Introduction to Geostatistics - Metamodelling with Gaussian processes

Bertrand Iooss
Main stakes of uncertainty management

• **Modeling phase**
  - Improve the model
  - Explore the best as possible different input combinations
  - Identify the predominant inputs and phenomena in order to prioritize R&D

• **Validation phase**
  - Reduce prediction uncertainties
  - Calibrate the model parameters

• **Practical use of a model**
  - Safety studies: assess a risk of failure (rare events)
  - Conception studies: optimize system performances and robustness

**Advantages of a probabilistic approach**

• to propagate the uncertainties, to perform global sensitivity analysis
• to design/optimize the system taking into account uncertainties
• to give rigorous safety margins, ...
Uncertainty management - The generic methodology

**Step A: Problem specification**
- **Input variables**
  - Uncertain: \( x \)
  - Fixed: \( u \)
- **Model**
  (or measurement process)
  \( f(x, u) \)
- **Variables of interest**
  \( Y = f(x, u) \)

**Step B: Quantification of uncertainty sources**
- Modelisation with probability distributions
- Direct methods, statistics, expertise

**Step B': Quantification of sources**
- Inverse methods, calibration, assimilation

**Step C: Propagation of uncertainty sources**
- Quantity of interest
  - Ex: variance, probability ..
- Observed variables
  \( Y_{\text{obs}} \)

**Step C': Sensitivity analysis, Prioritization**
- Decision criterion
  - Ex: Probability < 10^{-b}
- Feedback process
  - B. looss -
Uncertainties management for cpu time consuming models

$p$ input variables

\[ X = (X_1, \ldots, X_p) \]

Physical phenomena

Computer code

Metamodel

observed experiences

simulated experiences

Predicted experiences

Use of the metamodel:

- C': Sensitivity analysis

  Distribution of the inputs → Metamodle

  \[ Y_{SR} = f_{SR}(X) \]

  Distribution of the output

- C: Uncertainty propagation (via Monte Carlo methods)

  Identification of input parameters values

- B': Calibration

  Adequation between observed and simulated experiences
Metamodel : definition

[ Kleijnen 70's ]

A metamodel is a mathematical function

- which approximates the outputs of the model,
- with negligible cpu cost,
- which allows to make new output predictions with a good accuracy

• Synonyms:
  - Response surface
  - Simplified model
  - Emulator
  - Proxy model
  - Surrogate model
Metamodeling steps

Design of experiments:
Points to perform simulations

Simulation:
Performing simulations

Metamodelling:
Approximation of the computer code

\[ X_1 \quad X_2 \]

Computer code

Venez course 2012 – Geostatistics – B. Iooss - 6
Differents types of metamodels

- Linear regression
- Polynomials
- Splines

\[ \hat{G}(x) = \sum_{k=1}^{K} \hat{\beta}_k B_k(x) \text{ with } K \text{ the number of nodes} \]

- Additive models, GAM
  \[ \hat{G}(x) = \sum_{i=1}^{p} s_i(x_i) + \sum_{i<j}^{p} s_{ij}(x_i, x_j) + \cdots \]

- Regression trees
  \[ \hat{G}(x) = \sum_{k=1}^{K} \hat{\beta}_k I_k(x) \]

- Neural networks
- Chaos polynomials
- Support Vector Machines
- Kriging – Gaussian process

[Simpson et al. 2001]
[Storlie & Helton 2008]
Kriging metamodell

Kriging [Matheron 63] for computer codes relies on the idea to interpolate the code outputs in dimension $p$ [Sacks et al. 89] as a spatial cartography.

Kriging (or Gaussian process) is interesting because:
- it interpolates the outputs,
- it gives predictor associated with confidence bands.

Example in 1D:

Theoretical function ($p=1$):

$$Y = f(X) = \sin(X)$$

Simulation of $N=7$ computation points.

[Chevalier, 2011]
Introduction to Geostatistics
Introduction to Geostatistics

Objectives: treatment of numerical data with spatial support (or temporal) with uncertainty quantification

Principal aspects:

• Taking into account the spatial structure of data,
• Dimension 1, 2, 3, …,
• Irregular sampling,
• Integrating external information

2 types of methods:

• Estimation (prediction, …) at a given point

• Simulations reproducing the variability of the phenomenon
Example: porosity of a geological medium

- Reality
- Sample
- Optimal prediction (unique)
- One conditional simulation

[Chilès]
Spatial statistics: kriging interpolation

Linear combination of \( N \) data:
\[
Y^*(u) = \sum_{i=1}^{N} \lambda_i Y(u_i)
\]

Kriging can take into account the data configuration, the distance between data and target, the spatial correlations and potential external information.

