Traitement des incertitudes en simulation numérique

Cours 3 : Krigeage et modélisation d'expériences numériques

Bertrand looss

Module INSA Toulouse/GMM 5

Planification, risque et incertitudes

9 janvier 2013

EDF
CHANGER L'ÉNERGIE ENSEMBLE

Module Traitement des incertitudes en simulation numérique

Depuis une trentaine d'années, l'industrie a développé des processus et des codes de calcul parfois très lourds pour modéliser des phénomènes complexes !
La plupart des ingénieurs sont amenés à manipuler ces codes & processus

1) Il est nécessaire d'**optimiser leur utilisation pour prendre des décisions !**
=> *Analyse de sensibilité, planification d'expérience, développement de modèles réduits*

2) La **validation de leurs résultats** est un problème crucial lorsqu'ils sont utilisés dans des cycles industriels (conception, sûreté, prévision, etc.)
=> *Gestion des incertitudes, calculs fiabilistes*

3 cours de 3h15 pour INSA GMM 5 & Master Pro 2 UPS
1. Cours 1 : Introduction, modélisation et propagation d'incertitudes
2. Cours 2 : Planification et analyse d'expériences numériques
3. Cours 3 : Modélisation d'expériences numériques, krigeage

3 séances de TP pour INSA GMM 5
1. TP 1 :Exercices en R
2. TP 2 : Exercices en R
3. TP 3 : Exercices en R

Une note sera délivrée via des compte-rendus réalisés à l'issue des TPs

Plan du cours 3

1. **Introduction**
2. Méthode d'interpolation spatiale par krigeage
3. Le métamodèle « processus gaussien »
4. Un exemple d'application en hydrogéologie

Approches quantitatives : schéma générique introductif

Étape C : Propagation des sources d'incertitude

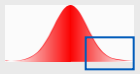
Étape A : Spécification du problème

Variables d'entrée
Incertaines : $\underline{x}$
Fixées : $\underline{d}$

Modèle
(ou processus de mesure)
$G(\underline{x},\underline{d})$

Variables d'intérêt
$Z = G(\underline{x},\underline{d})$

Quantité d'intérêt
Ex: variance, probabilité ..



Étape B: Quantification des sources d'incertitudes

Modélisation par des distributions

Étape C' : Analyse de sensibilité, Hiérarchisation

Rebouclage
(feedback)

Critère de décision
Ex: Probabilité < 10^{-b}

Rappels sur la planification d'expériences numériques

Enjeu : Echantillonner un espace de grande dimension de manière « optimale » (obtenir le plus d'informations possible sur le comportement de la sortie Z / $\mathbf{X} \in R^p$)

Problème : un échantillon aléatoire pur (Monte Carlo) remplit mal l'espace (surtout si p est élevé)

1. Plans « space filling » sont de bons candidats pour bien remplir l'espace :

   - Basés sur un critère de distances entre les points du plan (minimax, maximin, …) ; *ces plans ont une justification théorique pour la construction du métamodèle de krigeage (objet de ce cours)*

   - Basés sur un critère de répartition uniforme des points (discrépance) ; *ces plans ont une justification théorique lorsque l'on calcule la moyenne de la fonction G (**X**)*

2. Propriété de projections uniformes sur les marges peut être obtenue via les plans hypercubes latins (LHS)

3. Il est possible de coupler les 2 propriétés en construisant des LHS optimisés

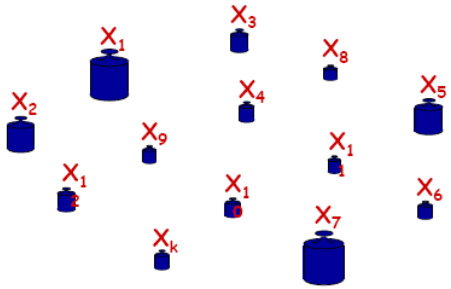
Rappels sur l'analyse de sensibilité

Enjeu : décomposer la variabilité globale de la sortie $Z = G(\mathbf{X})$ (due aux incertitudes sur les entrées $\mathbf{X}$) en part de variabilité due à chaque entrée X_i, $i=1,\ldots,p$

Problème : comme pour la planification, le coût en nombre N d'évaluations de $G(.)$ dépend de p

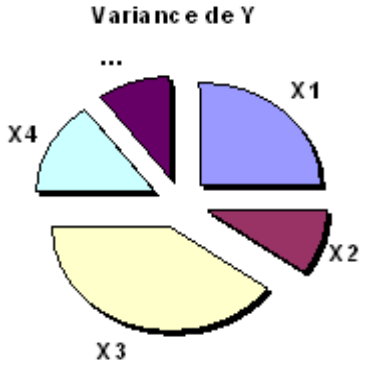
1. Le criblage (screening) :
   - plans d'expériences classiques,
   - plans d'expériences numériques

   $N \sim p/2$ à $10\,p$

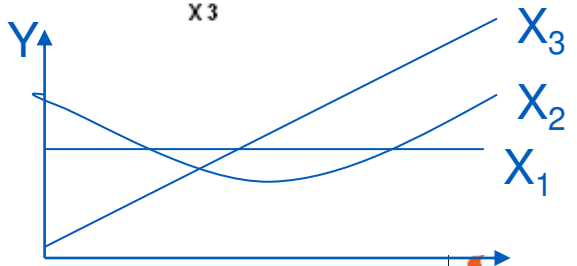


2. Les mesures d'influence globale :
   - corrélation/régression sur les valeurs/rangs,
   - décomposition de la variance fonctionnelle (Sobol),

   $N \sim 2p$ à $1e4\,p$



3. Exploration fine des sensibilités - $N \sim 10p$ à $100\,p$
   - Méthodes de lissage (param./non param.)
   - Métamodèles => objet de ce cours



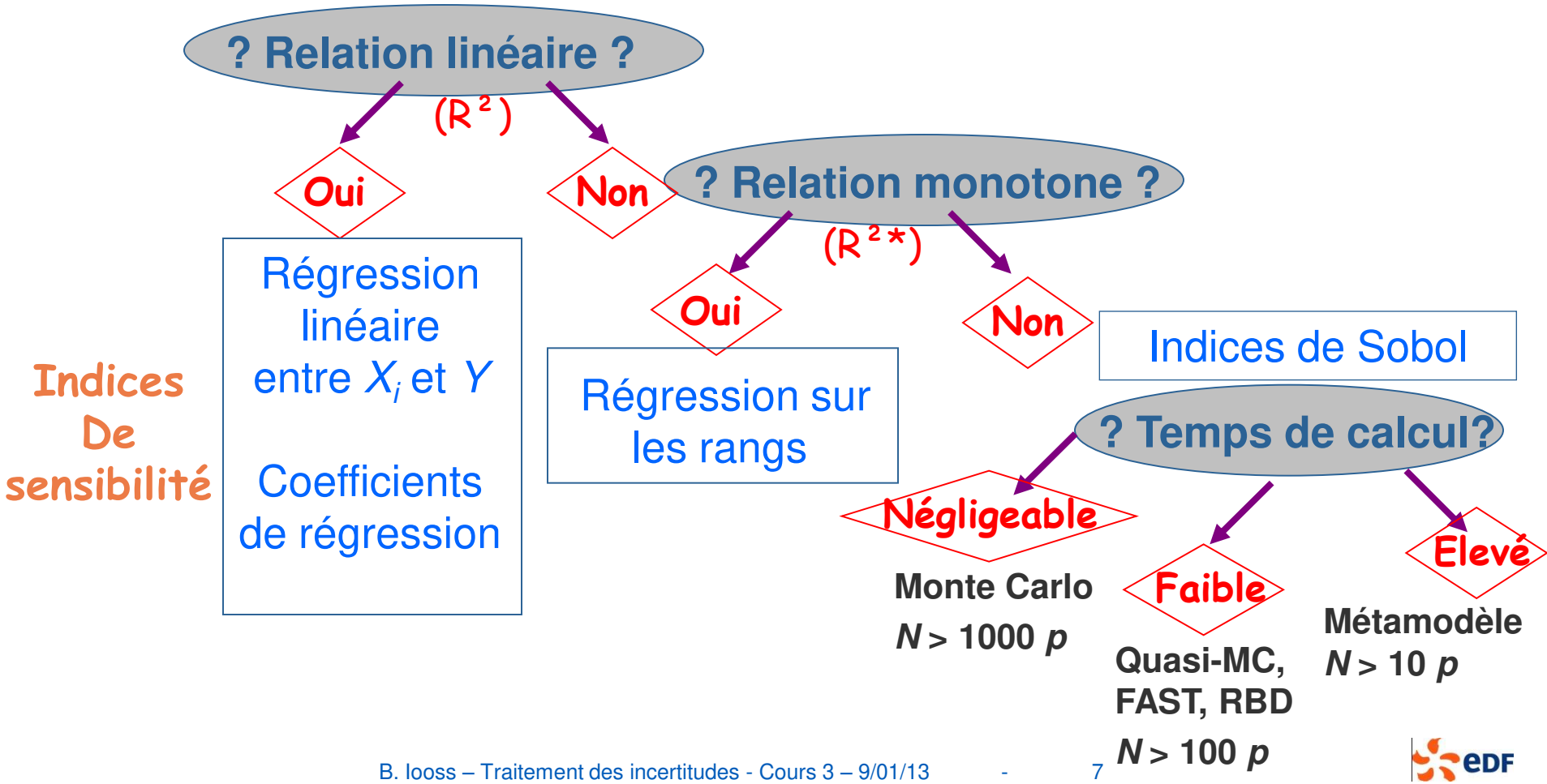
Analyse de sensibilité pour 1 sortie scalaire

Échantillon ($\mathbf{X}$, $Y(\mathbf{X})$) de taille $N > p$, de préférence de taille $N \gg p$

Étape préliminaire : visualisation graphique (par ex : scatterplots)

Méthodologie d'analyse de sensibilité quantitative

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



Métamodèle : définition

[Kleijnen 70's]

Un métamodèle est une fonction mathématique :

- **approximant les réponses du modèle étudié,**
- **de coût négligeable,**
- **permettant de prédire avec une bonne précision de nouvelles réponses**

Synonymes :

- Surface de réponse
- Modèle simplifié
- Émulateur
- Modèle proxy
- Surrogate model

Les étapes de la métamodélisation

[Le Gratiet, 2011]

Planification d'expériences :

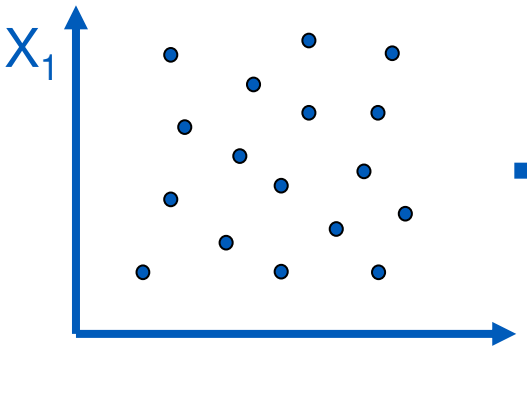
On choisit les points X où l'on va effectuer les simulations.

Simulation :

On effectue les simulations.

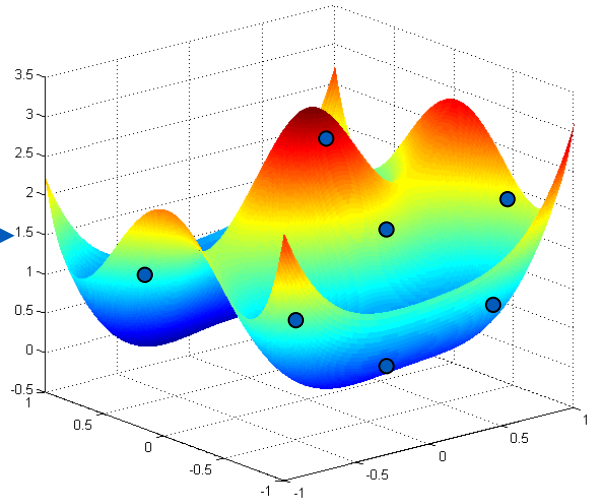
Méta-modélisation :

Approximation de la surface de réponse du code.



Code de calcul

Expérience

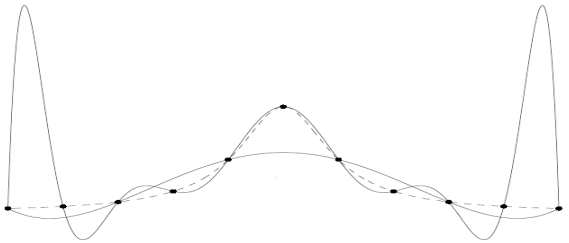


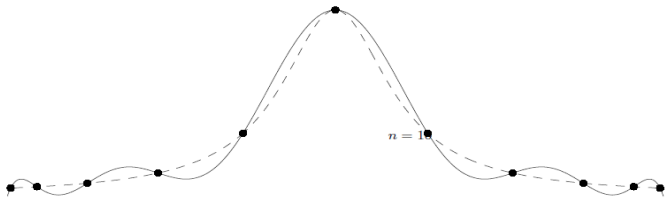
- Méthode : On va choisir parmi une famille de fonctions, la fonction qui est la plus proche de la fonction objectif

Différents types de métamodèles

- Modèle linéaire

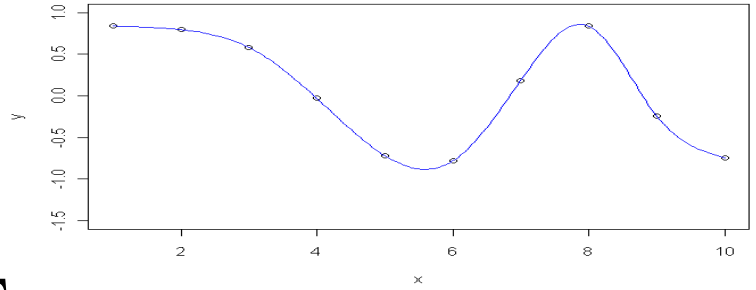
- Polynômes





- Splines

$$\hat{G}(\mathbf{x}) = \sum_{k=1}^{K} \hat{\beta}_k B_k(\mathbf{x}) \text{ avec } K \text{ le nb de noeuds}$$



- Modèles additifs, GAM $\hat{G}(\mathbf{x}) = \sum_{i=1}^{p} s_i(x_i) + \sum_{i<j}^{p} s_{ij}(x_i, x_j) + \cdots$

- Arbres de régression $\hat{G}(\mathbf{x}) = \sum_{k=1}^{K} \hat{\beta}_k I_k(\mathbf{x})$

- Réseaux de neurones
- Polynômes de chaos
- Régression à vecteur de supports
- Krigeage

Les polynômes de chaos

Base des polynômes du chaos = famille de polynômes orthogonaux $\psi_k(\mathbf{x})$ associés à une mesure $\mu_X(\mathbf{x})$

Ex. : densité gaussienne ➡ pol. d'Hermite ; uniforme ➡ pol. de Legendre

Le développement de chaque v.a. d'entrée sur sa base permet de représenter $Y = f(\mathbf{X})$ par tensorisation (chaos polynomial généralisé) :

$$Y(\mathbf{X}) = \sum_{k \in \mathrm{N}} \alpha_k \psi_k(\mathbf{X}) \qquad \text{avec } E[\psi_k(X)\psi_l(X)] = \delta_{k,l}$$

Limité à un ordre K de développement convenable (précision / nb de termes)

Par exemple : $p = 2$ et $K = 3$ ➡ 10 polynômes du chaos

Estimation de α_k par intégration numérique ou par moindres carrés

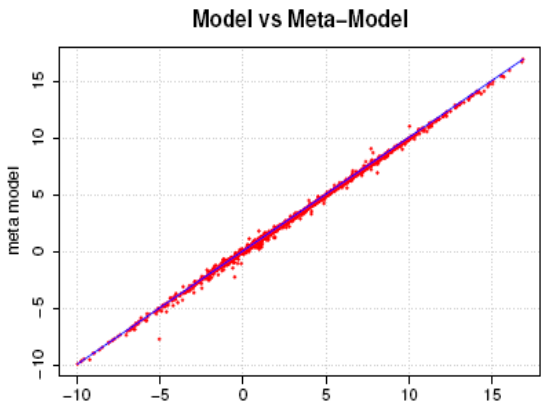
Plus : approche déterministe, taux de convergence connu, atteint un niveau arbitraire d'incertitudes, les coef. du chaos contiennent toute l'info. proba.

Minus : fléau de la dimension, adaptation des outils (en mode intrusif), problème de robustesse (non linéarités, discontinuités)

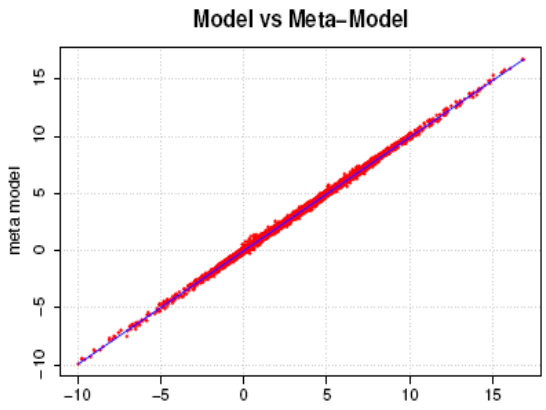
Exemple de métamodèle par Polynômes de Chaos

- $X = (X_1, X_2, X_3)$, $X_i \hookrightarrow \mathcal{U}(-\pi, \pi)$, independent
- $f(x_1, x_2, x_3) = sin(x_1) + asin^2(x_2) + bx_3^4 sin(x_1)$
- $a = 7, b = 1/10$
- budget: 500 computations

- Total degree of approx.: 9
- compute 220 coefficients



- Total degree of approx.: 12
- keep 16 best coefficients



Plan du cours 3

1. Introduction
2. **Méthode d'interpolation spatiale par krigeage**
3. Le métamodèle « processus gaussien »
4. Un exemple d'application en hydrogéologie

Introduction à la géostatistique

Objectifs : traitement de **données numériques** à **support spatial** (et/ou temporel) et **quantification des incertitudes**

Aspects principaux :

- prise en compte de la structure spatiale des données,
- espace de dimension quelconque,
- échantillonnage irrégulier et incomplet (données fragmentaires),
- prise en compte d'informations externes

2 familles de méthodes :

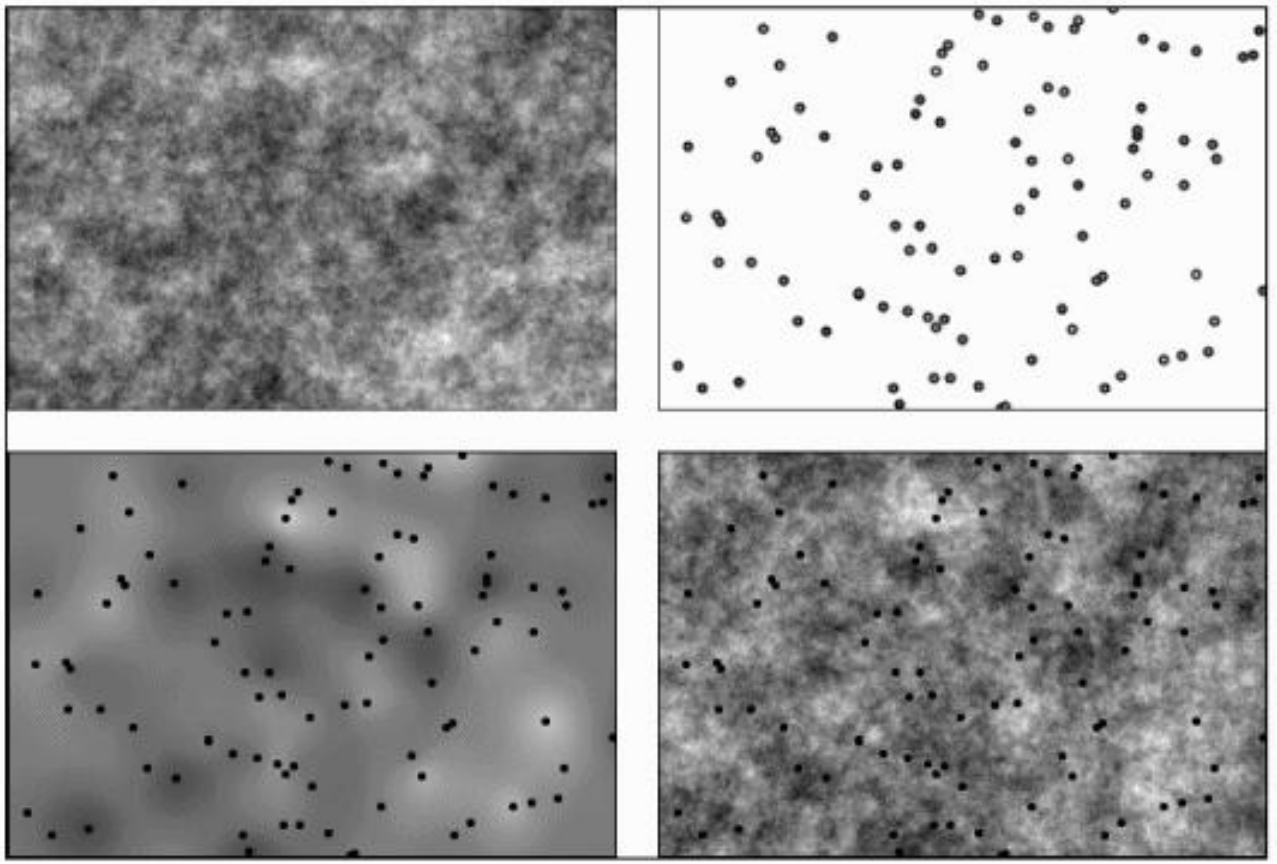
- **Estimation** (prédiction, …) en un point avec sa précision
  Estimation en tout point $\Rightarrow$ 1 carte M
  Estimation d'une moyenne dans un domaine spécifique

