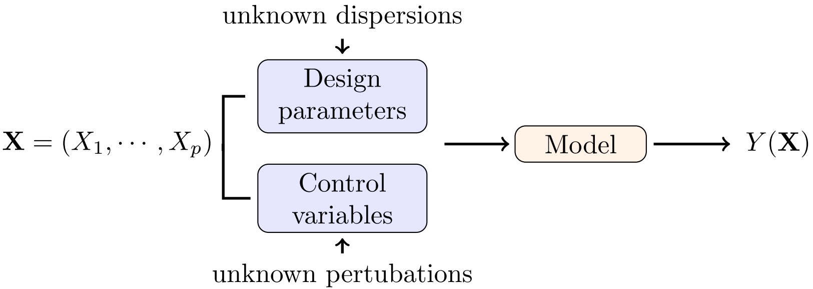


- Quantity of interest : Total annual thermal production, and treated waste
- 55 Input parameters : temperature, building features, waste, use of natural ressources,. . .

Why a sensitivity analysis ?

- The model is too time-consuming/complex because it depends on many input parameters
- Inputs are subject to different sources of variability ⇒ poor confidence in the response

The general context



unknown dispersions

$\mathbf{X} = (X_1, \cdots, X_p)$

Design parameters

Control variables

unknown pertubations

Model

$Y(\mathbf{X})$

Hypotheses

- The components of $\mathbf{X}$ are independent
- $\mathbf{X}$ are uniformely distributed $\mathcal{U}[0,1]^p$
- $Y \in L^2(\mathbb{R})$ i.e.

$$\int Y^2(\mathbf{x}) f_Y(\mathbf{x}) d\mathbf{x} = \mathbb{E}[Y^2(\mathbf{X})] < +\infty$$

The sensitivity analysis

What is sensitivity analysis ?

Identify the X_i's that most contribute to the variability of $Y(\mathbf{X})$

- Which one is the most/least influent ?
- What is the weight of the each contribution ?

What is sensitivity analysis for ?

- Understand/check if a model is a good approximation of the physical system/process
- Increase the confidence into the model
- Reduce the number of input parameters
- Determine if parameters interact with each other

Families of sensitivity methods

1. Screening methods
    - Morris algorithm
2. Local methods
    - Quadratic summation method
3. Global methods
    - The Sobol index

Screening methods

Goal

- Well suited to an important number of inputs and an expensive code
- Aim to determine a few influent factors and a majority of non influent ones ⇒ reduce the model dimension
- These are qualitative methods to rank (groups of) parameters in order of importance

The general idea

- Based on the discretization of the inputs space called levels
- Linked to the field of designs of experiments

The Morris method

Preliminary

- Set a regular grid Ω of $[0,1]^p$ into Q levels :
$\Omega = \left\{0, \frac{1}{Q-1}, \frac{2}{Q-1}, \cdots, 1\right\}^p$
- Choose a perturbation $\Delta \in [0,1]$ and $\Delta \propto \frac{1}{Q-1}$

The path

For $\mathbf{x}^* \in \Omega$ randomly chosen, One at A Time (OAT) design :

1. First point $P_0^* = Y(\mathbf{x}^*)$
2. Select a component, say x_i^*, disturbed by Δ

$$P_1^* = Y(x_1^*, \cdots, x_i^* \pm \Delta, x_{i+1}^*, \cdots, x_p^*)$$

3. Select $j \neq i$, and compute

$$P_2^* = Y(x_1^*, \cdots, x_i^* \pm \Delta, \cdots, x_j^* \pm \Delta, \cdots, x_p^*)$$

4. Repeat 3 until $x_1, \cdots, x_p$ successively vary $\Rightarrow P = \{P_0^*, \cdots, P_p^*\}$

The OAT design

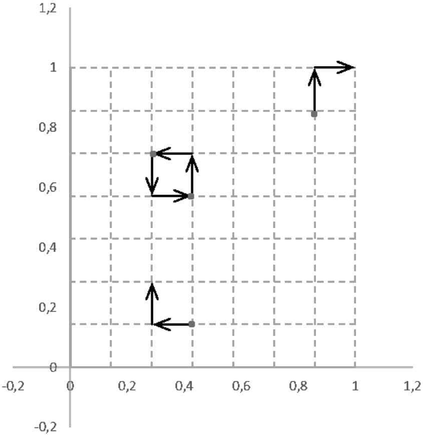


Figure: Examples of 4 paths in two dimensions. The support $[0, 1]$ is discretized into $Q = 8$ levels ; The perturbation is $\Delta = \dfrac{1}{7}$