Probabilistic framework

**Estimation without bias:**
\[
E [ Y^*(u) - Y(u) ] = 0
\]
The mean of the errors is zero.

**Estimation \( Y^*(u) \) is optimal:**
\[
\text{Var}[ Y^*(u) - Y(u) ] \text{ is minimal}
\]
The dispersion of the errors is reduced.
Stochastic model for $Y(x)$

The random field $Y(x)$, with $Y \in \mathbb{R}$ and $x \in \mathbb{R}^p$, is characterized by its mean and its covariance.

$Y(x)$ is stationary of second order:

1. $E[Y(x)] = m$ does not depend on $x$
2. Covariance: $\text{Cov}[Y(x), Y(x+h)] = E[Y(x+h)Y(x)] - E[Y(x+h)]E[Y(x)] = C(h)$

This covariance does not depend on $x$.

\[ C(h) \]

\[ \text{Covariance function} \]

\[ h \]

\[ \text{variance} \]
In practice

\[ C(h) = \sum_{i=1}^{N(h)} [Y(x_i + h)Y(x_i)] - m^2 \]

Var[\(Y(x)\)] = \(C(0)\)

\(Y(x) \rightarrow Y(x+h)\)

\(Y(x) \rightarrow Y(x+2h)\)

Covariance function

Range = maximal distance of correlation = correlation length
Examples of stochastic processes (Gaussian)

1D

[from: Marcotte]

2D

[from: Baig, 2003]
The variogram

- **Nugget effect**: (measurement error, microstructure) 0
- **Slope**: (spatial continuity of the phenomenon)
- **Range**: heterogeneity scale
- **Experimental variogram**: variogramme (model)

Venez course 2012 – Geostatistics – B. Iooss - 16
Simple kriging (known mean)

\[ Y^*(u) = \sum_{i=1}^{N} \lambda_i(u) [Y(u_i) - m] + m \]  
\( (m = \text{known constant}) \)

Min \{ \mathbb{E} [Y^*(u) - Y(u)]^2 \}
\( \lambda_i \rightarrow \text{multiple linear regression by least squares} \)

Best Linear Unbiased Predictor (BLUP)

Kriging weights \( \lambda_i(u) \) for \( Y(u_i) \) are obtained by:

\[
\begin{align*}
\sum_{j=1}^{N} \lambda_j(u) C(u_i - u_j) &= C(u_i - u) \quad \forall i = 1 \ldots N \\
\end{align*}
\]

System of \( N \) linear equations with \( N \) unknowns which have an unique solution (for non singular covariance matrix)

Kriging variance (estimation error):

\[ \sigma^2_K(u) = C(0) - \sum_{i=1}^{N} \lambda_i(u) C(u_i - u) \]

does not depend on the \( Y \) values

=> Visualisation of regions with imprecise estimations
=> Put new observation points in these regions
Example: cartography of air pollution

73 measures of benzene concentration (Rouen, France)
[from: Bobbia, Mietlicki & Roth, 2000]

\[ \gamma(h) = C(0) - C(h) \]
Simulations

- **Kriging give the optimal estimation** (unbiased, minimal error variance) of the variable at any point, from experimental data.

- **A simulation represents a possible realization** of the real phenomenon. It reproduces its true **variability** (distribution, variogram), with respect to experimental data (**conditional simulation**).

Main goal of simulation: **quantify the uncertainty via sampling** (as Monte Carlo).

Numerous methods of random fields simulation (LU decomposition, turning bands method, spectral method, Karhunen-Loève, etc.).
Kriging of observation + Non cond. Simul. - Kriging of non cond. simul = Conditional simulation
Example: profile of ocean bottom (1/5)

You have to put a cable on the ocean bottom

Question: what is the length of the cable?
Example : profile of ocean bottom (2/5)

The exact depth is uniquely known at the observation points (survey)
Example: profile of ocean bottom (3/5)

Kriging of the ocean depth

The true length is 110 km while kriging gives 104.6 km => some cable is missing
Example: profile of ocean bottom (4/5)

Another approach: the conditional simulations
Example: profile of ocean bottom (5/5)