- **Simulations** reproduisant la variabilité des échantillons
  $\Rightarrow$ N cartes M_i équiprobables $\Rightarrow$ N calculs Monte-Carlo $Y_i = f(M_i)$

Exemple : porosité d'un milieu géologique

Réalité

Échantillonnage



Estimation optimale
(unique)

Simulation conditionnelle
(infinité de réalisations possibles)

[Chilès]

Variables, fonctions et champs aléatoires

- **Notion de variable aléatoire :**

  Une variable aléatoire associe un élément de $\Re$ à chaque épreuve aléatoire ω

  $$X : \omega \rightarrow X(\omega)$$

- **Notion de variable aléatoire vectorielle :**

  A une épreuve, on associe n éléments de $\Re$

  $$\omega \rightarrow [\, X_1(\omega)\, ,\, X_2(\omega)\, ,\, \ldots\, ,\, X_n(\omega)\,]$$

- **Notion de fonction aléatoire :**

  Variable aléatoire vectorielle à une infinité de composants

  - dénombrable : Z(1), Z(2), …, Z(i), …
  - ou non dénombrable : fonction Z(t), fonction Z(x)

- **Notion de champ aléatoire :**

  Variable aléatoire vectorielle à une infinité de composants dans un espace à plusieurs dimensions : Z(x,y) ; Z(x,y,z) ; Z(x,t)

Variable régionalisée

Modèle constitutif de la géostatistique :

- Le phénomène étudié prend des valeurs dans un domaine D spatial et/ou temporel :

  fonction $z(x)$ avec x position dans $\Re^d$ ($d = 1, 2, 3$ ou 4)

En général, on connaît $z(x)$ en quelques points x_α

- La fonction $z(x)$ est une variable régionalisée

  = réalisation d'une fonction aléatoire $Z(x)$ implantée sur D

Les données sont considérées comme ayant été générées par un processus aléatoire :

$$Z(x_\alpha) = z(x_\alpha)$$

Inférence statistique

Connaître les propriétés statistiques de Z à partir d'une réalisation unique $z(x_\alpha)$ est impossible dans le cas général

Chaque $Z(x_i)$ est une variable aléatoire qui a sa loi de probabilité

==> impossible de définir la loi de Z

⇒ **Hypothèse de stationnarité**

Z est stationnaire ⇔ sa loi est invariante par translation

(les propriétés statistiques ne changent pas si on découpe le domaine en morceaux)

⇒ **Hypothèse d'ergodicité**

(la moyenne spatiale sur D de Z tend vers son espérance qd $D \to \Re^d$)

Stationnarité d'ordre deux

Z est stationnaire d'ordre 2 si ses moments d'ordre 2 existent et sont invariants par translation :

Moyenne constante : $E[Z(x)] = m$

Variance constante : $Var[Z(x)] = \sigma^2$

Covariance ne dépend pas de x : $Cov[Z(x_i), Z(x_j)] = C(x_i - x_j)$

On dira que Z est stationnaire

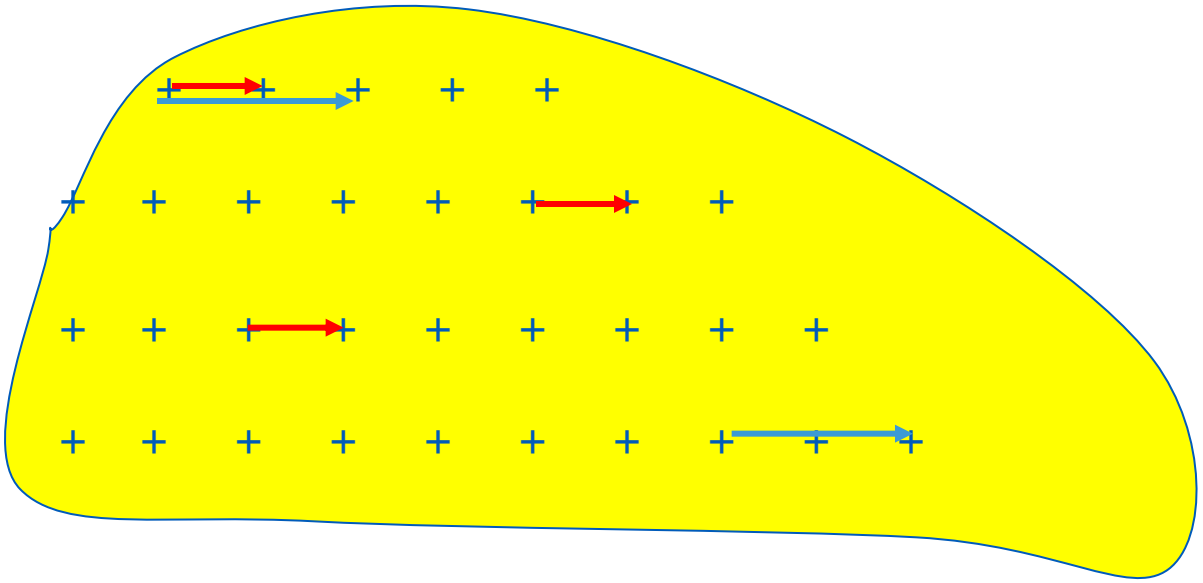
Stationnarité des variables régionalisées

- Si Z est stationnaire, tous les Z(x) ont la même loi, déterminée en utilisant les n observations $z(x_\alpha)$ (= n réalisations de Z(x))

- **En *géostatistique* linéaire, la moyenne et la covariance suffisent pour caractériser Z**

- **Stationnarité d'ordre 2 :** E[Z(x)] = m ne dépend pas de x

$$\mathrm{Cov}[Z(x), Z(x+h)] = \mathrm{E}\{Z(x+h)Z(x) - E[Z(x+h)]E[Z(x)]\} = C(h) \text{ ne dépend pas de x}$$

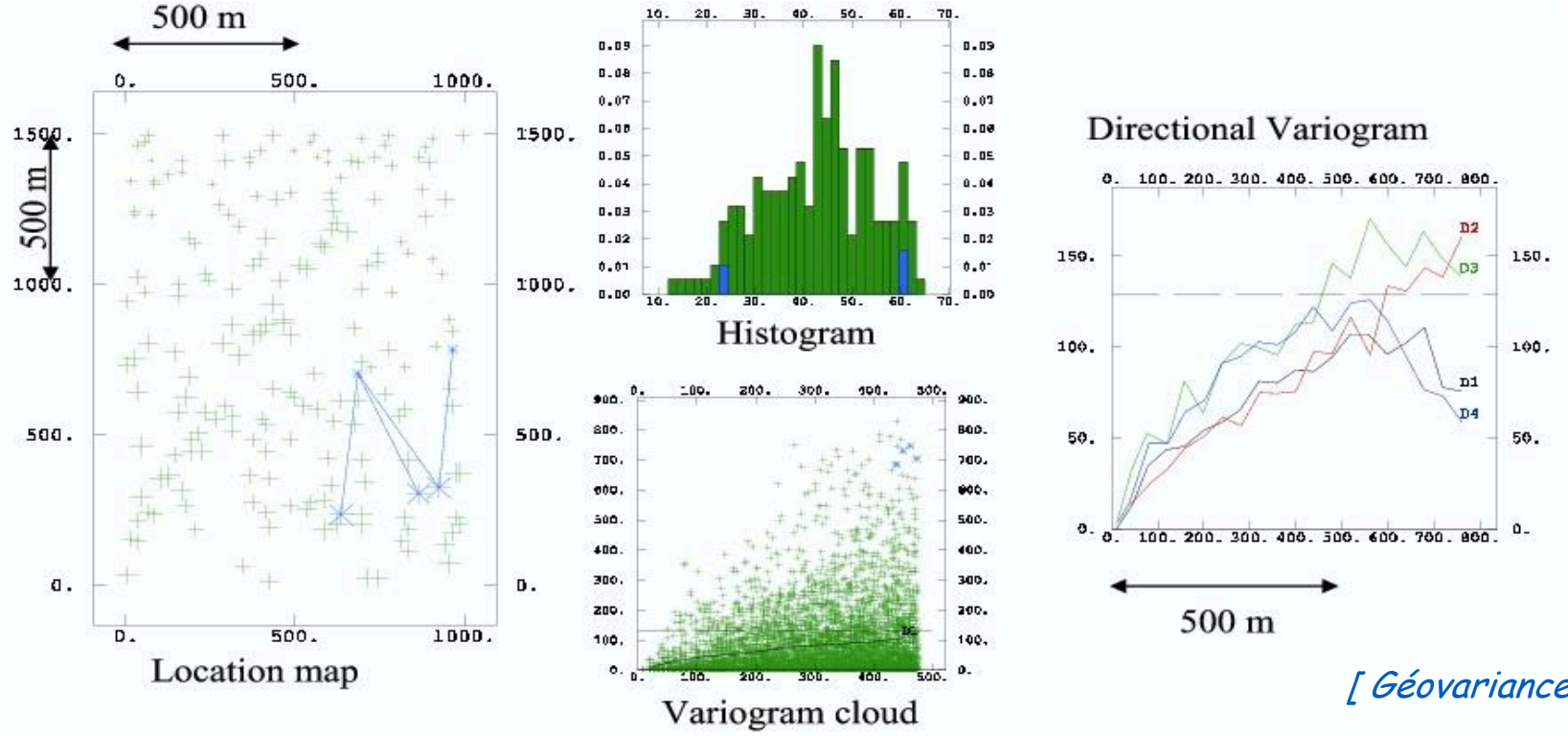


Var[Z(x)] = C(0)

Z(x) ⟶ Z(x+h)

Z(x) ⟶ Z(x+2h)

Analyse exploratoire des données



[Géovariances]

Statistiques élémentaires : moyenne, variance, médiane, histogramme, …

Nuée variographique : pour tous les couples de point (x_1,x_2), on calcule

$$\frac{1}{2}[Z(x_1)-Z(x_2)]^2$$ et on le reporte en fonction de x_1-x_2

Le variogramme

Hypothèse localement stationnaire (ou intrinsèque) :
les incréments de Z(x) sont stationnaires

Variogramme : $\gamma(h) = 0.5\ Var[\ Z(x+h) - Z(x)\ ]$

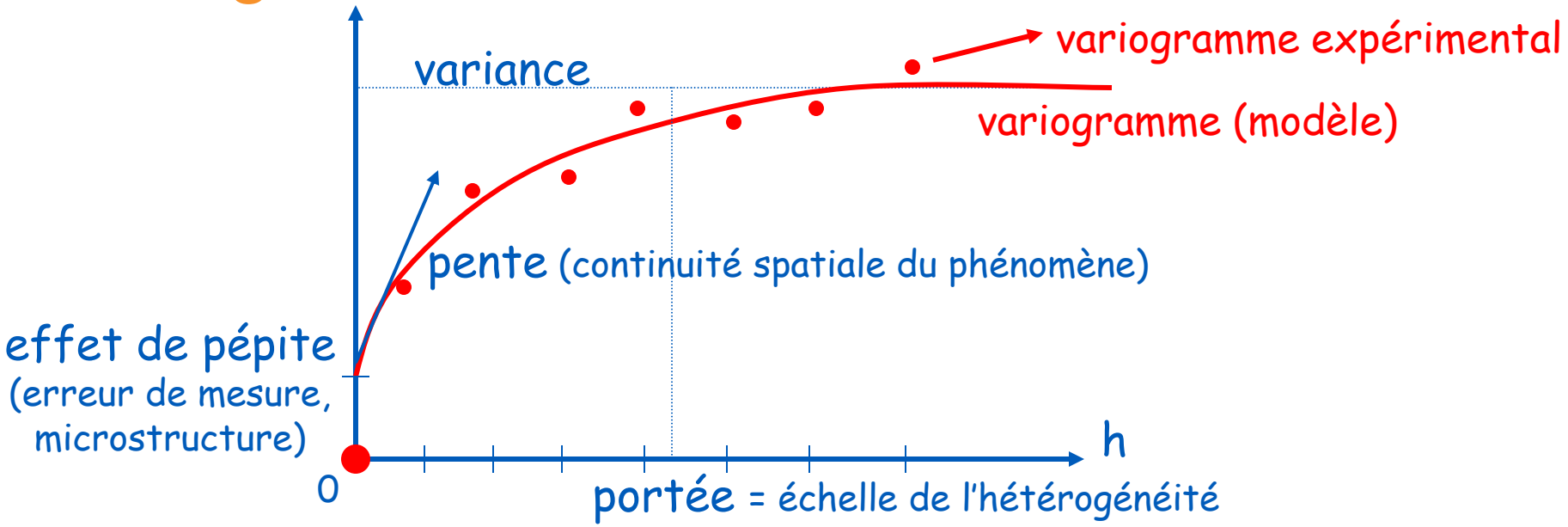
E[Z(x+h) – Z(x)] et $\gamma(h)$ caractérisent entièrement Z

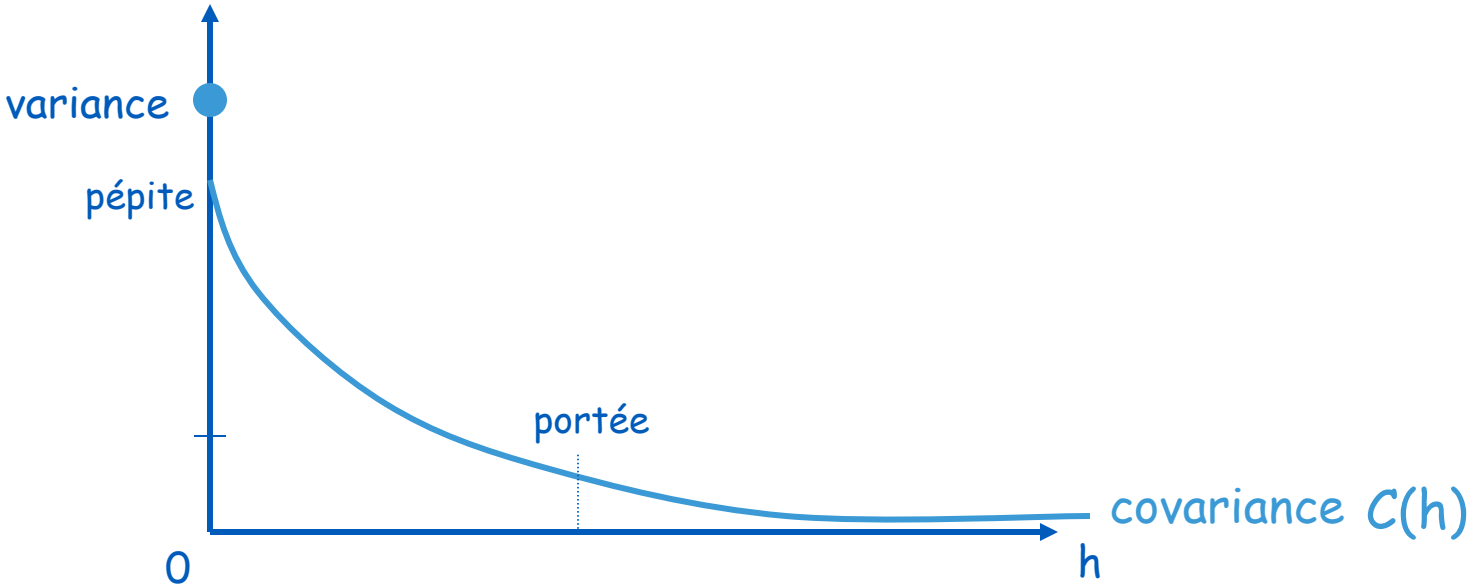
Stationnarité, $C(h)$
Moyenne et variance
de Z sont constantes

$\gamma(h) = C(0) - C(h)$

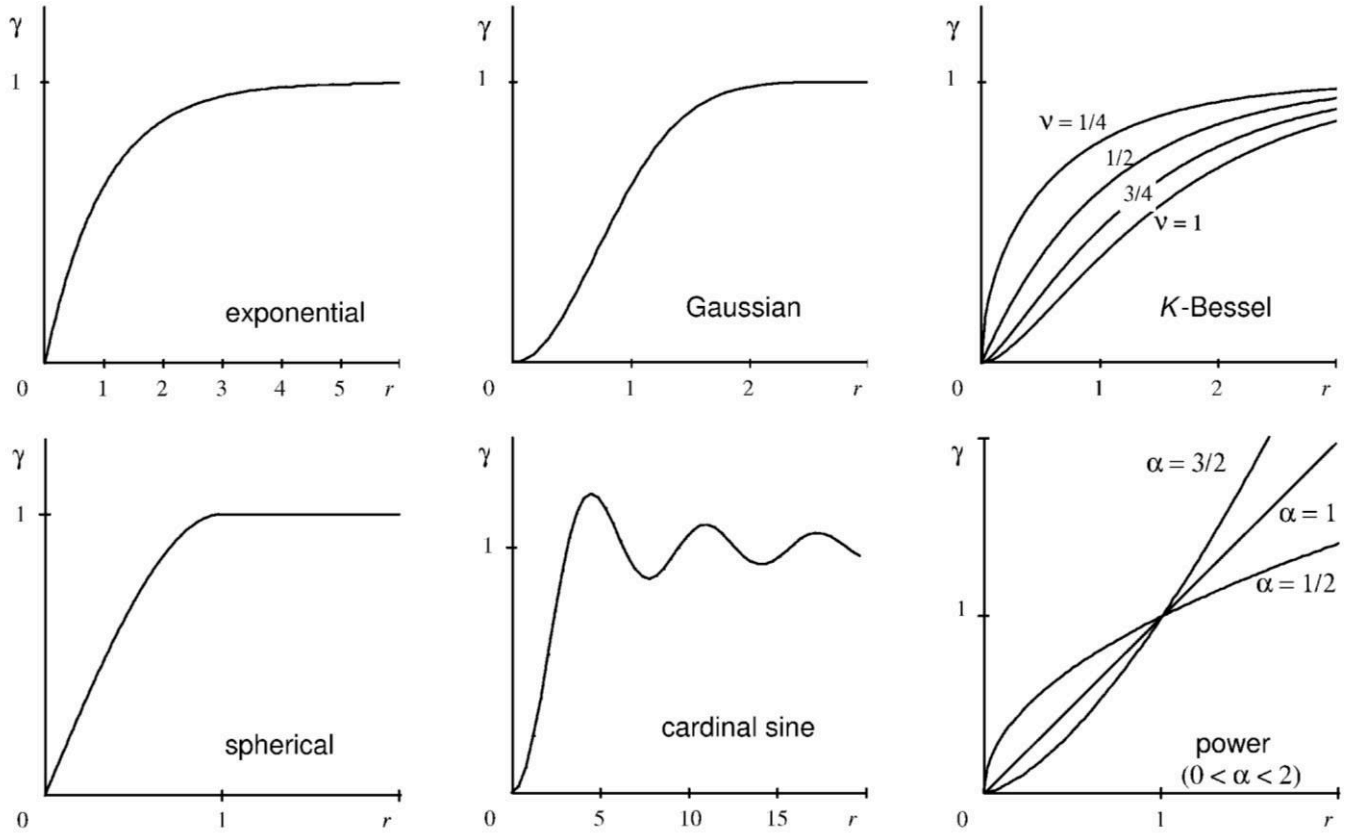
Stationnarité locale, $\gamma(h)$
Moyenne et variance
de Z sont linéaires