The Morris algorithm

Repeat the previous procedure r times, so that there is $P^1, \cdots, P^r$ paths

Elementary effect of X_i

For a path P^k

$$d_i^k = \frac{|Y(\cdots, x_i^1 \pm \Delta, \cdots) - Y(\cdots, x_i^1, \cdots)|}{\Delta}$$

The mean effect of X_i

$$\mu_i = \frac{1}{r} \sum_{k=1}^{r} d_i^k$$

The variation effect of X_i

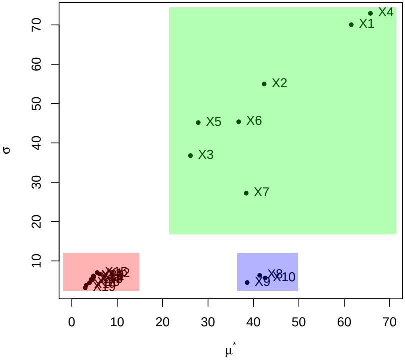
$$\sigma_i = \sqrt{\frac{1}{r-1} \sum_{k=1}^{r} (d_i^k - \mu_i)^2}$$

The cost

Requires $r(p+1)$ simulations

The Morris algorithm

Graphical representation of (μ_i, σ_i), for all $i \in \{1, \cdots, p\}$
Example on the Morris function ($p = 20$)



🟥 Non influent

🟦 Linear effects

🟩 Non linear or interaction effects

The Morris algorithm

To sum up

- Construct a OAT design and conpute the elementary effects from r paths
- If $(\mu_i, \sigma_i) \simeq 0$, the parameters are non influent
- If the variation effect $\sigma_i \simeq 0$ but $\mu_i \neq 0$, the parameters have linear effects

Advantages

- Intuitive and easy to use
- Only $r(p+1)$ simulations required
- Does not require any assumptions of the model's regularity

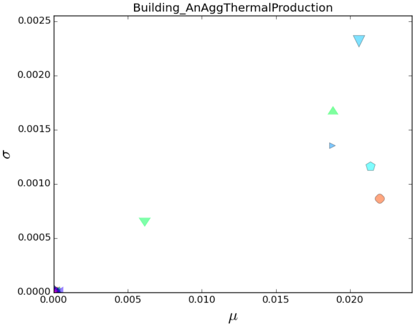
Drawbacks

- Qualitative method to detect the least influent parameters only
- Not able to split non linear effects from the interact ones

The simulation of sustainable cities

- Output : Total annual thermal production
- 55 Inputs : temperature, building features, waste, use of natural ressources,. . .

The screening results



Selected parameters : electricity demand, surface,luminaire.

Local methods

Goal

- Deterministic method
- Provide the slope of Y in the parameter space at given values
- Allow a rapid preliminary exploration of the model

The intuitive idea

1. Locally disturb one parameter X_i at a time
2. Run the model with the perturbated X_i
3. Compare with the response without perturbation, i.e.

$$\frac{Y(X_i + \Delta_{X_i}) - Y(X_i)}{\Delta_{X_i}} \simeq \frac{\partial Y}{\partial X_i}$$

The quadratic summation method

The Taylor expansion

$$\begin{aligned} Y(\mathbf{X}) &= Y(\mu) + \sum_{i=1}^{p} \left.\frac{\partial Y}{\partial X_i}\right|_{\mathbf{X}=\mu} (X_i - \mu_i) \\ &+ \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \left.\frac{\partial^2 Y}{\partial X_i \partial X_j}\right|_{\mathbf{X}=\mu} (X_i - \mu_i) \cdot (X_j - \mu_j) + o(\|\mathbf{X} - \mu\|) \end{aligned}$$

Hypotheses

- The model is linear around μ
- μ_i is the nominal value of X_i
- $V(X_i)$, $V(Y)$ are the variances of X_i and Y

The quadratic summation method

The Taylor expansion becomes

$$Y(\mathbf{X}) = Y(\mu) + \sum_{i=1}^{p} \left.\frac{\partial Y}{\partial X_i}\right|_{\mathbf{X}=\mu} (X_i - \mu_i)$$