The 95%-confidence interval from conditional simulations is $[108.8, 113.5]$

Same problem for probability of failure estimation (non linear transfer fct)

Venez course 2012 – Geostatistics – B. Iooss
Gaussian process metamodel
Gaussian process metamodel (1/2)

- **Idea:** Computer code results are interpolated with the kriging technique
- **Necessary hypothesis:** Gaussian process

**Definition:**
\[ Y(x) = \beta F(x) + Z(x) \]

- **Regression** stochastic part

**Parametric choices:**
- \( F \): polynomial of degree 1
  \[ \beta F(x) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i \]
- \( R \): stationary \( \Rightarrow \) covariance function

**Example:** Gaussian covariance
\[ R(x, u) = R(x - u) = \exp\left(-\sum_{i=1}^{p} \theta_i |x_i - u_i|^2\right) \]

**Anisotropy:** \( \theta_i \)'s are not equal (correlation length of each input variable)

Stochastic process \( Z \) with:
\[ E[Z(x)] = 0 \]
\[ \text{Cov}(Z(x), Z(u)) = \sigma^2 R(x, u) \]
where \( \sigma^2 \) is the variance and \( R \) the correlation function
\[ Z \sim N(0, \sigma^2 R) \]
Gaussian process metamodel (2/2)

- **Joint distribution:**

  - Gaussian process (Gp) model: \( Y(x) = \beta F(x) + Z(x), \; x \in \mathbb{R}^p \)
  
  - Learning sample (LS) of \( N \) simulations: \( (X_{LS}, Y_{LS}) \)

\[
X_{LS} = (x^{(1)}, \ldots, x^{(N)}), \quad F_{LS} = F(X_{LS}), \quad R_{LS} = (R(x^{(i)}, x^{(k)}))_{i,k} \\
Y_{LS} \sim \mathcal{N}(\beta F_{LS}, \sigma^2 R_{LS})
\]

- Conditional Gp metamodel:

\[
Y(x)_{|X_{LS}, Y_{LS}} \sim \text{Gp}
\]

\[
\text{Mean: } \hat{Y}(x) = \mathbb{E}[Y(x)|_{X_{LS}, Y_{LS}}] = \beta F(x) + r(x) R_{LS}^{-1} [Y_{LS} - \beta F_{LS}]
\]

\[
\text{with } r(x) = [R(x^{(1)}, x), \ldots, R(x^{(N)}, x)]
\]

\[
\text{Covariance: } \text{Cov}(Y(u)|_{X_{LS}, Y_{LS}}, Y(v)|_{X_{LS}, Y_{LS}}) = \sigma^2 (R(u,v) + r(u) R_{LS}^{-1} r(v))
\]

\( \Rightarrow \text{Variance} \Rightarrow \text{Mean Square Error (MSE)} \)
Conclusion: given a sufficient number of points, we obtain an accurate metamodel
Hyperparameters estimation

- **Maximum likelihood method**

  - Likelihood maximisation on the learning basis $(X_s, Y_s)$:
    \[
    (\beta^*, \theta^*, \sigma^*) = \text{Argmax}_{(\beta, \theta, \sigma)} \ln L(Y_{LS}, \beta, \theta, \sigma)
    \]

  - with
    \[
    \ln L(Y_{LS}, \beta, \theta, \sigma) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \ln(\det R_{LS}) - \frac{1}{2} \sigma^2 \left[ Y_{LS} - \beta F_{LS} \right] R_{LS}^{-1} \left[ Y_{LS} - \beta F_{LS} \right]
    \]

  - Joint estimation of $\beta$ and $\sigma$:
    \[
    \begin{align*}
    \beta^* &= \left[ t F_{LS} R_{LS}^{-1} F_{LS} \right]^{-1} t F_{LS} R_{LS}^{-1} Y_{LS} \\
    \sigma^2 &= \frac{1}{N} t \left[ Y_{LS} - \beta^* F_{LS} \right] R_{LS}^{-1} \left[ Y_{LS} - \beta^* F_{LS} \right]
    \end{align*}
    \]

  - Estimation of correlation parameters $\Theta$:
    \[
    (\theta^*) = \text{Argmin}_{\theta} \psi(\theta) \quad \text{with} \quad \psi(\theta) = \left| R_{LS} \right|^{-1/N} \sigma^2
    \]
Estimation and validation

\[ R(u, v) = R(u - v) = \exp \left( - \sum_{i=1}^{p} \theta_i |u_i - v_i|^2 \right) \]