Le variogramme





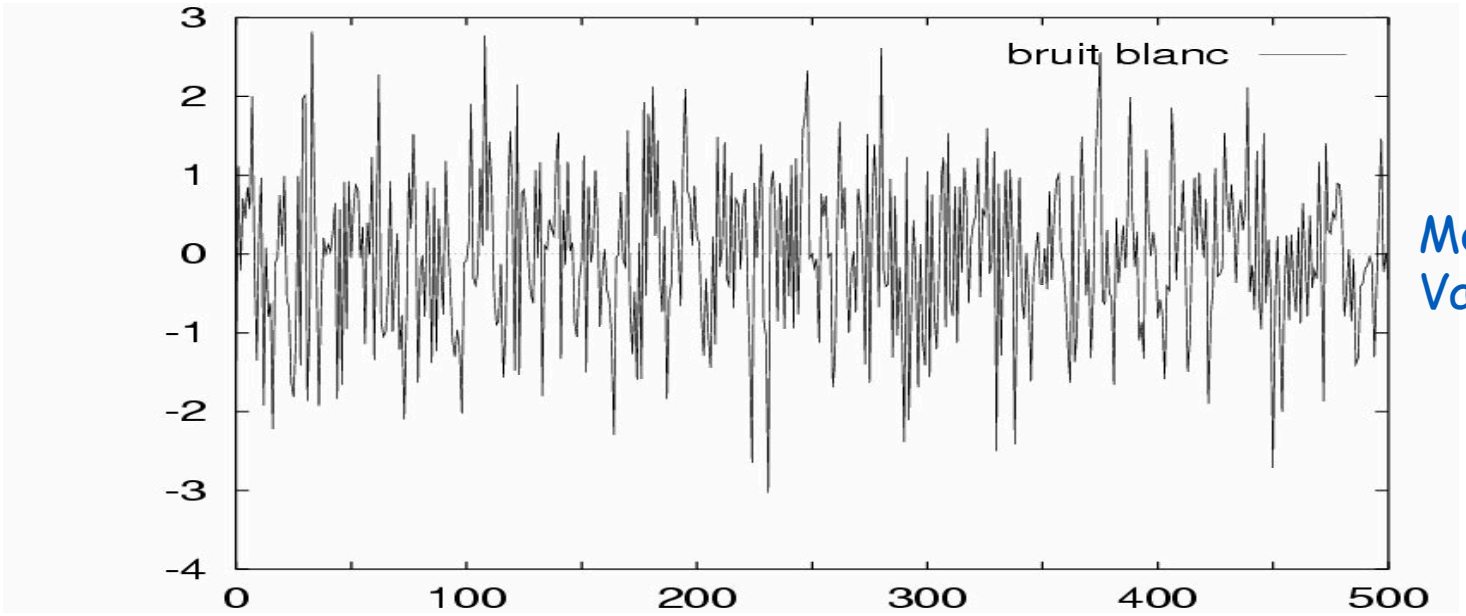
Modèles basiques de variogramme



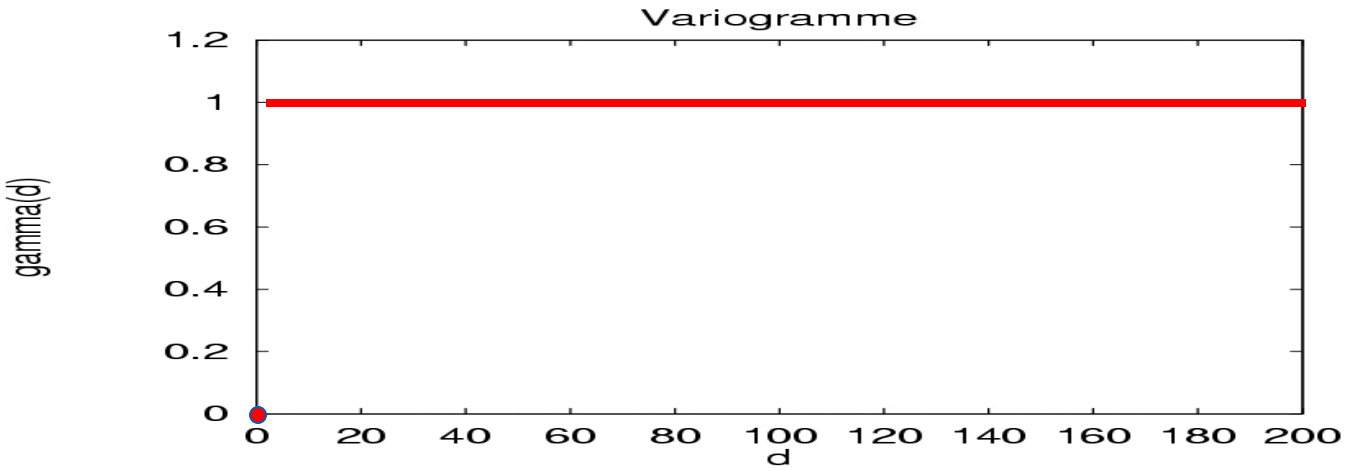
[Chilès]

Contraintes : fonction conditionnellement définie négative (pour garantir la positivité de la variance d'une combinaison linéaire de n variables aléatoires)

Exemple : profil purement aléatoire (bruit blanc)

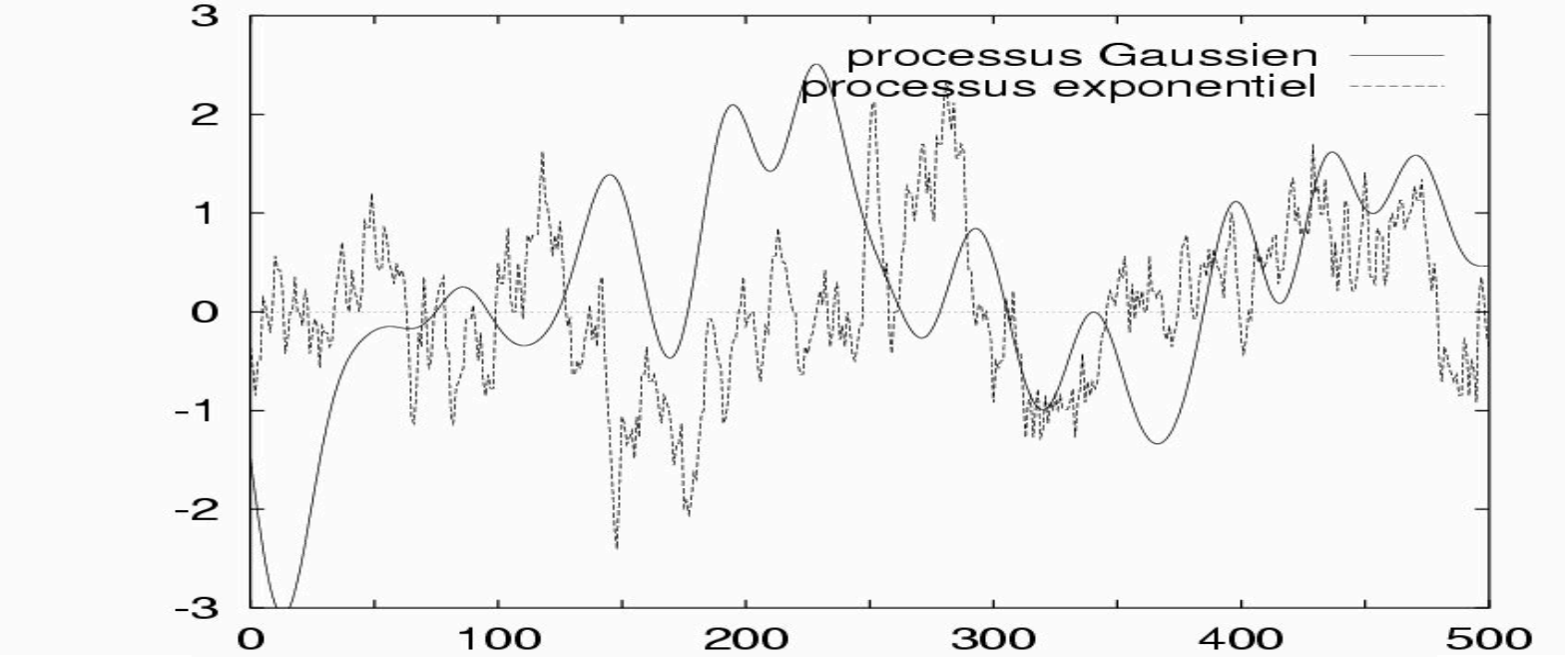


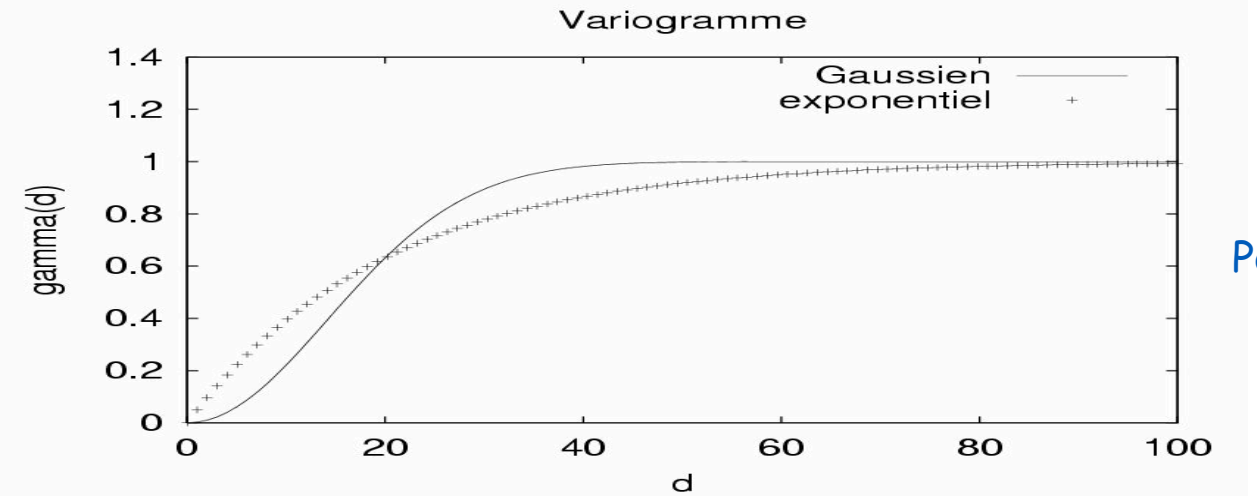
Moyenne = 0
Variance = 1



Remarque : la distribution des points du processus est gaussienne

Caractérisation de processus stochastiques

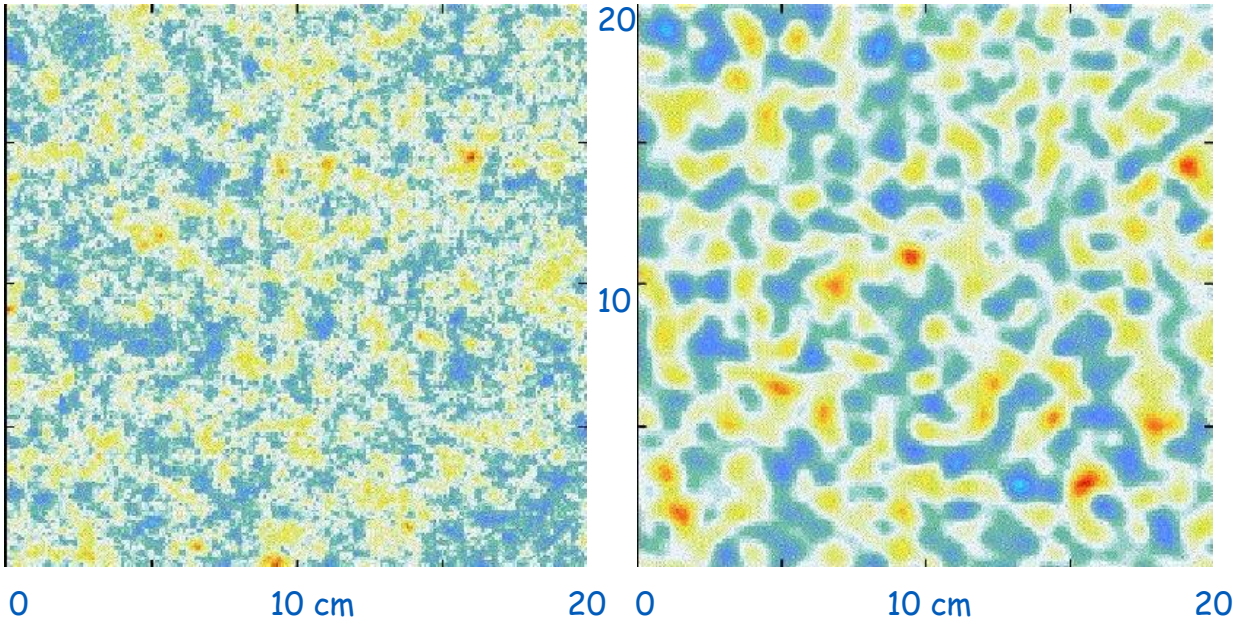




Portée = 20

Caractérisation de milieux aléatoires

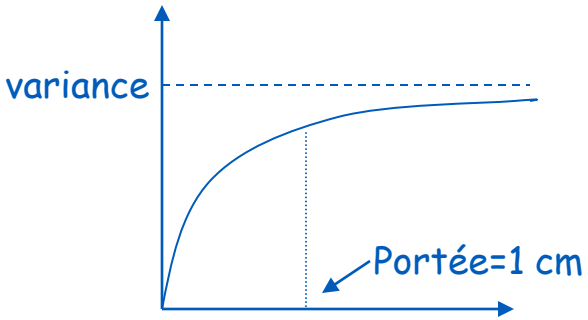
Exemples : turbulences thermiques dans un fluide



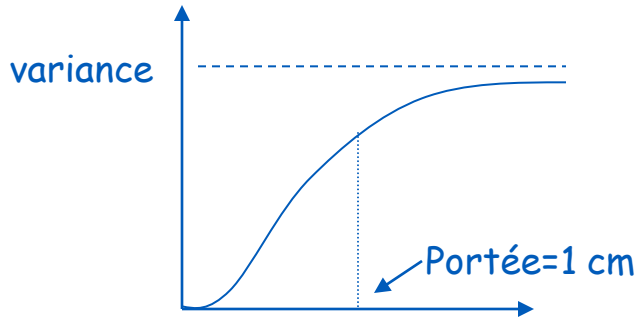
Niveaux de couleurs = températures

Variogrammes

Exponentiel : $\gamma(h) = \sigma^2 [\, 1 - \exp(-h) \,]$

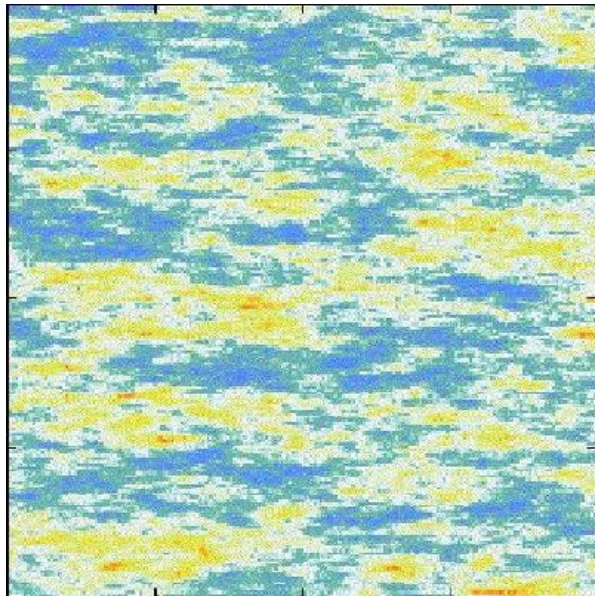


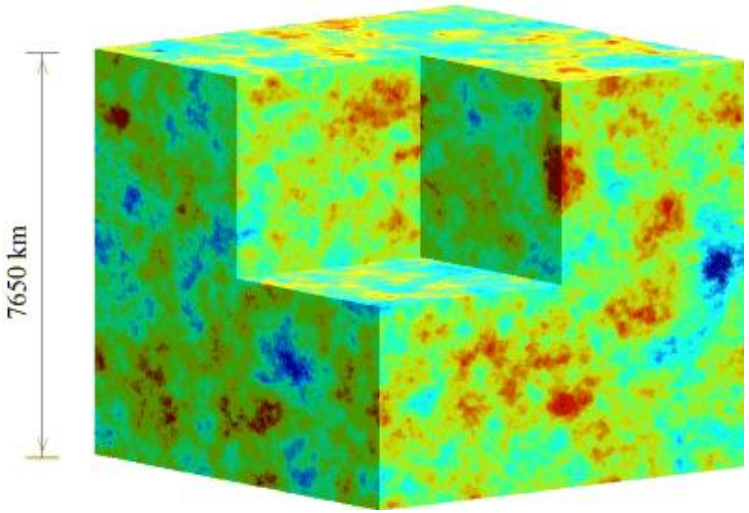
Gaussien : $\gamma(h) = \sigma^2 [\, 1 - \exp(-h^2) \,]$



Anisotropies (différences suivant les directions spatiales)

Exemples : milieux sédimentaires (champs de porosité)





[Baig et al., 03]

Variogrammes

exponentiel

$\gamma(h) = \sigma^2 \, [\, 1 - \exp(-h/a) \,]$

Portées horizontales : a_x=400 m

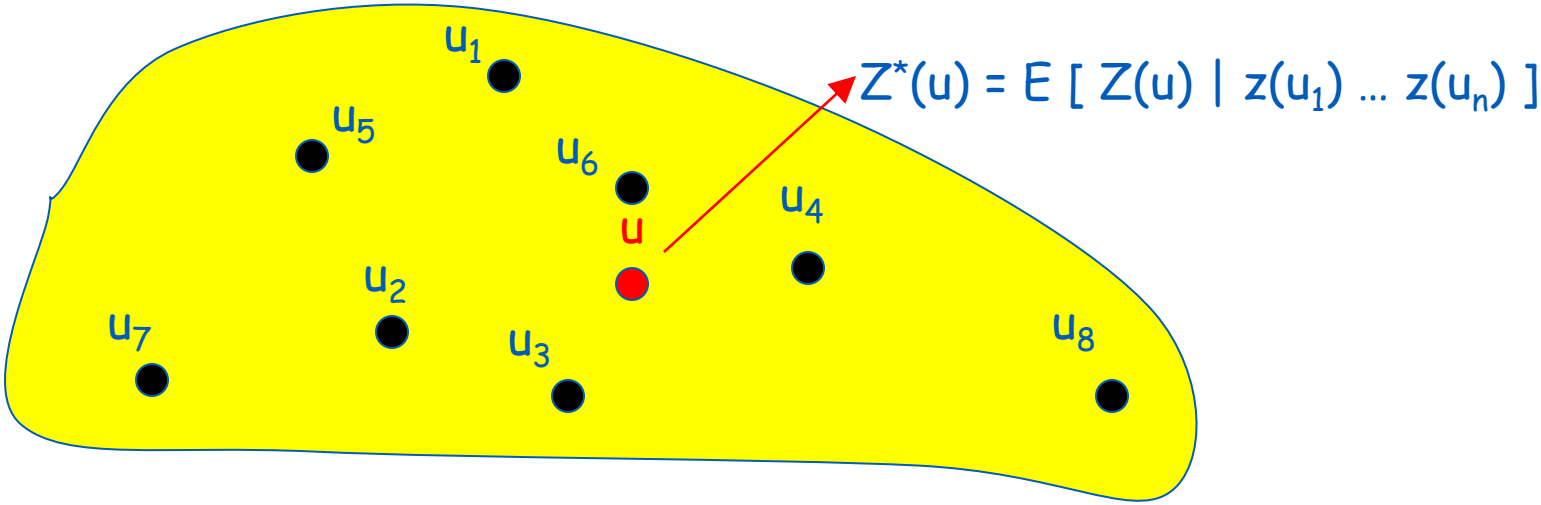
Portées verticales : a_y=100 m

Le krigeage

Combinaison linéaire des n données :

$$Z^*(u) = \sum_{i=1}^{n} \lambda_i Z(u_i)$$

Prise en compte de la configuration des données, de la distance entre données et cible, des corrélations spatiales, d'informations externes, …



Estimateur sans biais : $E[Z^*(u) - Z(u)] = 0$
la moyenne des erreurs est nulle

$Z^*(u)$ est optimal : $\mathrm{Var}[Z^*(u) - Z(u)]$ est minimale
la dispersion des erreurs est réduite

Le krigeage simple (moyenne connue)

$$Z^*(u) = \sum_{i=1}^{n} \lambda_i(u)\,[Z(u_i) - m] + m$$

(m constante et connue)

$\min_{\lambda_i} \{ E [Z^*(u) - Z(u)]^2 \}$

⇒ régression linéaire multiple (à résidus corrélés) par moindres carrés

Les poids de krigeage $\lambda_i(u)$ associés à $Z(u_i)$ sont déterminés par :

$$\begin{cases} \sum_{j=1}^{n} \lambda_j(u)\, C(u_i - u_j) = C(u_i - u) & \forall i = 1 \ldots n \\ \sum_{j=1}^{n} \lambda_j(u) = 1 \end{cases}$$

où C(h) est la covariance de Z(u)

Système de n équations linéaires à n inconnues qui possède une unique solution (si la matrice est non singulière)