Then

$$V(Y) = \sum_{i=1}^{p} \left(\left.\frac{\partial Y}{\partial X_i}\right|_{\mathbf{X}=\mu} \right)^2 V(X_i)$$

Scaled sensitivity index

$$\boxed{\eta_i^2 = \left(\left.\frac{\partial Y}{\partial X_i}\right|_{\mathbf{X}=\mu} \right)^2 \frac{V(X_i)}{V(Y)} \in [0;1]}$$

Estimation method

- Monte Carlo estimation to compute $V(X_i)$ and $V(Y)$
- Automatic differentiation, finite differences to get the partial derivatives

Summary

To sum up

- Local variation of the inputs around a nominal value
- Estimation of partial derivatives

Advantages

- Useful information on the behavior of Y near the nominal values of parameters
- Rapid preliminary exploration of the model
- Could be well suited to the probability of failure

Drawbacks

- Can only be used when the model is linear or when the variation around a baseline is small
- Not well suited to measure the effects of various parameters on the output

The simulation of sustainable cities

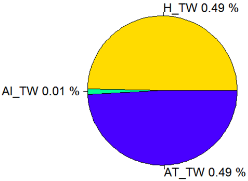
- Output : Total treated waste
- Inputs :
  1. ACT Households 1
     1. Population (H_pop)
     2. Total Treated Waste (H_TW)
  2. ACT Activities Industrials W1
     1. Population (AI_pop)
     2. Total Treated Waste (AI_TW)
  3. ACT Tertiary Activities W1
     1. Population (AT_pop)
     2. Total Treated Waste (AT_TW)

Taylor expansion around the mean

1-order	2-order
592,332	592,842

⇒ Linear model

Local sensitivity indices



Global methods

Goal

- Instead of local perturbation, global methods aim to consider the whole input space
- Able to quantify (and to rank) the sensitivity of each parameter on the model response
- Able to quantify the interactions among parameters

The general idea

- Based on the random distribution of inputs-output
- Generate a sample of observations from the inputs distribution

The Sobol index

Intuitive idea

- When X_i is fixed, how does the output Y behave ?

$$V(Y|X_i) \quad \text{and} \quad \mathbb{E}[V(Y|X_i)]$$

X_i is influent if $\mathbb{E}[V(Y|X_i)]$ is small

- The total variance decomposition

$$V(Y) = \mathbb{E}[V(Y|X_i)] + V[\mathbb{E}(Y|X_i)]$$

- The first-order Sobol index

$$\boxed{S_i = \frac{V[\mathbb{E}(Y|X_i)]}{V(Y)} \in [0,1]}$$

X_i is influent if S_i is close to 1.

The FANOVA

The functional decomposition ANOVA

$$Y(\mathbf{X}) = Y_0 + \sum_{i=1}^{p} Y_i(X_i) + \sum_{i<j=1}^{p} Y_{ij}(X_i, X_j) + \cdots + Y_{1\cdots p}(\mathbf{X}) \qquad (1)$$

The decomposition exists and is unique if

$$\mathbb{E}[Y_u(\mathbf{X}_u)Y_v(\mathbf{X}_v)] = 0,\ \forall u \neq v \subseteq \{1, \cdots, p\}$$

Applying the variance to (1),

$$V[Y(\mathbf{X})] = \sum_{i=1}^{p} V[Y_i(X_i)] + \sum_{i<j=1}^{p} V[Y_{ij}(X_i, X_j)] + \cdots + V[Y_{1\cdots p}(\mathbf{X})]$$

The Sobol index

The Sobol index of a group of parameters $\mathbf{X}_u$

$$S_u = \frac{V[Y_u(\mathbf{X}_u)]}{V[Y(\mathbf{X})]}$$

Properties of the FANOVA

$$\begin{array}{lcl} Y_0 & = & \mathbb{E}[Y(\mathbf{X})] \\ Y_i(X_i) & = & \mathbb{E}[Y(\mathbf{X})|X_i] - Y_0 \\ Y_{ij}(X_i, X_j) & = & \mathbb{E}[Y(\mathbf{X})|X_i, X_j] - \mathbb{E}[Y(\mathbf{X})|X_i] - \mathbb{E}[Y(\mathbf{X})|X_j] + Y_0 \\ & & \vdots \end{array}$$