- Hyperparameters \( \theta_i \) estimated by likelihood maximization

Simplex method, stochastic algorithms

Problems in high dimensional context \( (p > 10) \), can be solved by sequential algorithms [Marrel et al. 2008]

- Predictor validation:
  Predictivity coefficient

\[ Q_2(Y, \hat{Y}) = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y - Y_i)^2} \]

- Test sample
- or leave-one-out
- or \( k \)-fold cross validation

MSE validation: Percentage of predicted values inside confidence bounds
Effects of the hyperparameters $\theta$ and $\sigma$

$$f(x) = \sin(4\pi x)$$

$\sigma^2 = 1; \theta = 0.2$

$\sigma^2 = 1; \theta = 10^{-4}$

$\sigma^2 = 4; \theta = 0.2$

$\sigma^2 = 4; \theta = 10^{-4}$

Venez course 2012 – Geostatistics – B. Iooss
Effects of the covariance structure

\(\text{gauss}\)

\(\sin(x)\)

\(x\)

\(2\quad 4\quad 6\quad 8\quad 10\)

\(2\quad 1\quad 0\quad -1\quad -2\)

\(\text{exponential}\)

\(\sin(x)\)

\(x\)

\(2\quad 4\quad 6\quad 8\quad 10\)

\(2\quad 1\quad 0\quad -1\quad -2\)

\(\text{matern5}_2\)

\(\sin(x)\)

\(x\)

\(2\quad 4\quad 6\quad 8\quad 10\)

\(2\quad 1\quad 0\quad -1\quad -2\)

\(\text{matern3}_2\)

\(\sin(x)\)

\(x\)

\(2\quad 4\quad 6\quad 8\quad 10\)

\(2\quad 1\quad 0\quad -1\quad -2\)

[ Chevalier, 2011 ]
Adaptive designs using Gaussian process metamodel
The best way to build Gp: model-based adaptive designs

Example: criterion of the Gaussian process MSE (Mean Square Error)

\[
MSE(x) = \sigma^2 + r(x)R_{LS}^{-1}r(x) + u(x)\left(\beta F_{LS}R_{LS}^{-1}\beta F_{LS}\right)^Tu(x)
\]

\[u(x) = \beta F(x) - k(x)R_{LS}^{-1}\beta F_{LS}\]

\[x_{new} = \arg \max_{x \in D} MSE(x)\]

Remark: other criteria are possible (e.g. focusing to active variables)

Conclusion: Model-based adaptive designs are the most efficient ones, but are not always applicable.

In practice, we need to initiate the process with a space-filling design.
Estimation of rare events probability using kriging

Industrial problems: safety analysis with computer code (nuclear, transport, ...)

Problem: find $P_f = \text{Prob} \left[ f(X) > T \right]$ with $X =$ random inputs; $T =$ threshold

Reasonable variance everywhere
Large errors in the target region

New adaptive design

$X^* = \arg \min_X (IMSE_T)$

$IMSE_T = \int MSE(x) 1_{X_T}(x) dx$

$X_T$ is a small tube around $T$

Large variance in non-target region
Good accuracy in target region

Venez course 2012 – Geostatistics – E
Optimisation of a model output using kriging

Industrial problems: conception with costly computer code (automobile, nuclear, aeronautics, ...)

Problem: find the values of $X$ which minimize the model output

$$X^* = \arg \min_{X \in D} f(X)$$

If $f$ is costly, a natural solution would be to optimize a metamodel of $f$: dangerous idea because the metamodel tends to smooth the true model.

Gp metamodel allows to take into account the metamodel error, and to define the expected improvement $EI(X)$ for each $X \in D$.

$$EI(x) = E[ \max(0, \text{observed minimum} - f(x)) ]$$
Adaptive design for optimization: EGO algorithm

[Jones et al. 1998]

EGO: step 0

[Chevalier, 2011]
Adaptive design for optimization: EGO algorithm

EGO: step 1

Chevalier, 2011
Adaptive design for optimization: EGO algorithm

EGO: step 2

kriging the sinus function

[ Chevalier, 2011 ]
Adaptive design for optimization: EGO algorithm

EGO: step 3

kriing the sinus function

[Chevalier, 2011]
Conclusions on the Gaussian process metamodel

- *Gp model construction is possible even in high dimensional case*

- *Main advantage of Gp: probabilistic metamodel which gives confidence bands in addition to a predictor*