Variance d'estimation

Intérêt du krigeage : calcul de l'erreur d'estimation

Variance du krigeage :

$$\sigma_K^2(u) = C(0) - \sum_{i=1}^{n} \lambda_i(u)\, C(u_i - u)$$

où C(0) est la variance de Z et C(h) sa covariance

- On peut donc visualiser les régions où l'estimation est imprécise et où il conviendrait de placer des nouveaux points de mesure

- La variance du krigeage ne dépend pas de la valeur des données mais uniquement de la covariance et du schéma d'échantillonnage

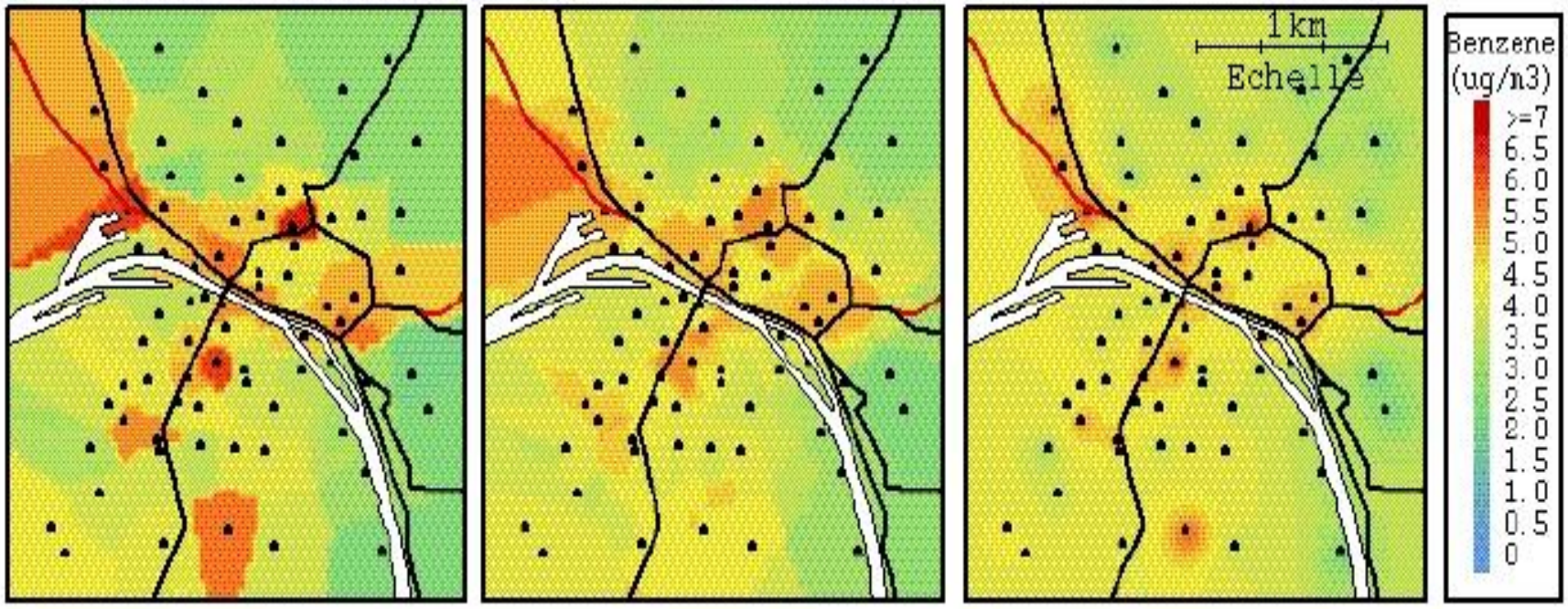
- Avant d'effectuer les mesures, on peut donc étudier la qualité des différents schémas d'échantillonnage possibles

Exemple : cartographie de polluants atmosphériques

73 données de concentration en benzène sur l'agglomération de Rouen

Méthodes classiques : moyenne pondérée des valeurs aux stations voisines

L'estimation dépend de la localisation des stations / nœud où on estime



(a) Méthode polygonale (b) Moyenne mobile (3 pts) (c) Inverse des (distances)**2

[Bobbia et al., 2000]

Ajustement du variogramme

Krigeage

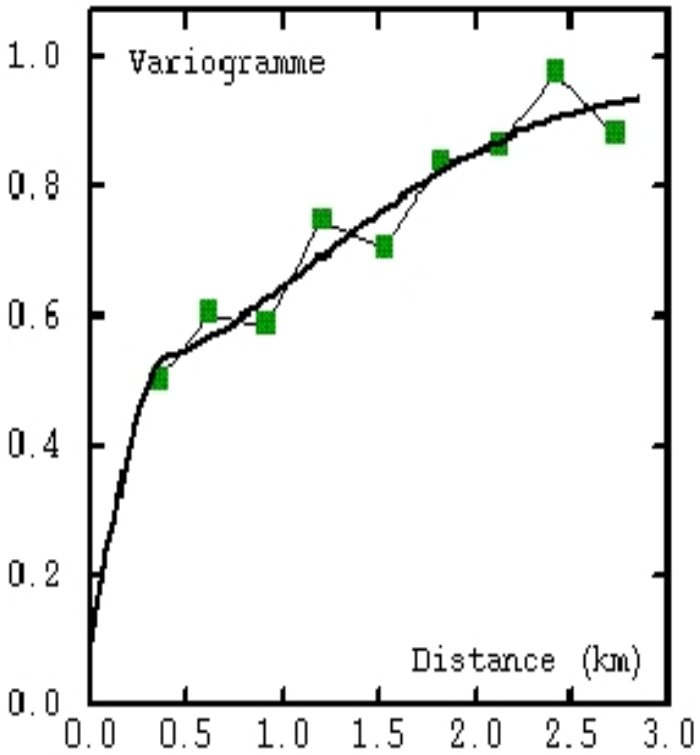
Prise en compte des caractéristiques spatiales du polluant (par l'intermédiaire du variogramme ou de la covariance)

Influence du modèle de variogramme

- Portée → voisinage du krigeage
- **Krigeage = interpolateur exact.**
  Présence d'un effet de pépite
  (bruit, microstructure < taille support,
  microstructure < distance minimale)

⇒ carte plus lisse

Introduction de dérive linéaire à l'aide de variables externes



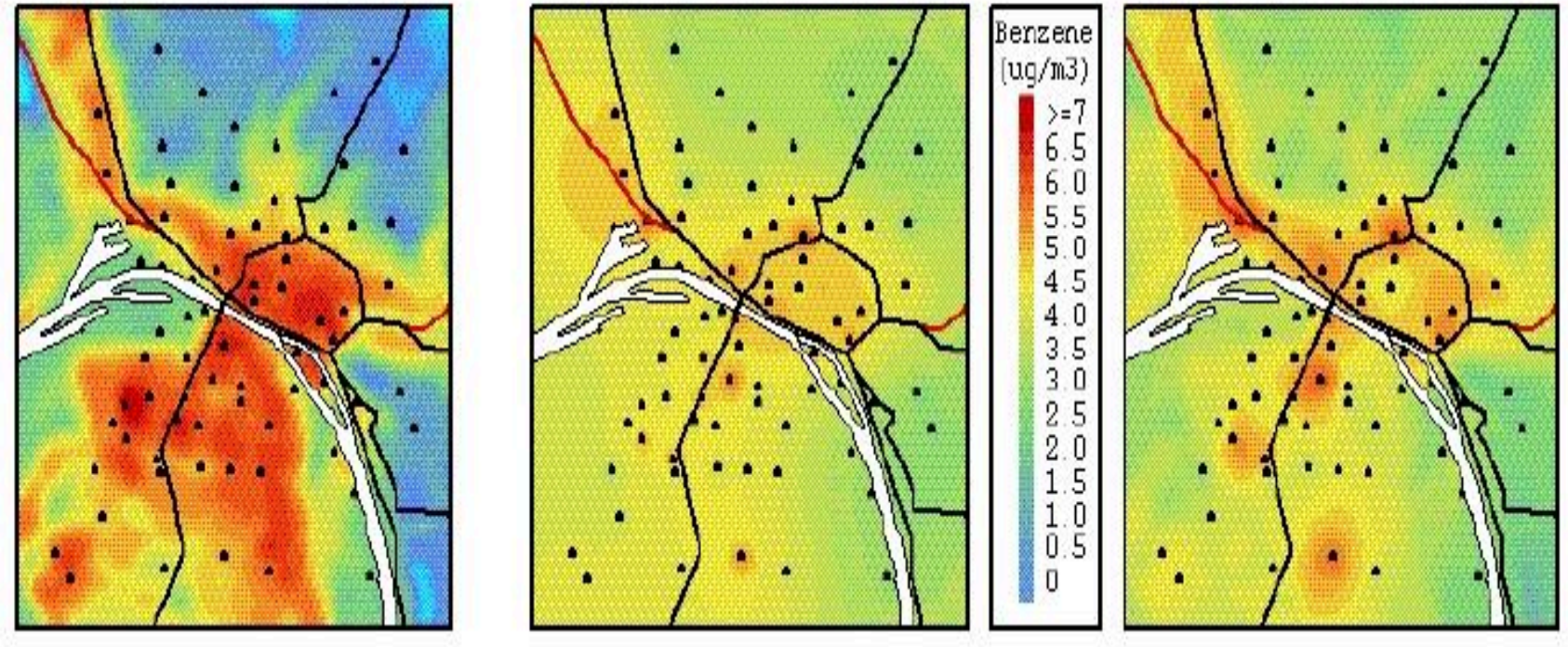
[Bobbia et al., 2000]

Exemples de krigeage

Le krigeage ordinaire se fait à moyenne constante mais inconnue

La dérive externe est une fonction de la densité de population et du relief (liés à la pollution automobile)

La pollution est supposée en dépendre linéairement



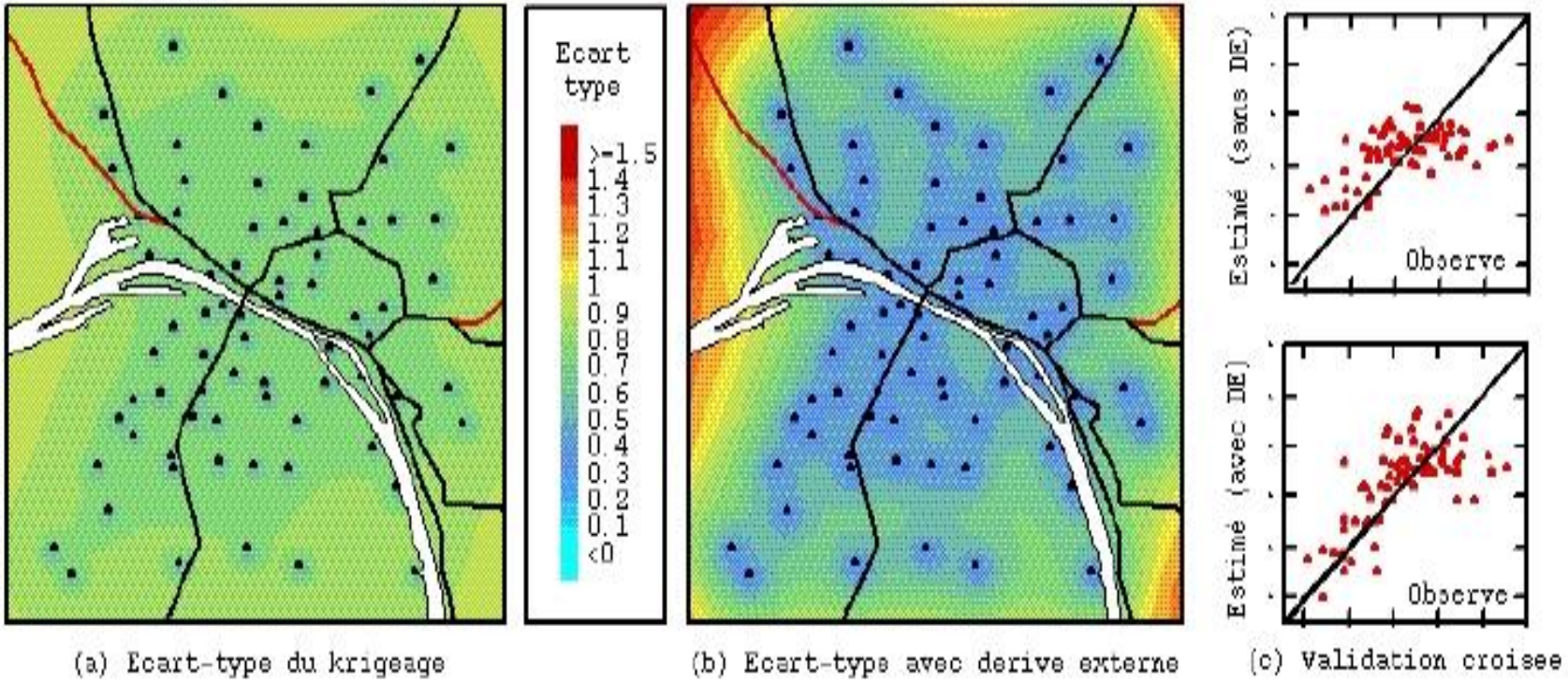
(a) Derive externe (b) Krigeage ordinaire (c) Krigeage avec dérive externe

[Bobbia et al., 2000]

Erreurs de krigeage

Deux méthodes quantitatives pour juger la qualité de l'estimation :

- Écart type du krigeage (→ fortes incertitudes en extrapolation)
- Validation croisée (→ la dérive externe améliore les résultats)



(a) Ecart-type du krigeage (b) Ecart-type avec derive externe (c) Validation croisee

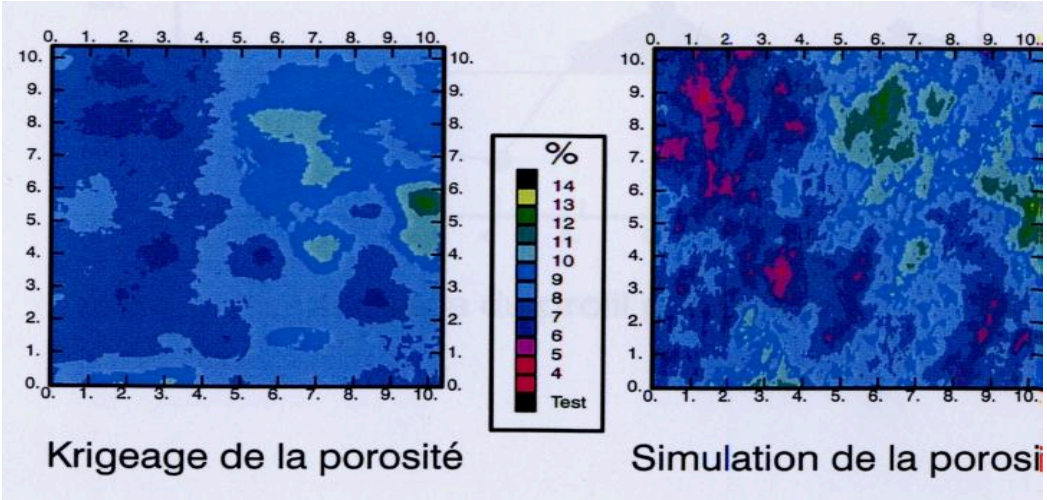
[Bobbia et al., 2000]

Simulations

- **Le krigeage cherche l'estimation optimale** (sans biais, variance d'erreur minimale) de la variable en tout point, à partir de données expérimentales

- **Une simulation représente une réalisation possible** du phénomène réel

  Elle cherche à reproduire sa **variabilité** (distribution, variogramme), tout en respectant les données expérimentales **(simulation conditionnelle)**



But principal de la simulation : quantifier les incertitudes par Monte Carlo
(N simulations équiprobables, calculs des moyenne, variance, distribution, …)

Il existe de nombreuses méthodes de simulations de champs aléatoires (décomposition LU, méthode spectrale, Karhunen-Loève, etc.)

Simulations conditionnelles

Krigeage obs.

+

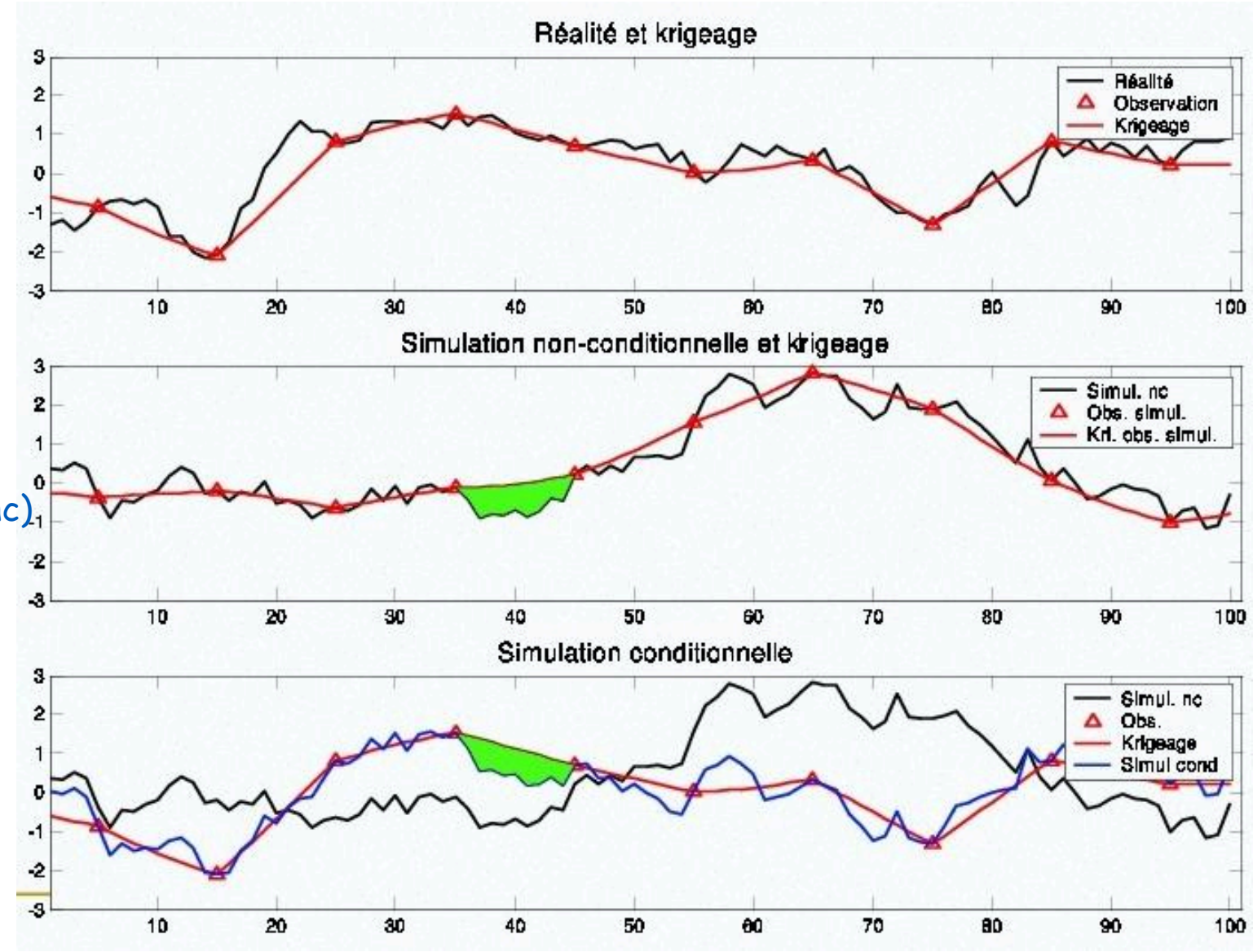
Simulation nc

-

Krigeage(simul. nc)