Finally, the first-order Sobol index of X_i is

$$\boxed{S_i = \frac{V[\mathbb{E}(Y(\mathbf{X})|X_i)]}{V[Y(\mathbf{X})]}}$$

The second-order Sobol index of the couple (X_i, X_j) is

$$\boxed{S_{ij} = \frac{V[\mathbb{E}(Y(\mathbf{X})|X_i, X_j)]}{V[Y(\mathbf{X})]} - S_i - S_j}$$

The Sobol index

Properties of the Sobol index

- $2^p - 1$ Sobol indices are constructed
- $S_i \in [0,1]$. The closer to 1, the more influent is X_i
- If $\sum_{i=1}^{p} S_i = 1$, the model is additive
- The total index measures the total contribution of X_i

$$S_{T_i} = \sum_{u \ni i} S_u$$

Sobol index estimation

- Monte Carlo estimation
- Spectral decomposition (FAST)
- Meta-modeling if the model is too expensive (linear model, polynomial chaos)

The Monte Carlo estimation

Let

- $\mathbf{X} = (X_i, \mathbf{X}_{-i})$ and an independent copy $\mathbf{X}^* = (X_i, \mathbf{X}^*_{-i})$
- $Y = Y(\mathbf{X})$, $Y^* = Y(\mathbf{X}^*)$

Then

$$S_i = \frac{\mathrm{Cov}(Y, Y^*)}{V(Y)}$$

- Take two independant n-samples $(\mathbf{x}^l)_{l=1}^n$, $(\mathbf{x}^{*,l})_{l=1}^n$
- Set $y^l = y(\mathbf{x}^l)$ and $y^{*,l} = y(x_i^l, \mathbf{x}_{-i}^{*,l})$

The estimation of S_i is

$$\hat{S}_i = \frac{\sum_{l=1}^n (y^l - \bar{y}) \cdot (y^{*,l} - \bar{y^*})}{\sum_{l=1}^n (y^l - \bar{y})^2}, \quad \bar{y} = \frac{1}{n} \sum_{l=1}^n y^l, \ \bar{y^*} = \frac{1}{n} \sum_{l=1}^n y^{*,l}$$

The numerical cost is $n(p+1)$ for first order indices

Summary

To sum up

- Sensitivity indices based on variance decomposition
- S_u, for $|u| \geq 2$ quantifies the interaction of parameters $\mathbf{X}_u$
- The closer is S_u to 1, the more influent the group $\mathbf{X}_u$ is

Advantages

- Global sensitivity measure able to detect interactions
- Clear and unambiguous definition of sensitivity
- Many techniques have been developped to estimate Sobol indices

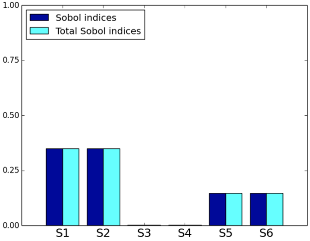
Drawbacks

- The estimation can be very expensive
- The FANOVA expansion is not true if input parameters are dependent

The simulation of sustainable cities

- Output : Total treated waste
- Inputs :
  1. ACT Households 1
     1. Population (H_pop)
     2. Total Treated Waste (H_TW)
  2. ACT Activities Industrials W1
     1. Population (AI_pop)
     2. Total Treated Waste (AI_TW)
  3. ACT Tertiary Activities W1
     1. Population (AT_pop)
     2. Total Treated Waste (AT_TW)

First-order Sobol indices



User's guide

The aim of the sensibility analysis is to...

	Screening	Local	Global method
Rank variables	x	x	x
Quantify sensitivity		x	x
Look at around a nominal value		x	
Explore the whole inputs space	x		x

What to do if the model is/has ...

	Expensive	Large p
Expensive	Sobol[2] + meta-modeling[4]	
Large p	Morris[1]	Morris[1] + Sobol

Bibliography

R. Faivre, B. Iooss, S. Mahévas, D. Mokowski, and H. Monod.
Analyse de sensibilité et exploration de modèles.
Quæ, France, 2013.

A. Saltelli, K. Chan, and E.M. Scott.
Sensitivity Analysis.
Wiley, West Sussex, 2000.

T.J. Santner, B.J. Williams, and W.I. Notz.
The design and analysis of computer experiments.
Springer Science & Business Media, 2013.

B. Sudret.
Global sensitivity analysis using polynomial chaos expansion.
Reliability engineering and system safety, 93(7):964–979, 2008.

W. Tinsson.
Plans d'expérience: constructions et analyses statistiques, volume 67.
Springer Science & Business Media, 2010.