- *Fitting quality is dependent of the initial design*
  - *Gp model is well adapted to sequential and adaptative designs*

- *Caveats: it can require a large amount of effort during the fitting process and cases with more than 1000 points begin to be difficult (matrix inversion)*

- *Designs for specific objectives (optimization, quantile, probability, etc.)*
Application
Motivating example: hydrogeological modeling

Collaboration Kurchatov Institute/CEA [ Volkova et al. 2008 ]

- Site (2 ha) near Moscow
- From 1943 to 1974: radioactive waste repositories
- 1990: site recognition with 20 piezometers
- Upper aquifer contamination in $^{90}\text{Sr}$

Questions:
- impact assessment of the contamination on the environment
- degree of rehabilitation of the site
Uncertainty management in hydrogeological modeling

- Computation of the spatio-temporal evolution of $^{90}\text{Sr}$ concentration in an ancient radwaste disposal site between 2002 and 2010

- Goal: Estimate the contamination impact on the environment
  Identify the influent inputs on predicted outputs

Step A

Numerical modeling:
Hydrogeological transport (Darcy’s law) scenario of $^{90}\text{Sr}$ with the MARTHE software

$\rho = 20$ uncertain inputs $X$: Permeability, dispersivity, $K_d$, infiltration intensity, ...

Output of interest: Concentration values $Y$

Quantity of interest: Distribution, variance

Goal: Estimate the contamination impact on the environment
Identify the influent inputs on predicted outputs

Venez course 2012 – Geostatistics – B. Iooss - 45
Step B: Quantification of uncertainty sources

Modelisation with probability distributions

Input variables
Uncertain: x
Fixed: u

Model (or measurement process)
f(x,u)

Variables of interest
Y = f(x,u)

Quantity of interest
Ex: variance, probability...

Step A: Problem specification

Step C: Propagation of uncertainty sources

Step C': Sensitivity analysis, Prioritization

Step B': Quantification of sources

Inverse methods, calibration, assimilation

Observed variables
Y_{obs}

Decision criterion
Ex: Probability < 10^{-b}

Feedback process
Step B: Quantification of uncertainty sources

Statistical modeling of the uncertainty of each input

Different cases:

1. A lot of data
   - Fitting of probability distributions
   - Statistical hypothesis test
   
   MARTHE case: hydraulic conductivity (lognormal distribution)

2. Few data ($n < 10$)
   - Bayesian inference

3. No data
   - Bibliography
   - Expert judgment techniques
   
   MARTHE case: dispersivity, permeability, ... (uniform distributions)
Step B: Quantification of uncertainty sources