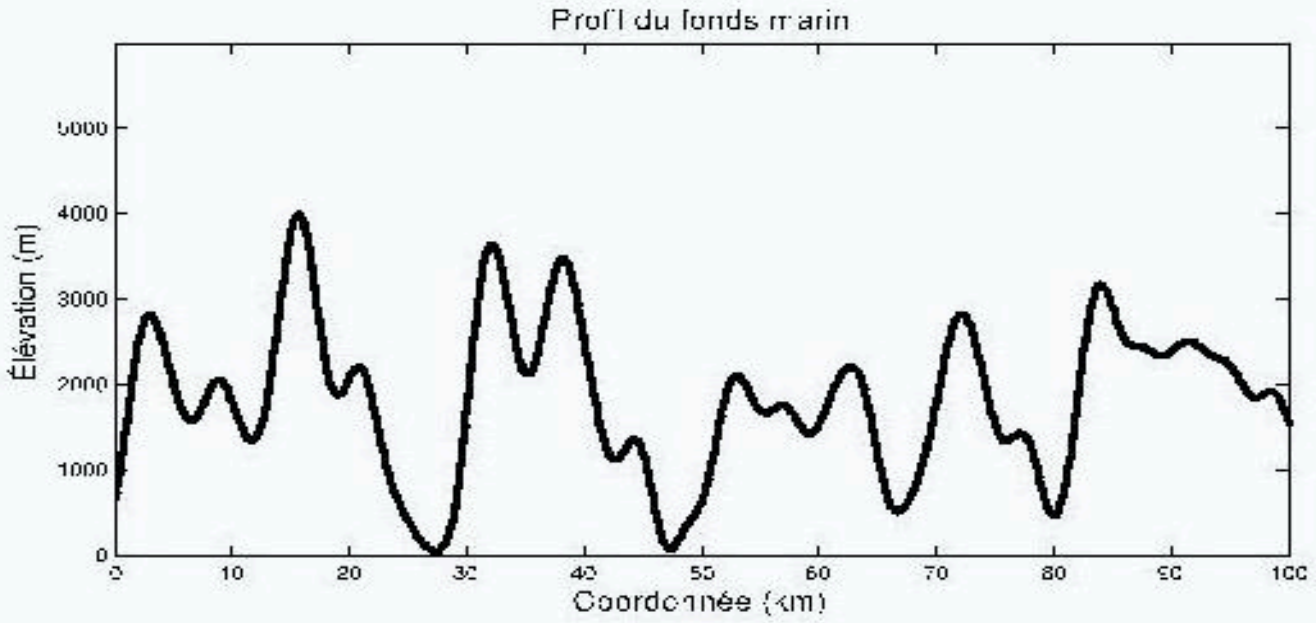
=

Simulation conditionnelle



[Marcotte, Cours EPM]

Exemple de simulations : profil du fond marin (1/6)

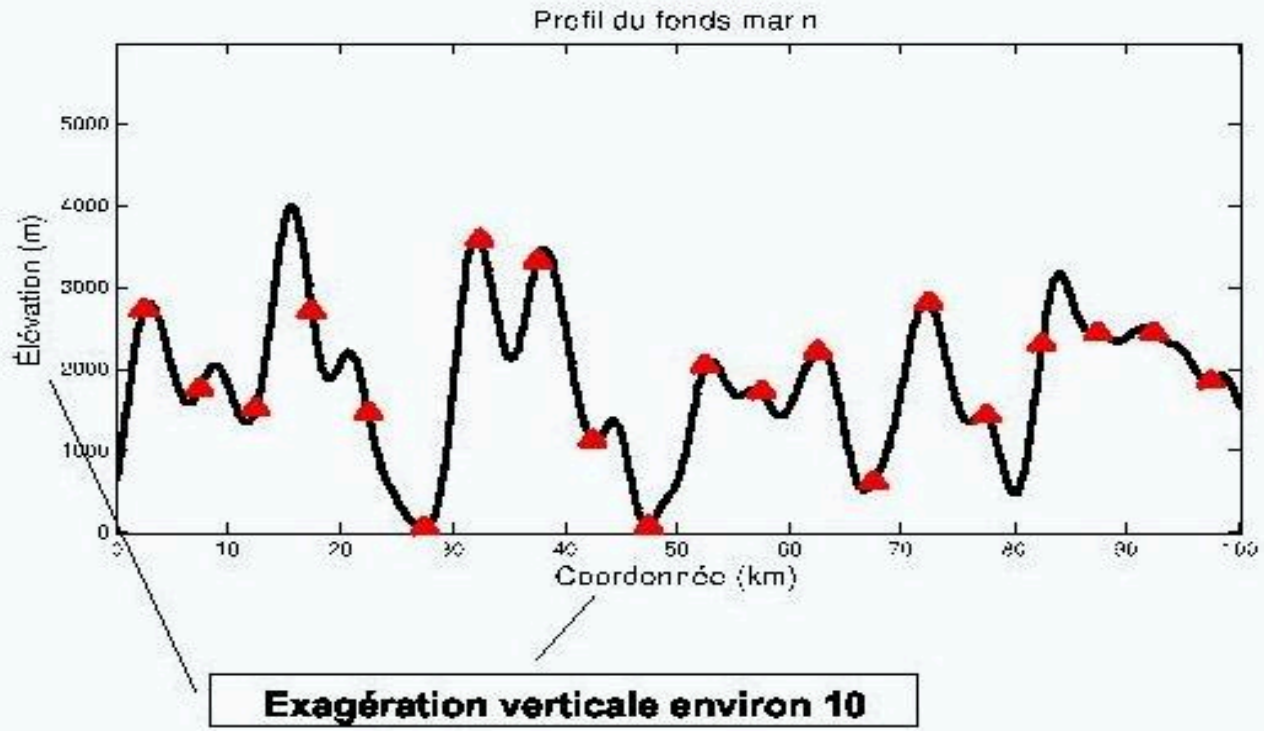


Vous déposez un câble de fibre optique sur le fonds marin.
Quelle longueur devrait avoir ce câble ?

[Marcotte, Cours EPM]

Exemple de simulations : profil du fond marin (2/6)

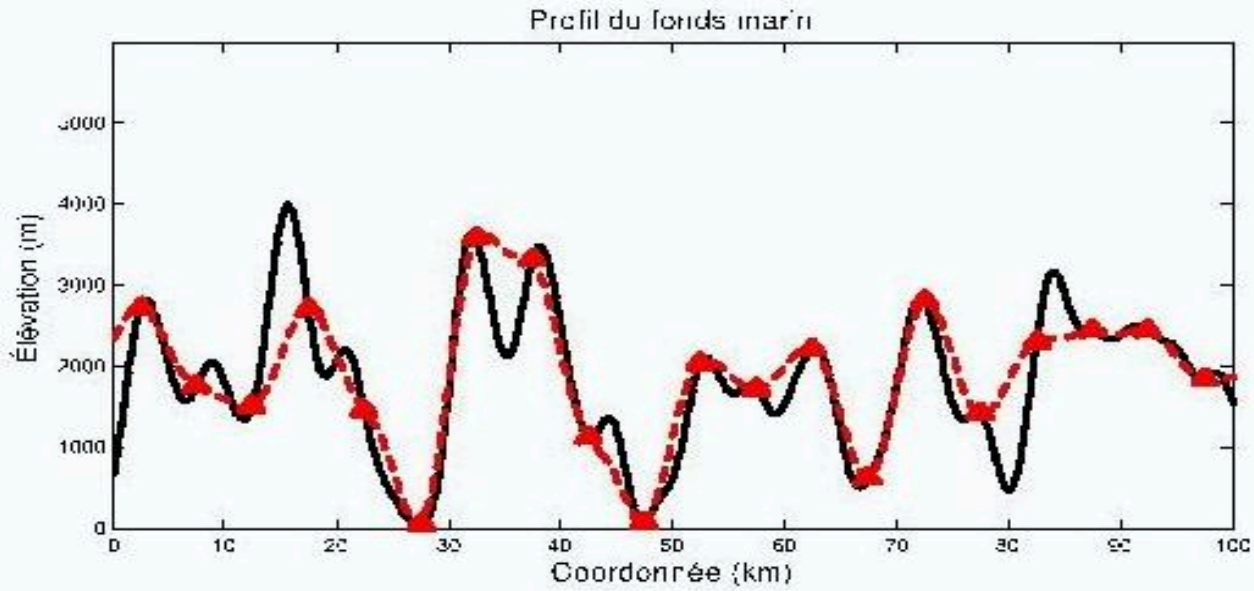
La profondeur exacte est connue uniquement aux points observations



[Marcotte, Cours EPM]

Exemple de simulations : profil du fond marin (3/6)

On effectue un krigeage

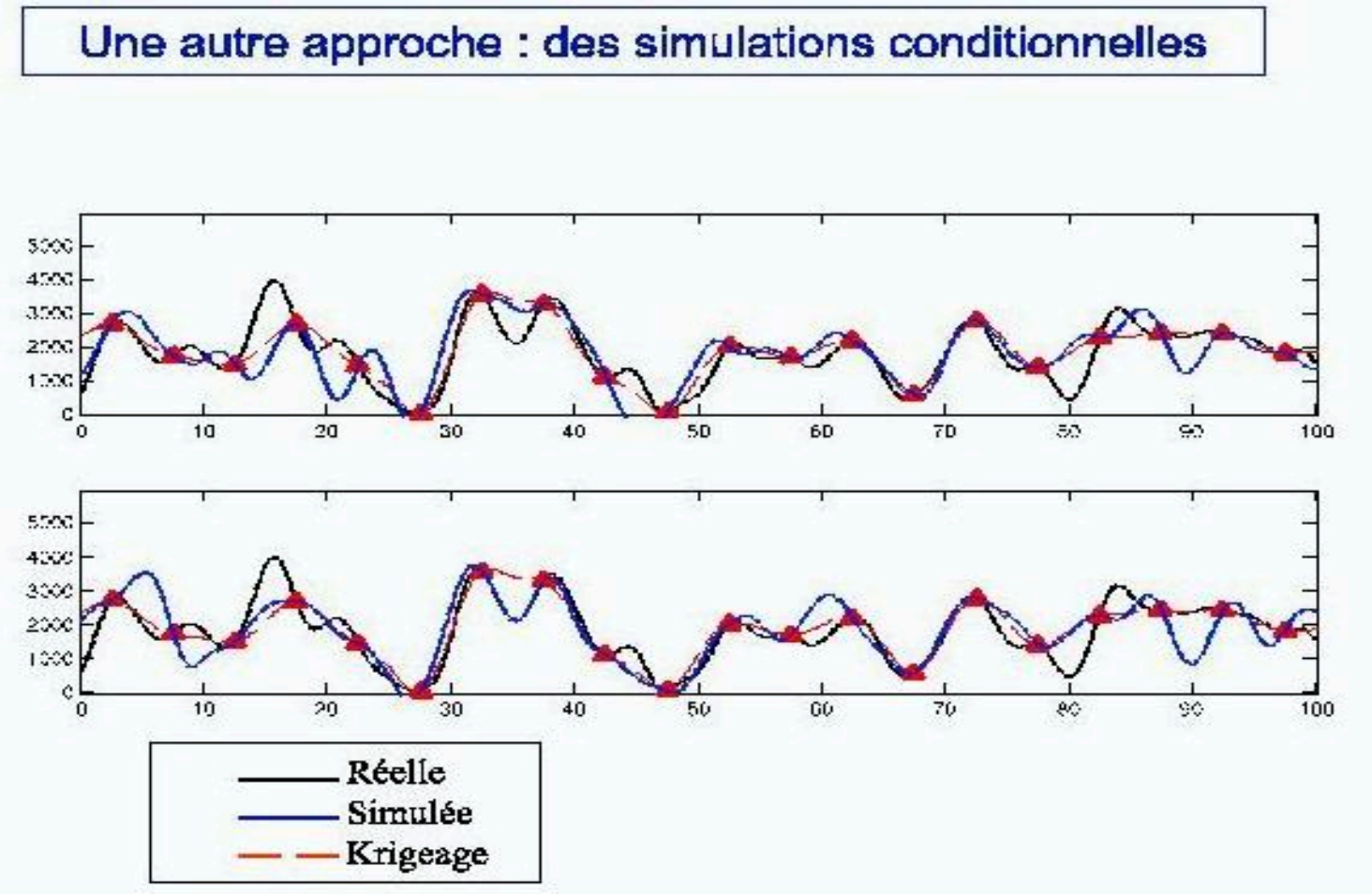


La longueur réelle est 110 km, le krigeage donne 104.6 km

On va manquer de câble !

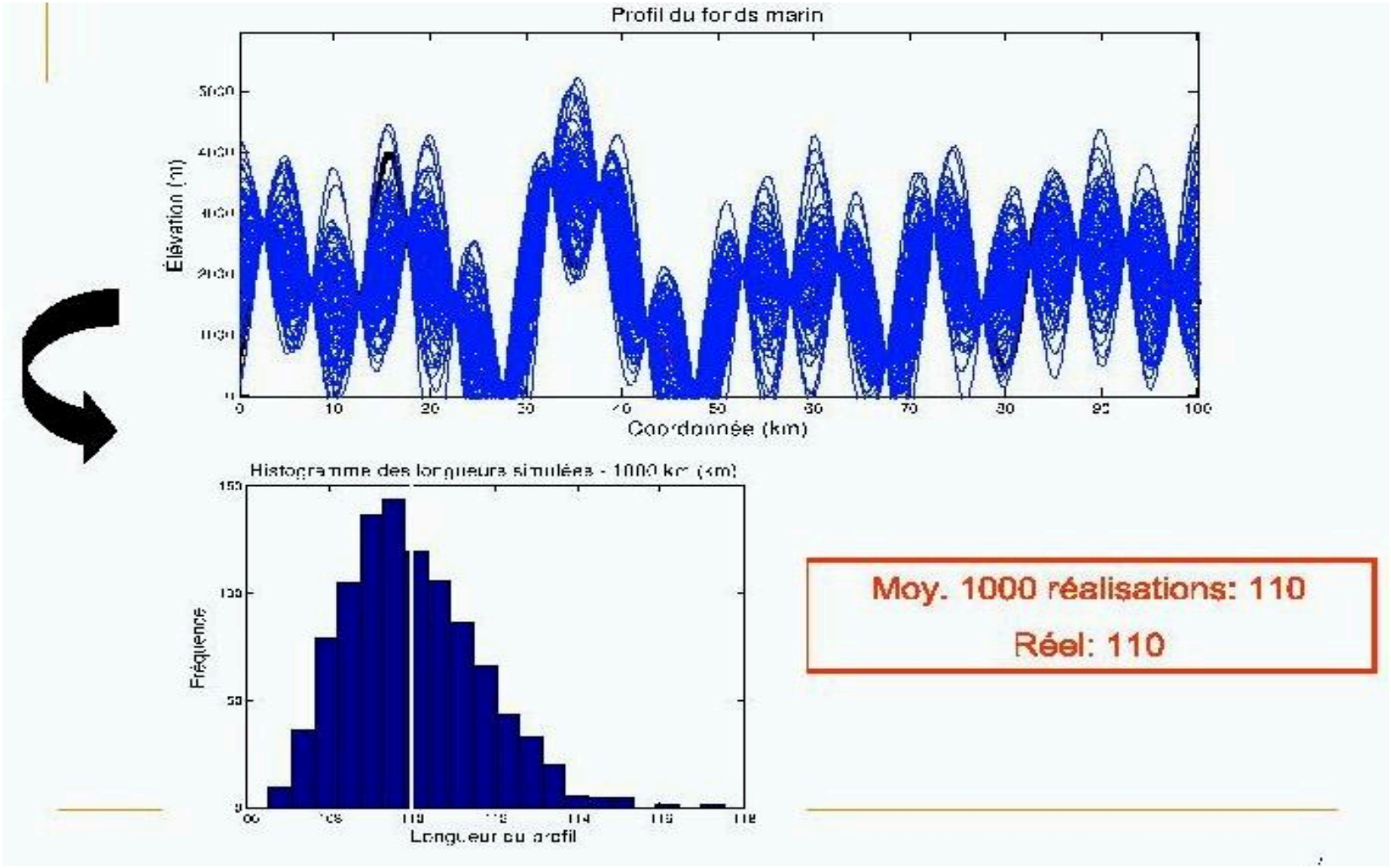
[Marcotte, Cours EPM]

Exemple de simulations : profil du fond marin (4/6)



[Marcotte, Cours EPM]

Exemple de simulations : profil du fond marin (5/6)



[Marcotte, Cours EPM]

Exemple de simulations : profil du fond marin (6/6)

L'intervalle de confiance à 95% obtenu par les simulations est:

[108.8, 113.5] =>

la valeur par krigeage 104.2 n'est même pas dans l'intervalle !

Longueur de câble = f (prof. du fonds marin) avec f la fonction de transfert
Ici f est non linéaire

Or le krigeage est un estimateur linéaire (qui gomme les variabilités locales)

Pour l'estimation d'une moyenne, on peut utiliser le krigeage car la fonction de transfert est linéaire

Comment calculer la probabilité de dépasser un seuil ou un quantile ?

- **à partir des simulations** (hypothèse gaussienne et calculs lourds),
- krigeage d'indicatrices : estimation d'indicatrices $1_{Z(x) \geq s}$

Plan du cours 3

1. Introduction
2. Méthode d'interpolation spatiale par krigeage
3. **Le métamodèle « processus gaussien »**
4. Un exemple d'application en hydrogéologie

Métamodèle du krigeage

L'idée du krigeage pour les codes est d'interpoler les réponses du code en dimension *p* comme pour une cartographie spatiale

Métamodèle intéressant car :

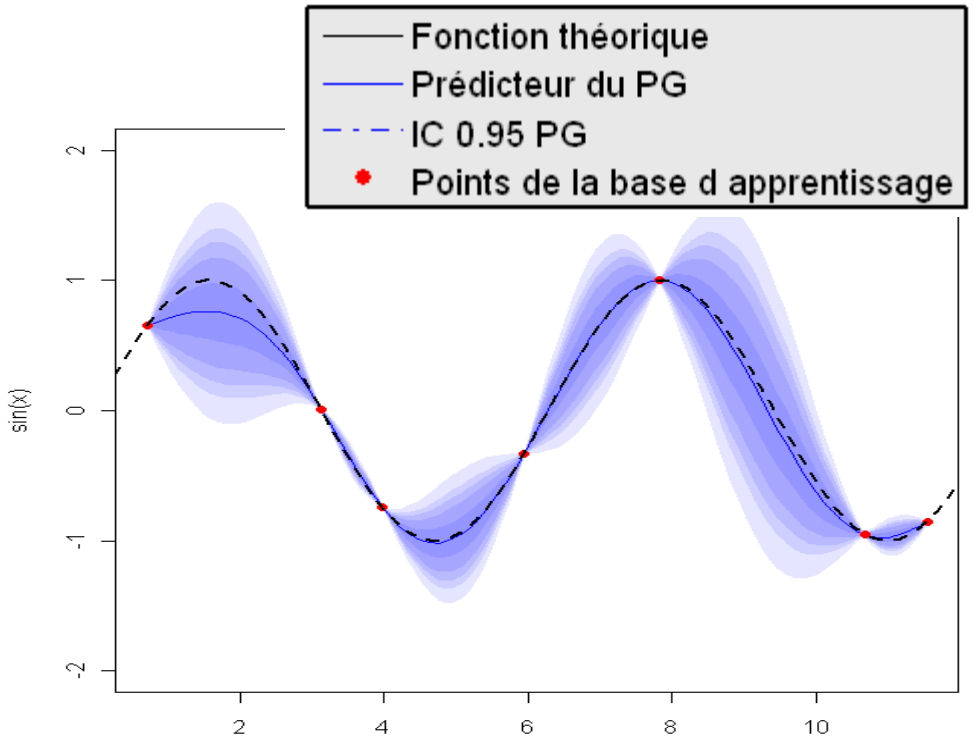
- interpole les réponses (pas d'hypothèse sur le bruit de mesure),
- évalue une nouvelle réponse très rapidement,
- fournit en plus d'une prédiction, une estimation de son erreur

Exemple en 1D :

Fonction théorique (p =1) :

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

Simulation de N = 7 calculs



[Chevalier, 2011]

Le Processus Gaussien (PG) et krigeage

Un Processus Gaussien est un processus aléatoire réel $\{Y(x)\}_{\in D \subset R^d}$
dont toutes ses lois finies-dimensionnelles $(Y(x_1),\ldots,Y(x_n))$ *sont gaussiennes*

$Y(x) \sim GP(\, m(x), C(x,x')\,)$ où $m(x) = \mathrm{E}(Y(x))$ et $C(x,x') = \mathrm{E}[(\, Y(x)\text{-}m(x)\,)\,(\, Y(x)\text{-}m(x'))]$

Différentes hypothèses de modélisation :

Les sorties correspondent à des observations de la trajectoire d'un Processus Gaussien, dont la fonction de covariance vérifie : $C(x,x') = C(x - x')$ *et*

$m(x) = m$ avec m connue pour le *Krigeage Simple, ou*

$m(x) = m$ avec m inconnue pour le *Krigeage Oridinaire, ou*

$m(x) = \beta\, F(x)$ avec β inconnue pour le *Krigeage Universel,*

Métamodèle Processus Gaussien (PG)

Idée : le code est la réalisation d'un champ aléatoire gaussien

Definition :

Processus Gaussien défini sur $R^p \times \Omega$

- $Y(x,\omega) = H(x) + Z(x,\omega)$

Regression **Partie stochastique**

Z processus stochastique avec :

$E_\Omega[Z(x)] = 0$

$Cov_\Omega(Z(x), Z(u)) = \sigma^2 R(x, u)$

où σ^2 est la variance

et R la fonction de corrélation

$Z \sim \mathcal{N}(0, \sigma^2 R)$

Exemples de choix paramétriques :

- **H** : polynôme de degré 1

$$H(\mathbf{x}) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

- **Z** : processus stationnaire avec covariance exponentielle généralisée

$$\mathrm{R}(\mathbf{x}, \mathbf{u}) = \mathrm{R}(\mathbf{x} - \mathbf{u}) = \exp\left(- \sum_{i=1}^{p} \theta_i |x_i - u_i|^{q_i} \right)$$

Loi jointe et loi conditionnelle

- **Loi jointe :**
  - Modèle PG : $Y(X) = \beta F(X) + Z(X)$
  - Base d'apprentissage (BA) $X_S = [x_1,...,x_N]$, Y_S , $F_S = F(X_S)$, $R_S = \left(R(x_i, x_k)\right)_{i,k}$
  - Loi jointe sur la BA : $Y_S \sim N(\beta F_S, \sigma^2 R_S)$
  - Loi conditionnelle pour un nouveau point x^* : $Y(x^*)_{|X_S, Y_S} \sim \mathcal{N}(\mu(x^*), \tilde{\sigma}^2(x^*))$

$$\mu(x^*) = \beta F(x^*) + r(x^*) R_S^{-1} \left[Y_S - \beta F_S\right]$$

$$\tilde{\sigma}^2(x^*) = \sigma^2\left(1 - {}^t r(x^*) R_S^{-1} r(x^*)\right) \quad \text{avec} \quad r(x^*) = \left[R(x_1, x^*),..., R(x_N, x^*)\right]$$

- **Prédicteur et erreur associée :**
  - Meilleur prédicteur linéaire sans biais (BLUP) : $\hat{Y}(x^*) = \mu(x^*)$
  - Interpolateur exact des points de la BA
  - Formulation de l'erreur : $MSE[\hat{Y}(x^*)] = \tilde{\sigma}^2(x^*)$

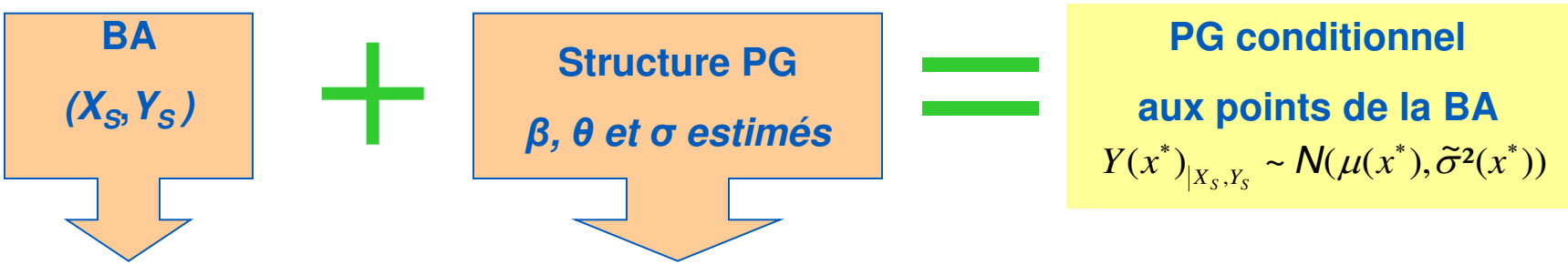
Construction du métamodèle PG

Estimation des paramètres de régression et de corrélation β, θ, σ

- Estimation sur la BA
- Maximum de vraisemblance + algorithme d'optimisation (par ex. stoch)
- Maximisation du R^2 sur une base de test pour θ

$$Q_2(Y,\hat{Y}) = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(\overline{Y} - Y_i\right)^2}$$

Formulation et compréhension du modèle :

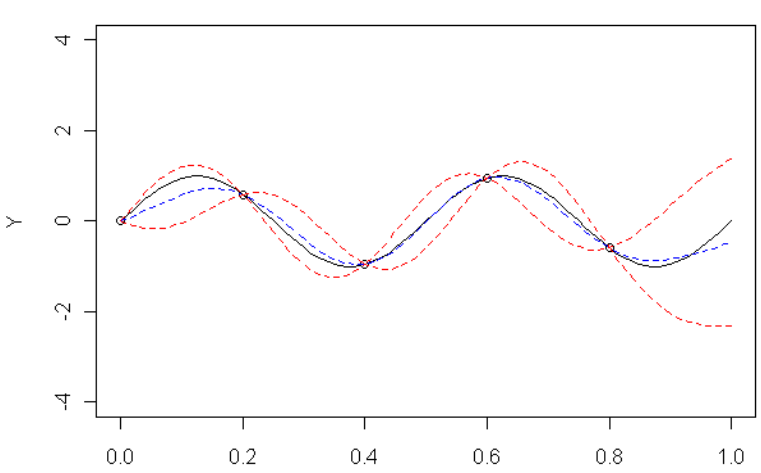


- Résultat des simulations du code
- Points « d'ancrage »
- Points Interpolés exactement
- Sélection optimale de ces points

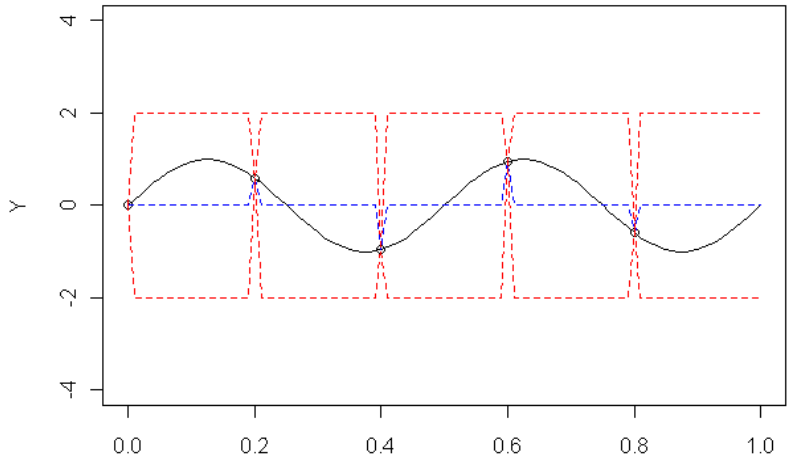
- Structure du réseau, de la « toile »
- Estimation des paramètres par maximum de vraisemblance sur les points de la BA
- Sélection des paramètres β et θ significatifs

Impact du choix des hyperparamètres θ et σ

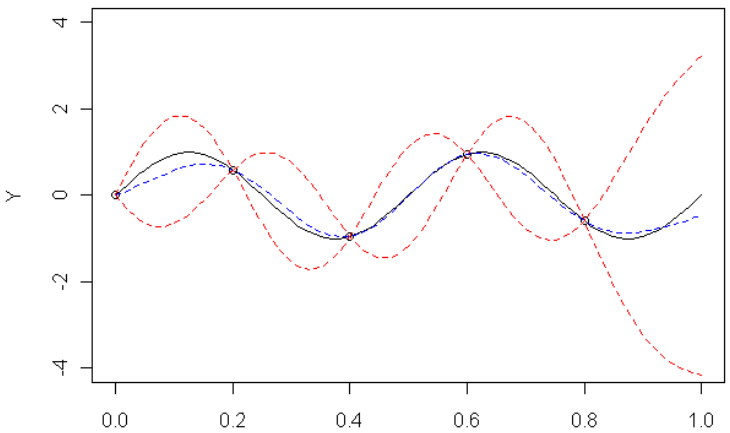
$$f(x) = \sin(4\pi x)$$



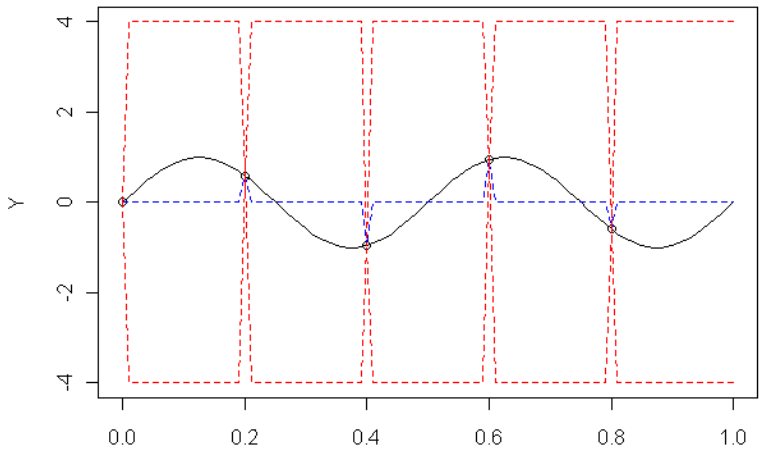
$$\sigma^2 = 1; \theta = 0.2$$



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



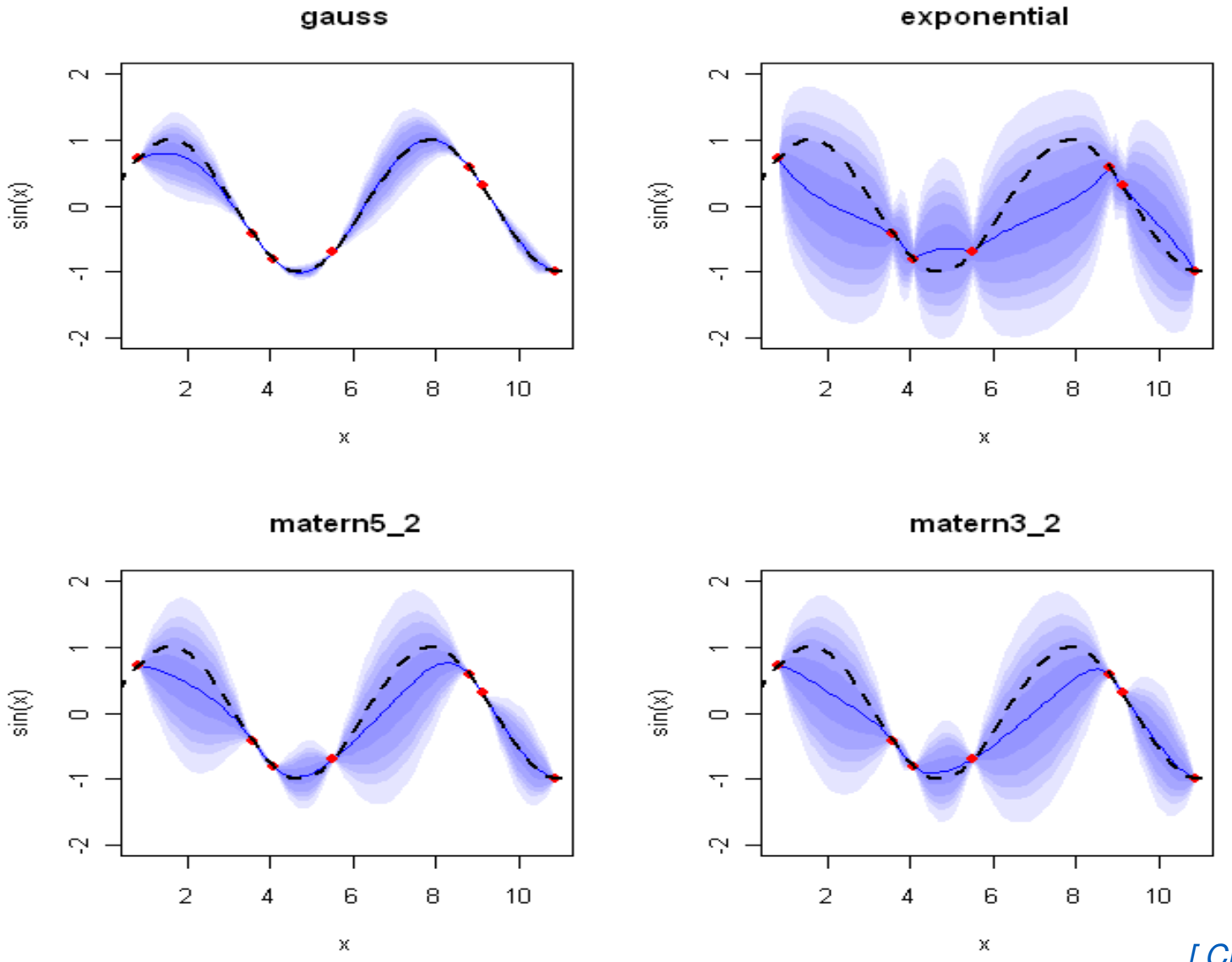
$$\sigma^2 = 4; \theta = 0.2$$



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

[Le Gratiet, 2011]

Impact du choix de la fonction de covariance



[Chevalier, 2011]

Estimation des paramètres par maximum de vraisemblance

- Maximisation de la vraisemblance sur la base d'apprentissage (X_s,Y_s):

$$\left(\beta^*,\theta^*,\sigma^*\right)=\underset{(\beta,\theta,\sigma)}{\text{Argmax}}\ \ln\left[L\left(Y_S,\beta,\theta,\sigma\right)\right]$$

avec

$$\ln L\left(Y_S,\beta,\theta,\sigma\right)=-\frac{N}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2}\ln\left(\det R_S\right)-\frac{1}{2}\sigma^2\ {}^t\left[Y_S-\beta F_S\right]R_S^{-1}\left[Y_S-\beta F_S\right]$$

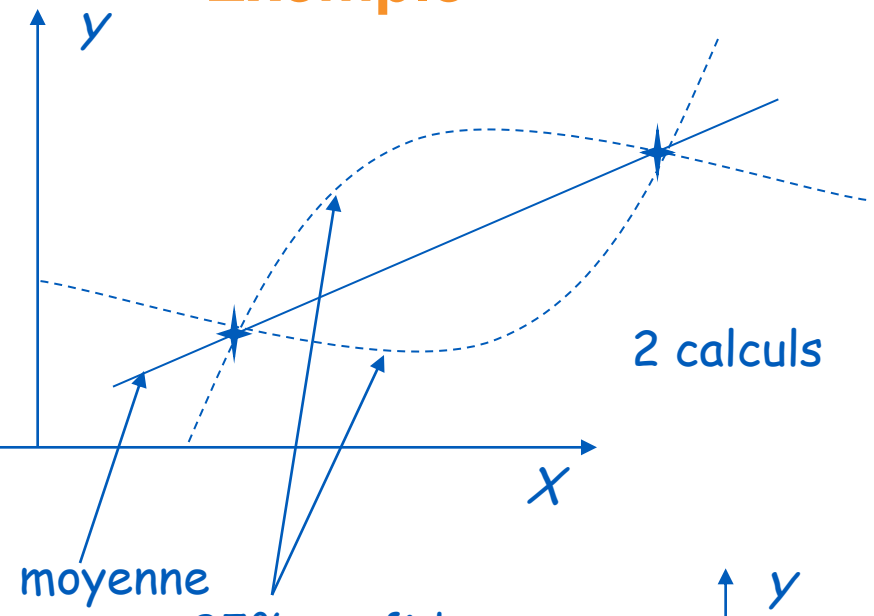
- Estimation conjointe de β et σ :

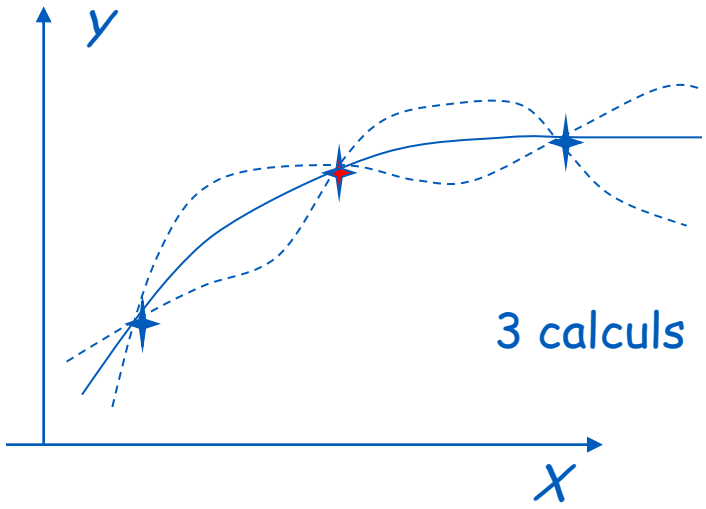
$$\begin{cases}\beta^*=\left[{}^tF_S R_S^{-1}F_S\right]^{-1}{}^tF_S R_S^{-1}Y_S\\ \sigma^{2*}=\dfrac{1}{N}{}^t\left[Y_S-\beta^*F_S\right]R_S^{-1}\left[Y_S-\beta^*F_S\right]\end{cases}$$

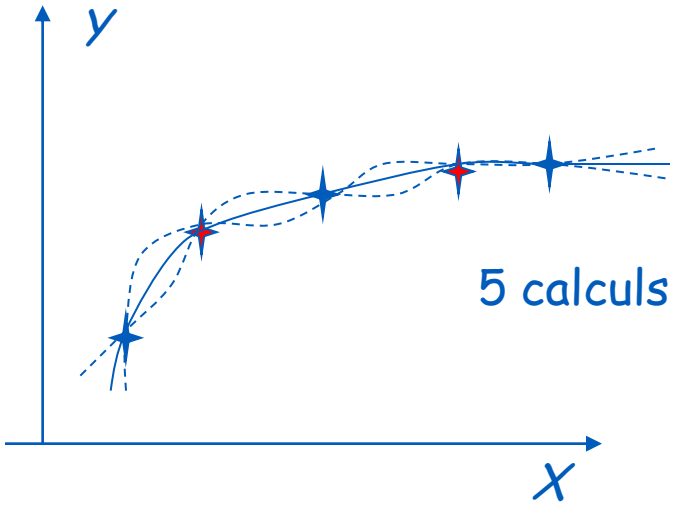
- Estimation des paramètres de corrélation θ :

$$\left(\theta^*\right)=\underset{\theta}{\text{Argmin}}\ \ \psi(\theta)\quad\text{avec}\quad\psi(\theta)=\left|R_S\right|^{-1/N}\sigma^{2*}$$

Exemple







Conclusion: à partir d'un certain nb de points, le métamodèle devient précis

Problème : fléau de la dimension *p* (dans l'optimisation des hyperparamètres)

Introduction à la planification adaptative

- Les suites de quasi-Monte Carlo et les plans LHS optimisés (par exemple maximin ou à discrépance faible) sont performants pour l'analyse de sensibilité et la construction de métamodèles

- La planification adaptative (séquentielle) est la **voie royale** pour réduire le nombre d'appels au code :

  1 on part d'un plan initial de quelques calculs

  2 on ajuste un premier métamodèle

  3 on ajoute de nouvelles simulations là où le plus d'information (vis-à-vis de l'objectif de l'étude) sera apportée

- En fonction de l'objectif de l'étude d'incertitudes, diverses stratégies de planification adaptative peuvent être proposées. La plupart requiert un **métamodèle**

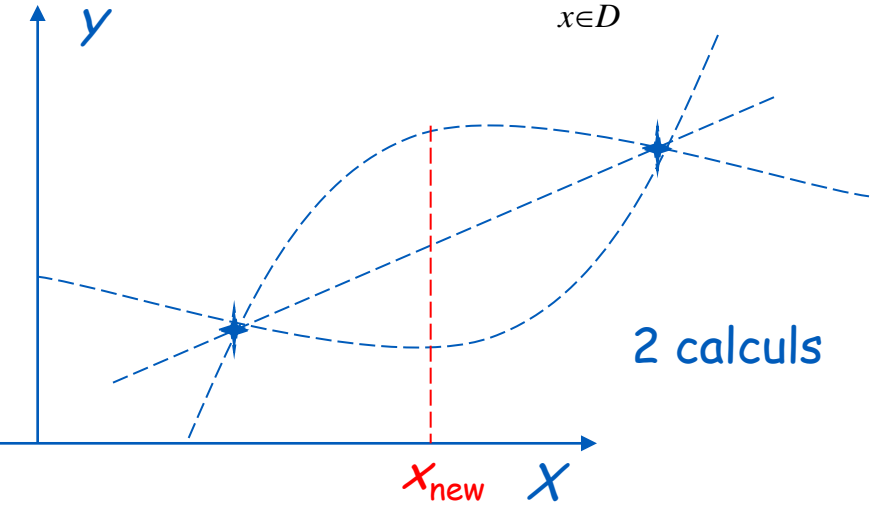
Plans adaptatifs

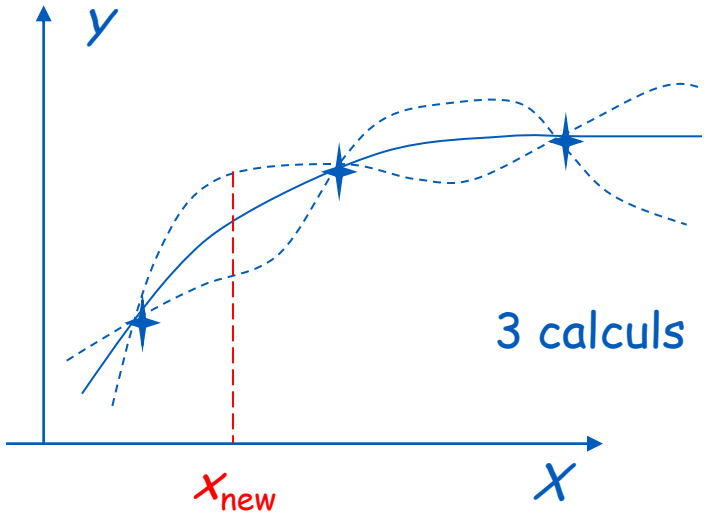
Example: utilisation du critère MSE

$$MSE(x) = \sigma^2 + {}^t r(x) R_{LS}^{-1} r(x) + u(x)({}^t(\beta F_{LS}) R_{LS}^{-1} \beta F_{LS})^t u(x)$$