- **Input variables of the model**

  ➔ nominal value, type of distribution and parameters

<table>
<thead>
<tr>
<th>Paramètres</th>
<th>Indicateur</th>
<th>Valeur du modèle</th>
<th>Type de distribution</th>
<th>Intervalle ou paramètres de distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Permeabilité couche 1</td>
<td>per1</td>
<td>8</td>
<td>Uniforme</td>
</tr>
<tr>
<td>2</td>
<td>Permeabilité couche 2</td>
<td>per2</td>
<td>15</td>
<td>Uniforme</td>
</tr>
<tr>
<td>3</td>
<td>Permeabilité couche 3</td>
<td>per3</td>
<td>8</td>
<td>Uniforme</td>
</tr>
<tr>
<td>4</td>
<td>Permeabilité zone 1</td>
<td>perz1</td>
<td>8</td>
<td>Uniforme</td>
</tr>
<tr>
<td>5</td>
<td>Permeabilité zone 2</td>
<td>perz2</td>
<td>8</td>
<td>Uniforme</td>
</tr>
<tr>
<td>6</td>
<td>Permeabilité zone 3</td>
<td>perz3</td>
<td>8</td>
<td>Uniforme</td>
</tr>
<tr>
<td>7</td>
<td>Permeabilité zone 4</td>
<td>perz4</td>
<td>8</td>
<td>Uniforme</td>
</tr>
<tr>
<td>8</td>
<td>Dispersivité longitudinale couche 1</td>
<td>d1</td>
<td>0,8</td>
<td>Uniforme</td>
</tr>
<tr>
<td>9</td>
<td>Dispersivité longitudinale couche 2</td>
<td>d2</td>
<td>0,8</td>
<td>Uniforme</td>
</tr>
<tr>
<td>10</td>
<td>Dispersivité longitudinale couche 3</td>
<td>d3</td>
<td>0,8</td>
<td>Uniforme</td>
</tr>
<tr>
<td>11</td>
<td>Dispersivité transversale couche 1</td>
<td>dt1</td>
<td>0,08</td>
<td>Uniforme</td>
</tr>
<tr>
<td>12</td>
<td>Dispersivité transversale couche 2</td>
<td>dt2</td>
<td>0,08</td>
<td>Uniforme</td>
</tr>
<tr>
<td>13</td>
<td>Dispersivité transversale couche 3</td>
<td>dt3</td>
<td>0,08</td>
<td>Uniforme</td>
</tr>
<tr>
<td>14</td>
<td>Coefficient de partage volumique c. 1</td>
<td>kd1</td>
<td>5,1</td>
<td>Weibull</td>
</tr>
<tr>
<td>15</td>
<td>Coefficient de partage volumique c. 2</td>
<td>kd2</td>
<td>0,34</td>
<td>Weibull</td>
</tr>
<tr>
<td>16</td>
<td>Coefficient de partage volumique c. 3</td>
<td>kd3</td>
<td>5,1</td>
<td>Weibull</td>
</tr>
<tr>
<td>17</td>
<td>Porosité tous les couches</td>
<td>poros</td>
<td>0,3</td>
<td>Uniforme</td>
</tr>
<tr>
<td>18</td>
<td>Infiltration type 1</td>
<td>i1</td>
<td>0,0001</td>
<td>Uniforme</td>
</tr>
<tr>
<td>19</td>
<td>Infiltration type 2</td>
<td>i2</td>
<td>0,004</td>
<td>Uniforme</td>
</tr>
<tr>
<td>20</td>
<td>Infiltration type 3</td>
<td>i3</td>
<td>0,02</td>
<td>Uniforme</td>
</tr>
</tbody>
</table>
Step C: Uncertainty propagation

Initial concentration

$N = 300$ runs from Latin Hypercube Sample

concentration distributions at the 20 piezometers (Bq/l)

Histogram & density - p23

Histogram & density – p31K

Venez course 2012 – Geostatistics – B. Iooss - 49
Step C': Sensitivity analysis - Sampling-based approaches

Sample \((X \in \mathbb{R}^p, Y(X) \in \mathbb{R})\) of size \(N > p\)

? Linear relation ?

\((R^2)\)

Yes

Linear regression between \(X\) and \(Y\)

Regression coefficients

? Monotonic relation ?

\((R^2*)\)

Yes

Regression on ranks

Sobol indices

? CPU time cost of the model ?

Monte Carlo

\(N > 1000p\)

Quasi-MC, FAST, RBD,

\(N > 100 p\)

Smoothing Metamodell

\(N > 10 p\)

negligible

small

large
Step C’: Regression-based approach

- Outputs with larger $R^2$ (linear relation)
  - p23 ($R^2 = 0.78$)  p104 ($R^2 = 0.68$)
  - p4-76 ($R^2 = 0.71$)

- Outputs with larger $R^2*$ (monotonic relation)
  - p4-76 ($R^2* = 0.95$)  p102K ($R^2* = 0.90$)
  - p107 ($R^2* = 0.92$)  p23 ($R^2* = 0.90$)
  - p104 ($R^2* = 0.91$)  p29K ($R^2* = 0.83$)

- Most influential inputs
  - Distribution coefficient, layer 2
  - Distribution coefficient, layer 1
  - Infiltration intensity
  - Permeability, layer 2

Problem: 14 outputs are non monotonic and we have only 300 simulation runs
Sensitivity analysis results for one scalar output « p104 »

Gaussian process (Gp) metamodel:
Predictivity coefficient: $Q_2 = 93\%$ - Linear regression : $Q_2 = 68\%$

Sobol indices estimation + confidence intervals

<table>
<thead>
<tr>
<th>(en %)</th>
<th>$\text{SRC}_{i}^2$ (linear regression)</th>
<th>$\mu_i = \mathbb{E}[\tilde{S}_i]$ Gaussian process</th>
<th>IC- 90% ($\tilde{S}_i$) Gaussian process</th>
</tr>
</thead>
<tbody>
<tr>
<td>per1</td>
<td>2</td>
<td>8</td>
<td>[ 5 ; 11 ]</td>
</tr>
<tr>
<td>kd1</td>
<td>52</td>
<td>69</td>
<td>[ 56 ; 83 ]</td>
</tr>
<tr>
<td>i3</td>
<td>13</td>
<td>13</td>
<td>[ 10 ; 17 ]</td>
</tr>
</tbody>
</table>