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

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





Remarque : d'autres critères sont possibles

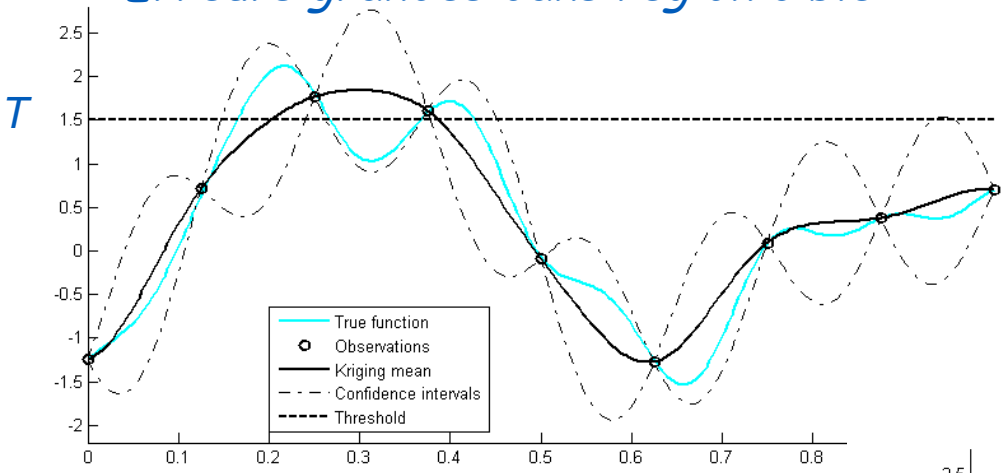
Conclusion: Les plans adaptatifs sont les plus efficace,
Mais en pratique, besoin d'un plan initial (space filling)

Plan adaptatif pour l'évaluation d'événements rares

Métamodèle de krigeage : predicteur + bornes de confiance

Problème : estimer P_f = Prob [$f(X) > T$] avec X = entrées aléatoires ; T = seuil

Variance raisonnable partout
Erreurs grandes dans région cible



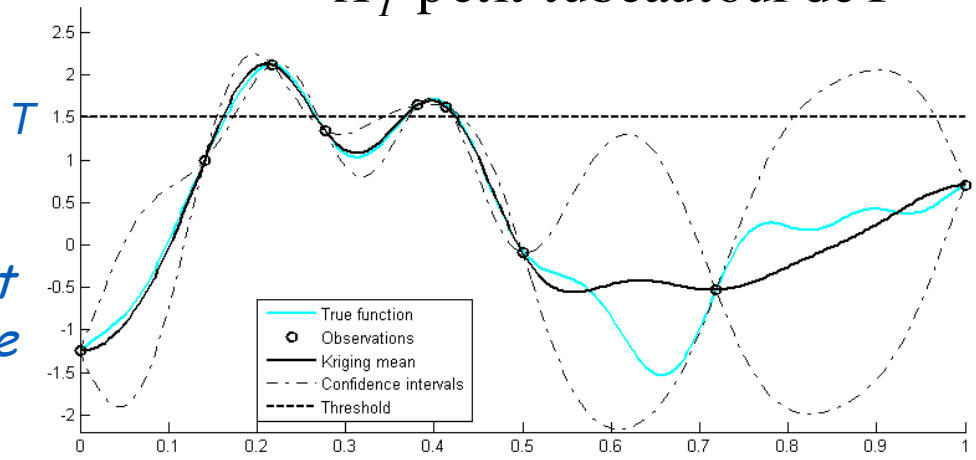
[Picheny et al., 2010]

☺ Plan adaptatif

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

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

X_T petit tubeautour de T



Variance élevée partout
sauf dans la région cible

Plan pour l'optimisation d'une sortie d'un modèle

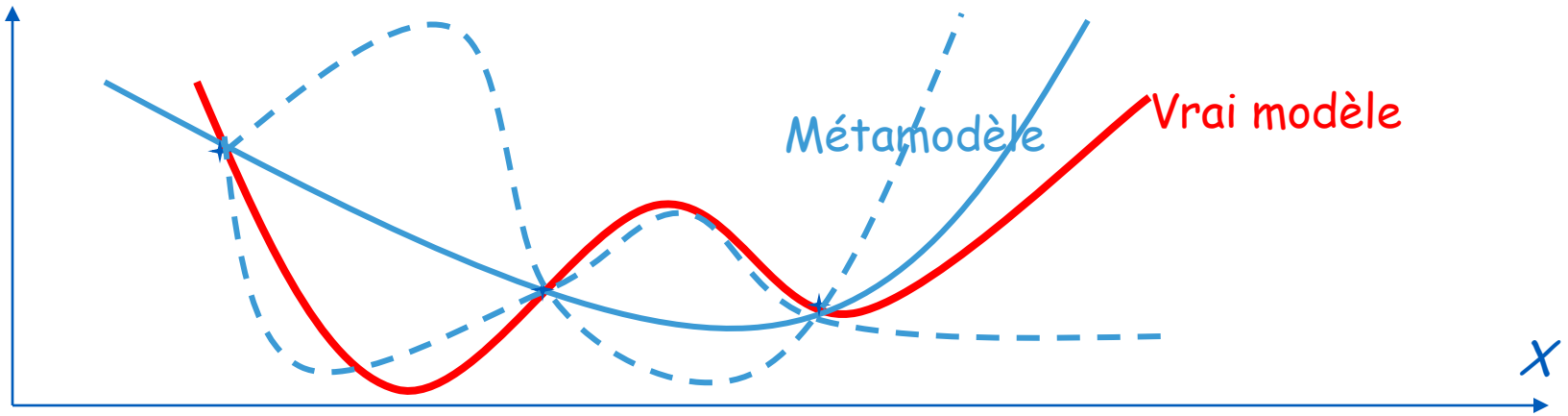
Problématique industrielle : conception à l'aide de modèles numériques lourds (automobile, nucléaire, aéronautique, …)

Le problème : trouver les paramètres X qui minimisent une sortie du modèle G

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

☠ **Si G est coûteux, une idée naturelle est d'optimiser sur un métamodèle** : dangereux car le métamodèle lisse le vrai modèle

☺ Le krigeage permet de prendre en compte l'erreur du métamodèle, et de définir l'amélioration espérée $EI(X)$ en chaque $X \in D$

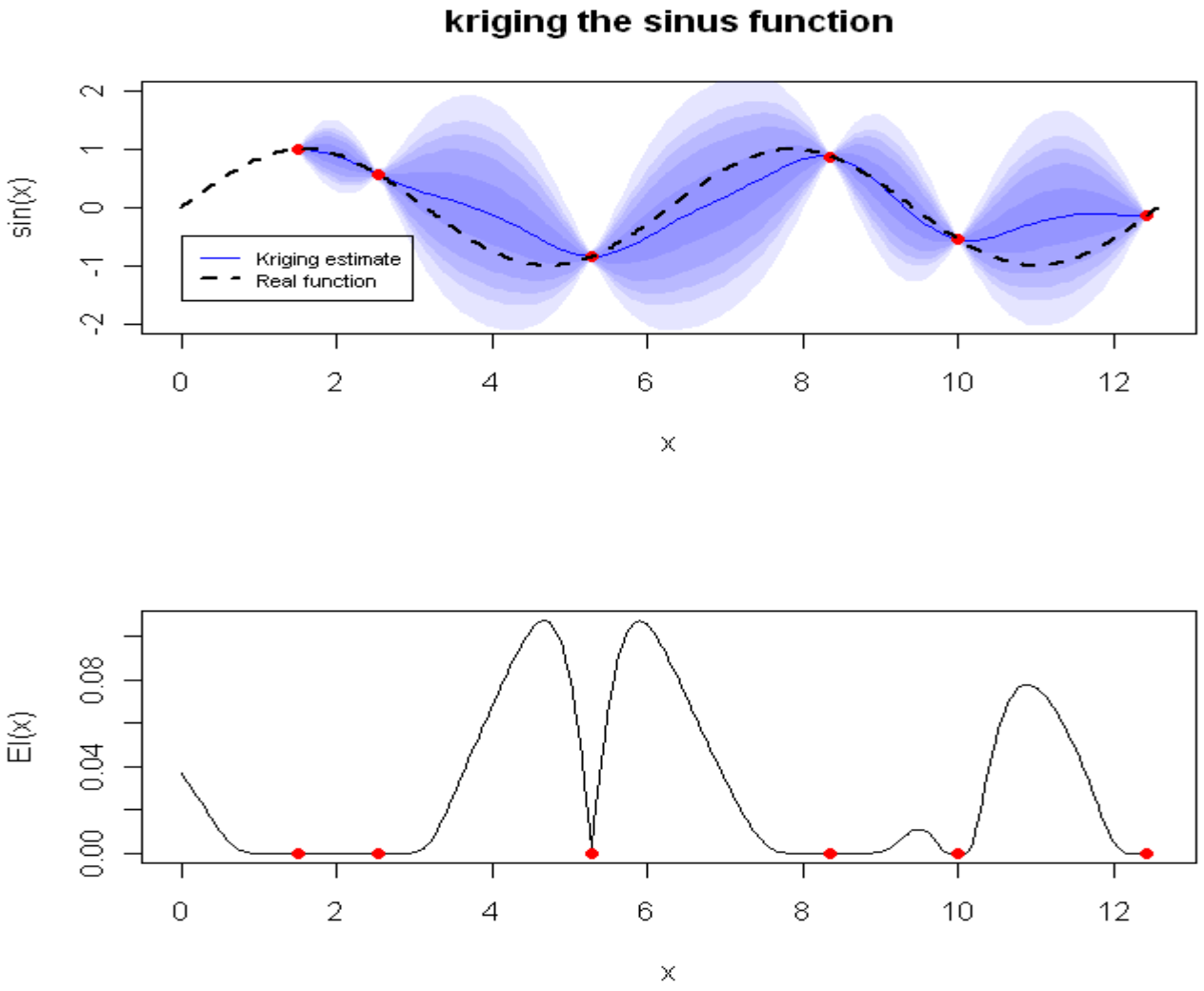


EI(x) = E[max(0 ,minimum observé jusqu'à présent – G(x))]

Algorithme EGO

[Chevalier, 2011]

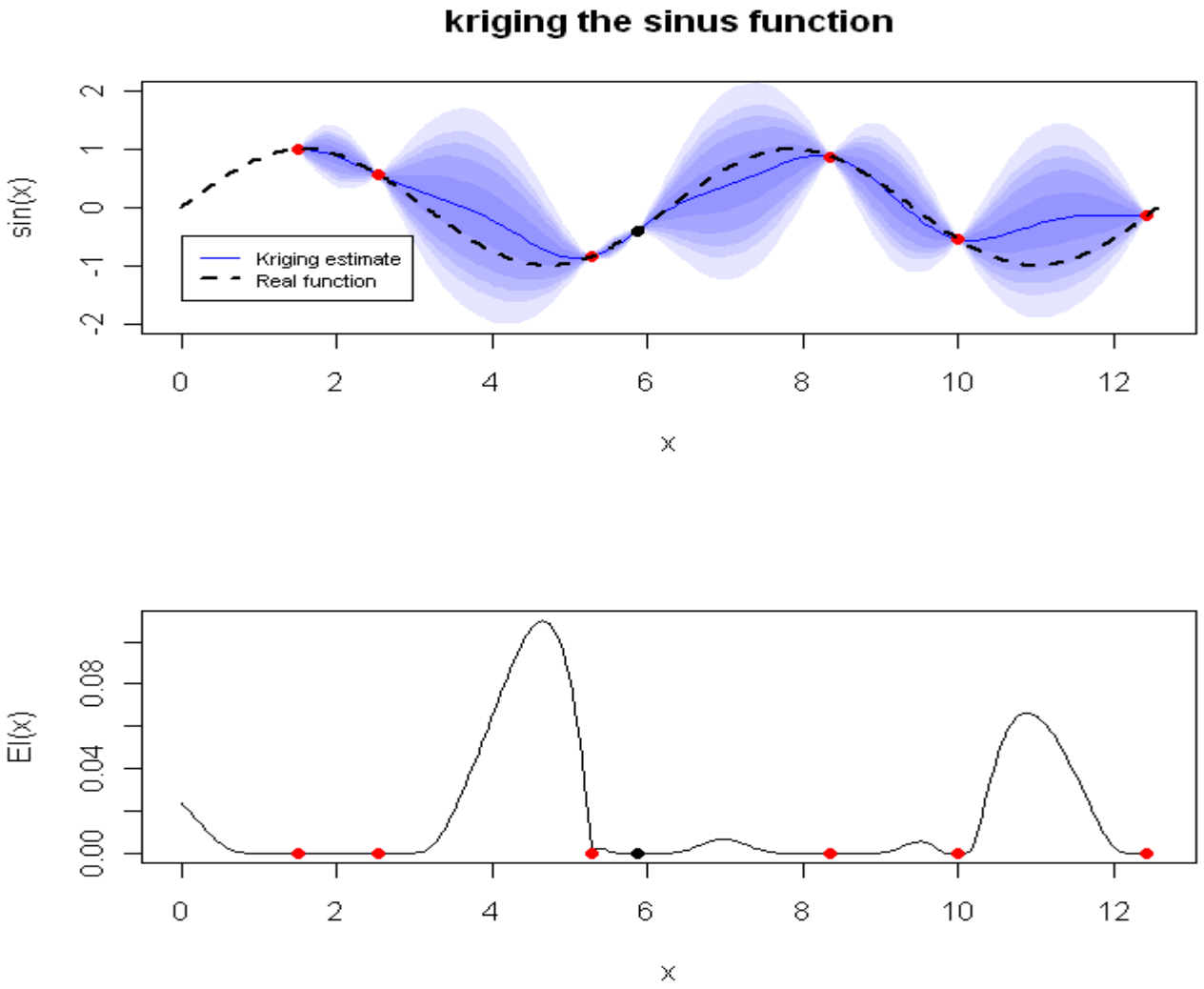
- EGO: step 0



Algorithme EGO

[Chevalier, 2011]

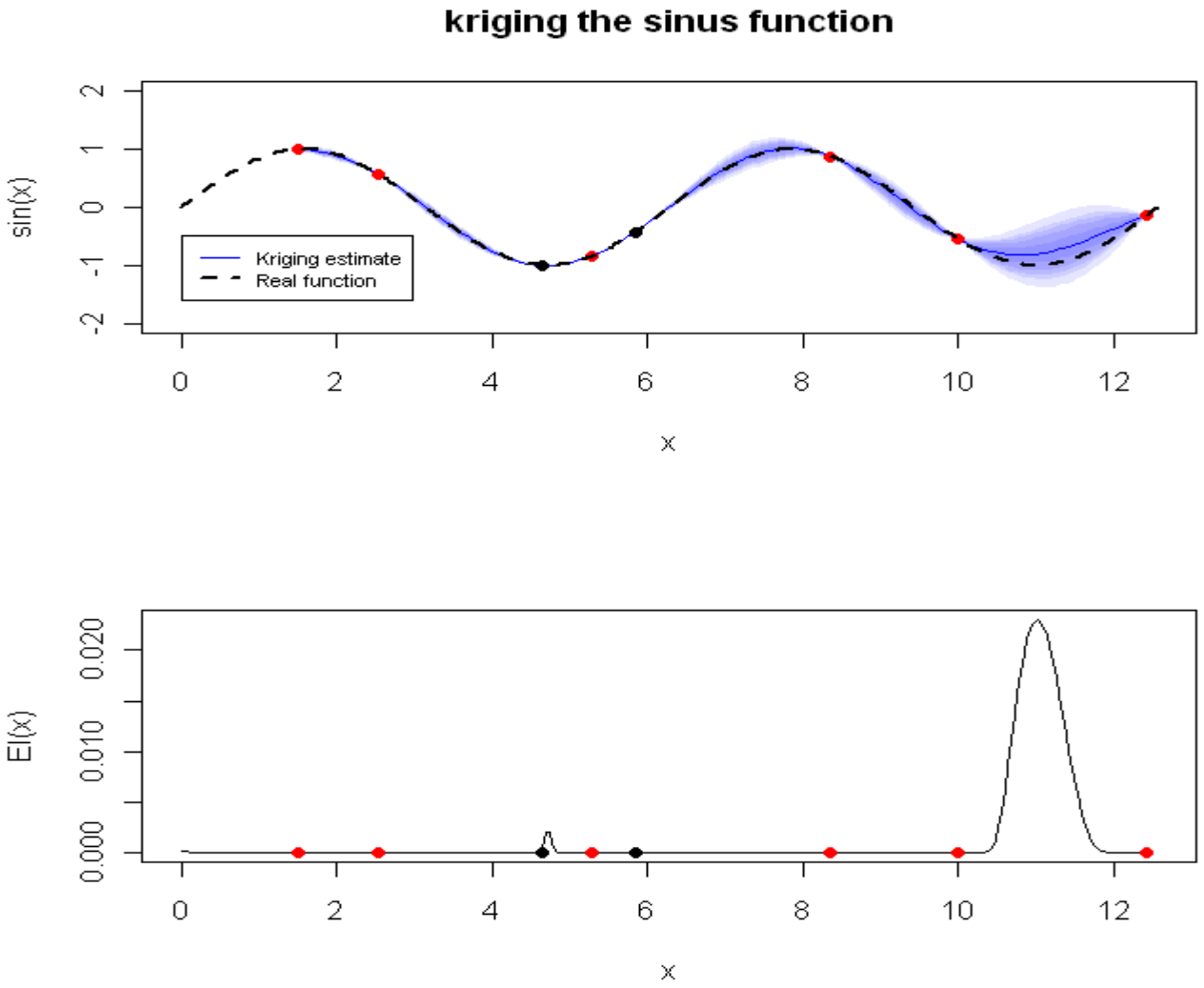
- EGO: step 1



Algorithme EGO

[Chevalier, 2011]

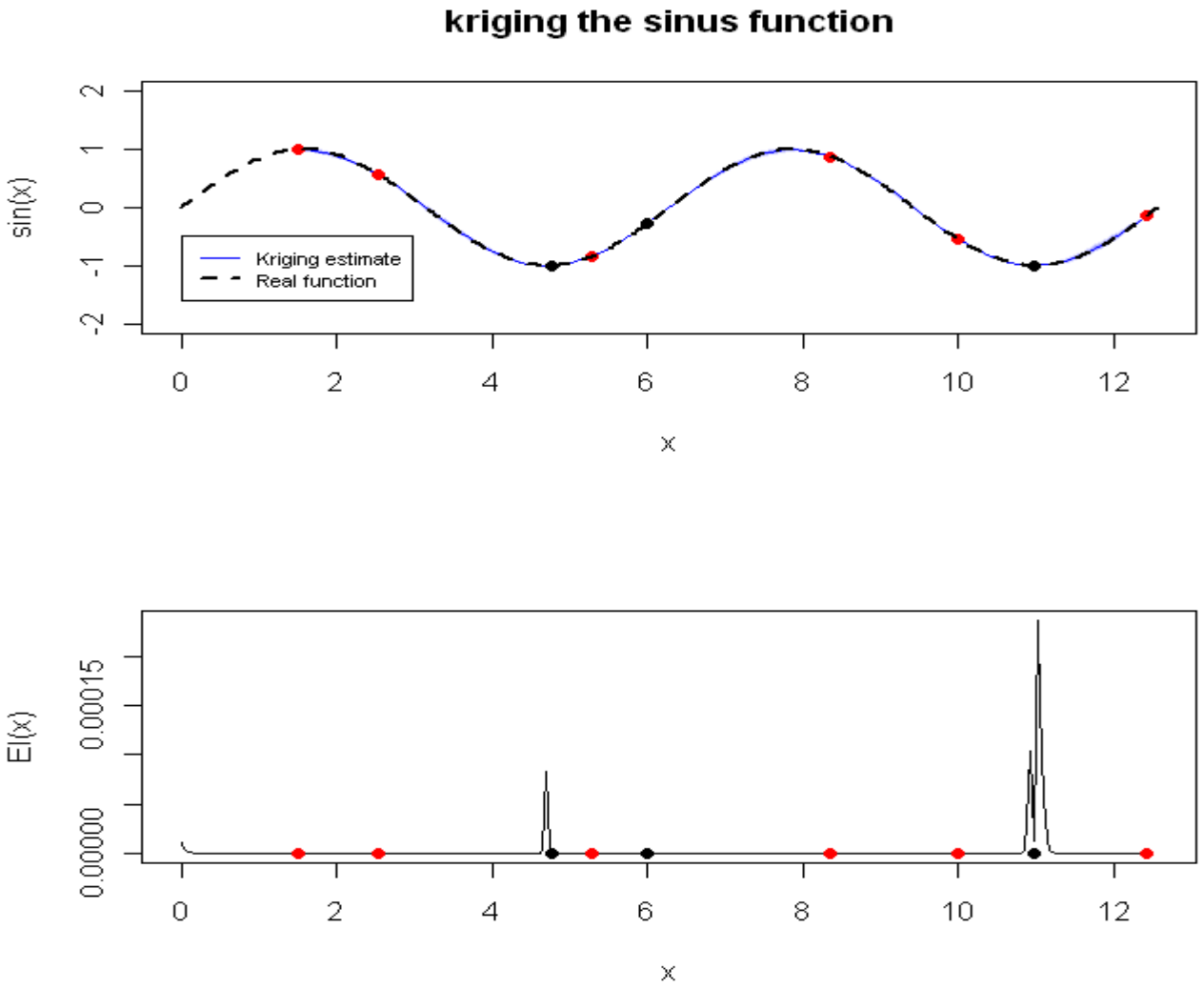
- EGO: step 2



Algorithme EGO

[Chevalier, 2011]