This can be done for several outputs ...
Step C’: sensitivity analysis for 20 scalar outputs

Main influent inputs

- **Group 1**: kd1 (distribution coef. of layer 1)
- **Group 2**: kd2 (distribution coef. of layer 2)
- **Group 3**: i3 (infiltration intensity)

... but the results are difficult to synthesize, then to interpret
We have simulated $N = 300$ maps (Monte Carlo runs) with $p = 20$ random inputs. Considering functional (spatial) output, concentration in $^{90}\text{Sr}$ predicted in 2010.

Discretized spatial output can be considered as a functional 2D output.
Sensitivity analysis when model outputs are functions

**Elementary cases** (sensitivity analysis on each scalar output):

- Very small CPU time consuming model
- Linear or monotonic model

**Difficult cases:**

- Complex/Non linear model → need of Sobol indices
- CPU time expensive model → need of metamodel
Sensitivity analysis for spatial outputs: methodology

- **Computer code** $f(.)$:
  - Input: $X = (X_1, \ldots, X_p)$ random vector
  - Output for input $x^* \mapsto y = f(x^*, z)$, $z \in D_z \subset \mathbb{R}^2$

In practice, $D_z$ is discretized in $n_z$ points (here: $64 \times 64 = 4096$ points)

- **Decomposition of $Y(z)$ on an orthogonal function basis** (fixed basis)
  - For example, a wavelet basis is well-suited if there are discontinuities

- **Modeling of the decomposition coefficients by a Gp metamodel**
  - Selection procedure of the most important coefficients

- **Prediction**: $x^* \mapsto$ prediction of coeff. $\Rightarrow$ spatial output map reconstruction

**Sensitivity analysis**: Spatial maps of sensitivity indices

Venez course 2012 – Geostatistics – B. Iooss - 56
Metamodel fitting: methodology for spatial output

- **Step 1**: Wavelet decomposition of each map (300 maps)
- **Step 2**: Kriging metamodelling of the main wavelet coefficients (the most variables) in function of X; constant for the other coefficients.
- **Step 3**: Prediction for a new input \( x^* \)
  \[ x^* \Rightarrow \text{prediction of the coefficients} \Rightarrow \text{spatial output map reconstruction} \]
- Example of use: Sensitivity analysis (spatial map of sensitivity indices)

**Sensitivity indices (Sobol indices, based on functional ANOVA):**

\[
S_i = \frac{\text{Var}_{X_i}[E(Y|X_i)]}{\text{Var}(Y)}; \quad S_{ij}; \quad ...; \quad S_{Ti} = S_i + \sum_j S_{ij} + \sum_{j,k} S_{ijk} + ...
\]

1st order 2nd order Total indices

The estimation of Sobol indices require a large number of simulations

computation of the Sobol indices by the way of the metamodel
Application on our test case (hydrogeological pollution)

$N = 300$ simulations

$p = 20$ random input variables

$K = 4096$ pixels

$k = 100$ wavelet coefficients modeled by Gaussian process

Mean predictivity (functional metamodel): $Q_2 = 72\%$

Estimation of first order and total Sobol’ indices maps by Monte Carlo (22000 runs with the functional metamodel)

20 maps of sensitivity indices
Spatial output: results of sensitivity analysis

Spatial maps of Sobol sensitivity indices of first order, for 6 inputs

Permeability layer 1

Permeability layer 2

Permeability layer 3

Kd layer 1

Kd layer 2

Strong infiltration intensity

Venez course 2012 – Geostatistics – B. Iooss - 59
Bibliography

• Chilès & Delfiner, Geostatistics, Wiley, 1999

• Fang, Li & Sudjianto, Design and modeling for computer experiments, Chapman, 2006

• Hastie, Tibshirani & Friedman, The elements of statistical learning theory, Springer, 2002

• Kleijnen, Design and analysis of simulation experiments, Springer, 2008

• Marcotte, Cours de l’Ecole Polytechnique de Montreal

• Marrel et al., Global sensitivity analysis for models with spatially dependent outputs, Environmetrics, 2011