- EGO: step 3



Plan du cours 3

1. Introduction
2. Méthode d'interpolation spatiale par krigeage
3. Le métamodèle « processus gaussien »
4. **Un exemple d'application en hydrogéologie**

Métamodèles : les différentes étapes

1. Détermination du domaine de variation des variables d'entrée influentes (et éventuellement de leur loi de proba)
2. Choix d'un type de métamodèle
3. Choix d'un type de plan d'expériences numériques
4. Évaluation du code pour le plan d'expériences choisi
5. Construction du métamodèle à partir des expériences simulées
6. Validation du métamodèle
7. Exploitation du métamodèle

Les étapes 2 et 3 sont guidées par le problème traité *(analyse d'incertitudes, calcul de sensibilité, outil de prédiction, évaluation d'événements rares, optimisation, …)*

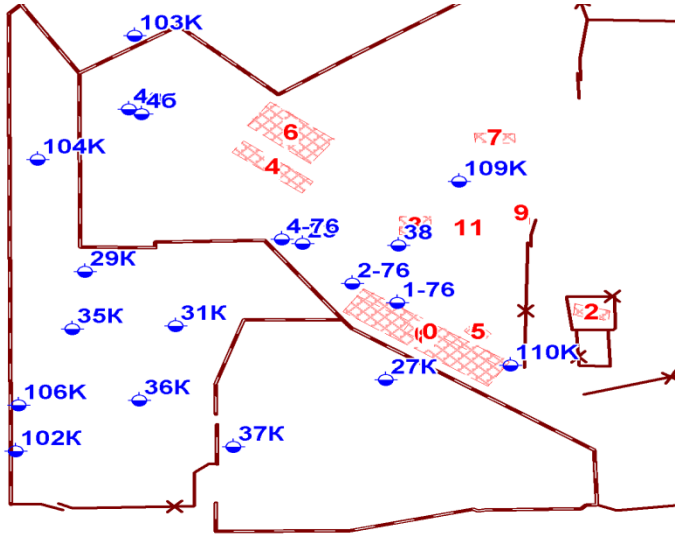
A l'issue de l'étape 6, on peut revenir à l'étape 3 (plan adaptatif)

Nota Bene : l'un des intérêts importants du métamodèle est de donner la possibilité d'étudier l'impact du choix de la distribution des entrées

Site de stockage temporaire de déchets radioactifs

Collaboration Institut Kurchatov/CEA
[Volkova et al. 08]

- Site (2 ha) situé aujourd'hui en banlieue de Moscou
- De 1943 à 1974 : stockage de déchets radioactifs solides
- Reconnaissance du site vers 1990 : réseau de 20 piézomètres
- Contamination nappe supérieure ^{90}Sr



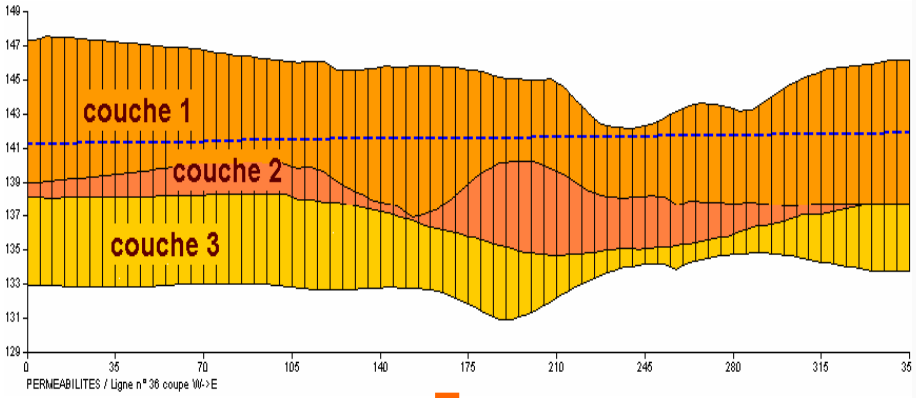
Estimation de l'impact de la contamination sur l'environnement

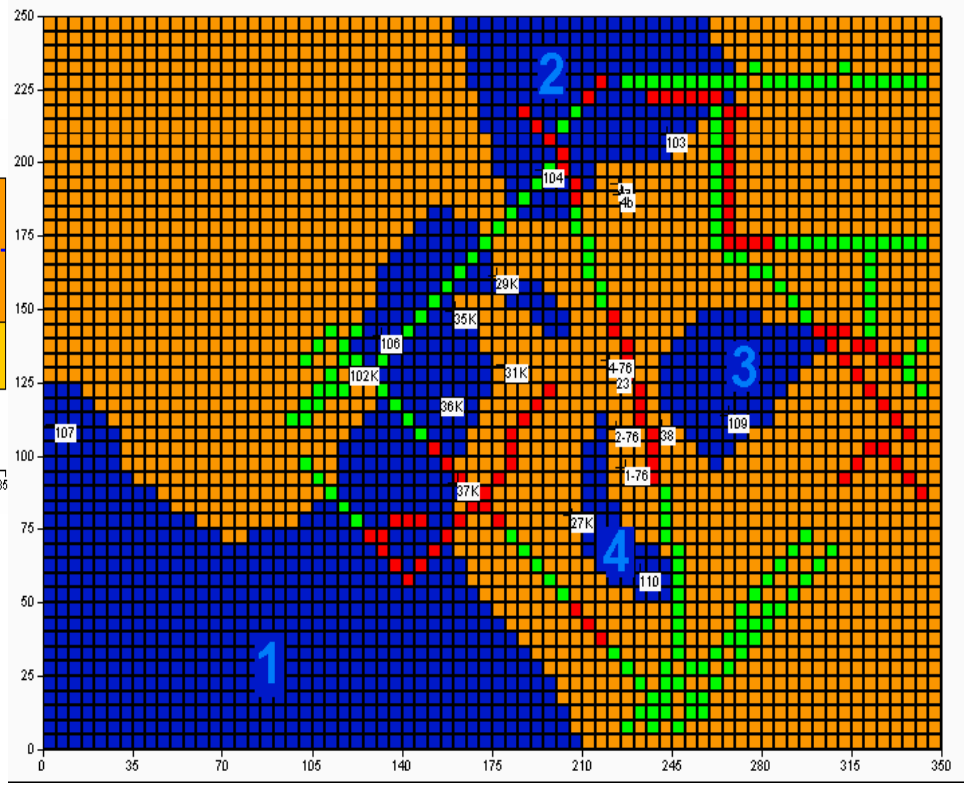
(degré de contamination de la nappe)

Faut-il réhabiliter le site ?
(excavation des déchets et traitement des sols)

Présentation du modèle du site de Kurchatov (1/2)

Modélisation géologique





Identification de 20 paramètres :

Perméabilité, porosité, coefficient Kd, intensité d'infiltration…

&

Incertitudes associées

(lois de proba. obtenues par des données ou de l'avis d'expert)

- zones d'absence de la couche 2 (zones numérotées de 1 à 4)
- présence de la couche 2
- localisation d'infiltrations modérées au niveau de canalisations
- localisation d'infiltrations fortes au niveau de canalisations
- p4a piézomètres

Présentation du modèle du site de Kurchatov (2/2)

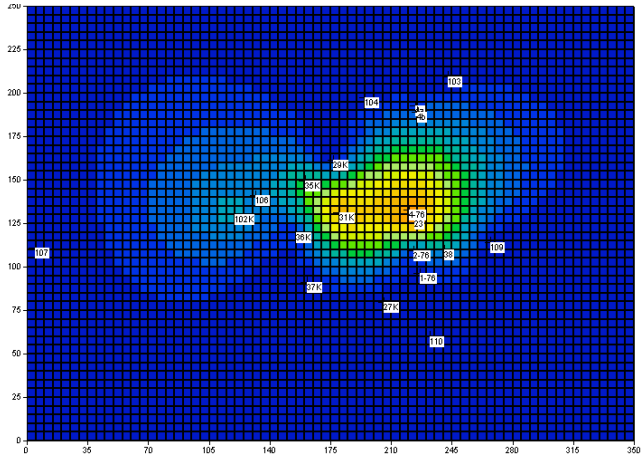
Logiciel MARTHE (BRGM) : **M**odélisation d'**A**quifère par un maillage **R**ectangulaire en régime **T**ransitoire pour le calcul **H**ydrodynamique des **E**coulements

250m × 350m, 200 pas de temps

Ecoulement transitoire 3D ; convection-dispersion ; sorption linéaire ; décroissance radioactive ; pas de terme source

Exemple d'un calcul (20mn)

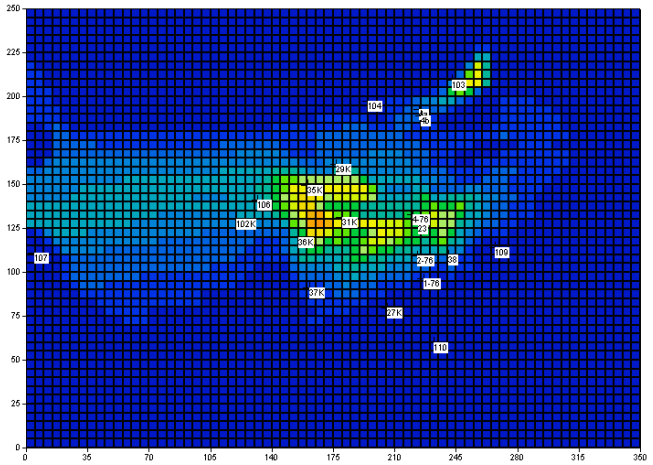
August 2002



Cartes de Concentrations

+ Jeu de 20 variables d'entrée

December 2010



Variables de sortie intéressantes :

1. Concentrations aux 20 piézomètres (20 sorties scalaires)
2. **Concentration spatiale discrétisée (64x64 = 4096 valeurs)**

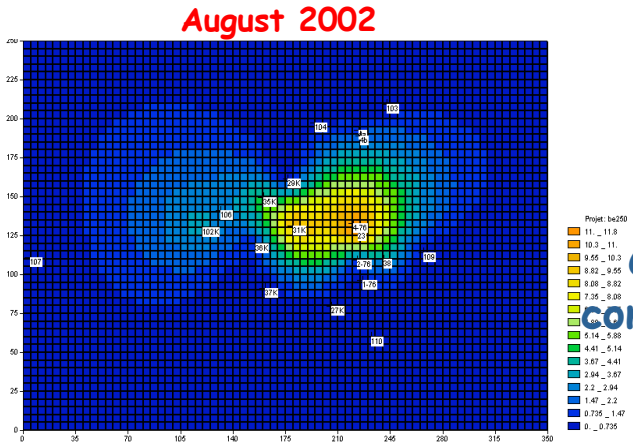
Etape B: Quantification des sources d'incertitudes

- Variables d'entrée du modèle

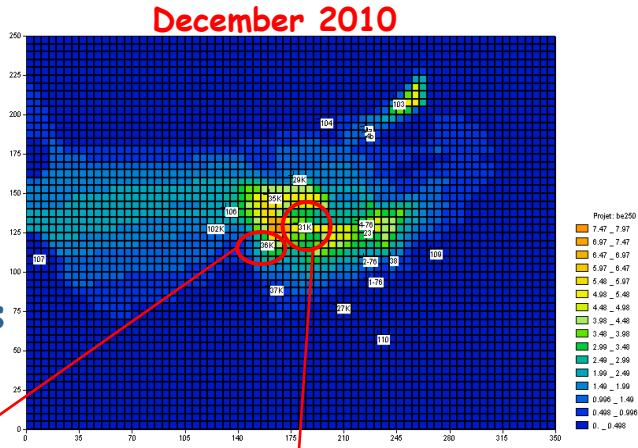
	Paramètres	Indicateur	Valeur du modèle	Type de distribution	Intervalle ou paramètres de distribution
1	Perméabilité couche 1	per1	8	Uniforme	1 - 15
2	Perméabilité couche 2	per2	15	Uniforme	5 - 20
3	Perméabilité couche 3	per3	8	Uniforme	1 - 15
4	Perméabilité zone 1	perz1	8	Uniforme	1 - 15
5	Perméabilité zone 2	perz2	8	Uniforme	1 - 15
6	Perméabilité zone 3	perz3	8	Uniforme	1 - 15
7	Perméabilité zone 4	perz4	8	Uniforme	1 - 15
8	Dispersivité longitudinale couche 1	d1	0,8	Uniforme	0,05 - 2
9	Dispersivité longitudinale couche 2	d2	0,8	Uniforme	0,05 - 2
10	Dispersivité longitudinale couche 3	d3	0,8	Uniforme	0,05 - 2
11	Dispersivité transversale couche 1	dt1	0,08	Uniforme	0,01*d1 - 0,1*d1
12	Dispersivité transversale couche 2	dt2	0,08	Uniforme	0,01*d2 - 0,1*d2
13	Dispersivité transversale couche 3	dt3	0,08	Uniforme	0,01*d3 - 0,1*d3
14	Coefficient de partage volumique c. 1	kd1	5,1	Weibull	1.1597, 19.9875
15	Coefficient de partage volumique c. 2	kd2	0,34	Weibull	0.891597, 24.4455
16	Coefficient de partage volumique c. 3	kd3	5,1	Weibull	1.27363, 22.4986
17	Porosité tous les couches	poros	0,3	Uniforme	0,3 - 0,37
18	Infiltration type 1	i1	0,0001	Uniforme	0 - 0,0001
19	Infiltration type 2	i2	0,004	Uniforme	i1 - 0,01
20	Infiltration type 3	i3	0,02	Uniforme	i2 - 0,1

Cas 1 – Sorties scalaires : analyse d'incertitudes en 2010

Concentration initiale

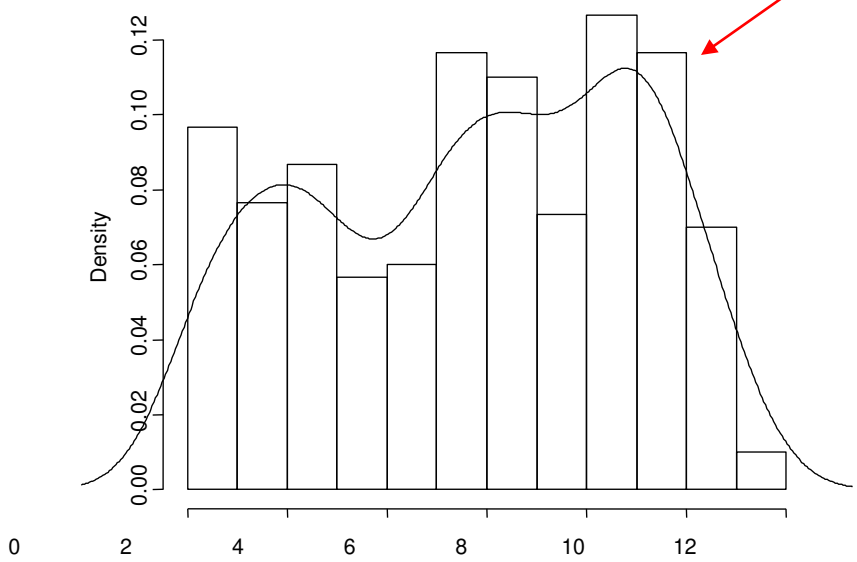


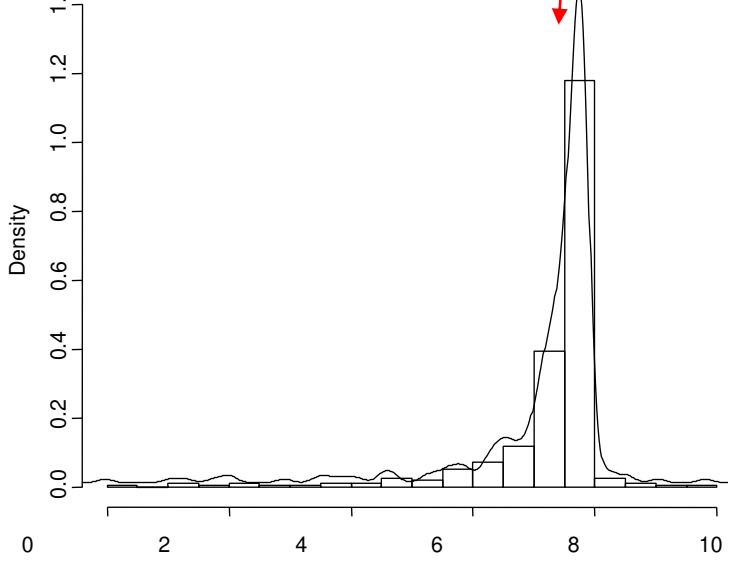
Cartes de concentrations



N = 300 calculs de type LHS

distributions des concentrations aux piézomètres (Bq/l)

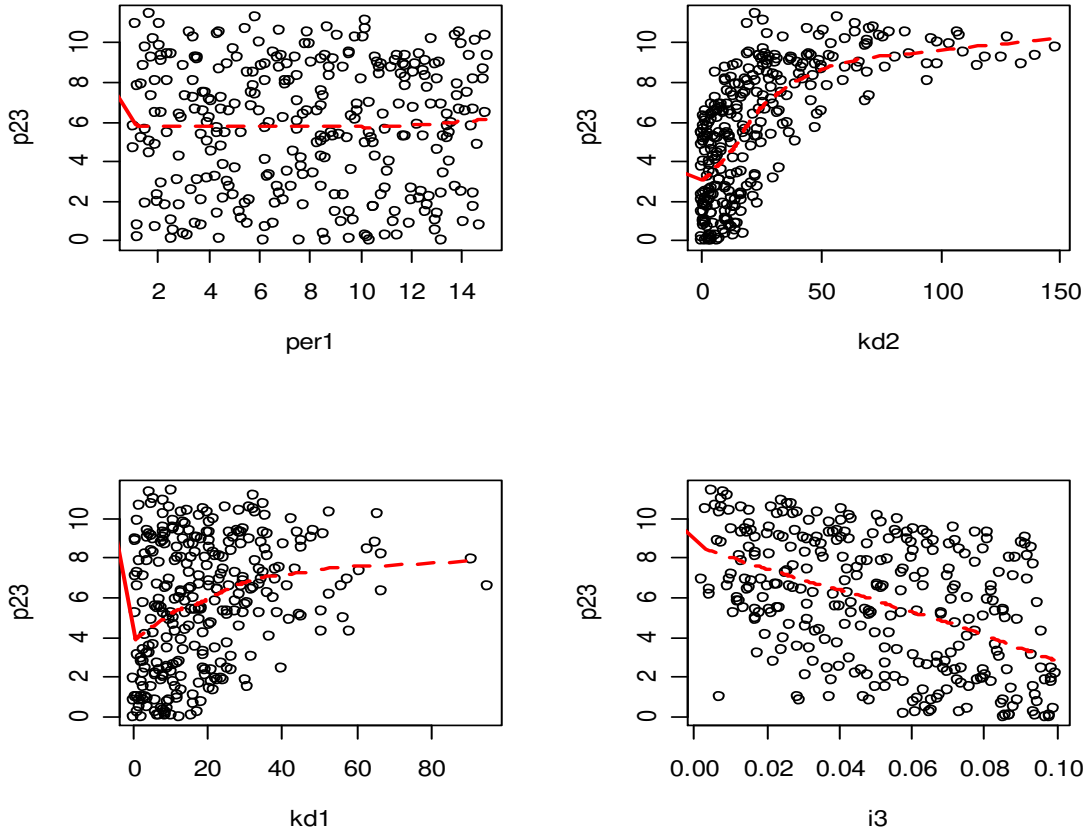




Analyse graphique : scatterplots piézomètre p23

Parfois de simples diagrammes sortie / entrées sont suffisants

1 sortie (p23) - p = 4 entrées
N = 300 calculs de type Monte Carlo (toutes les entrées X_i varient en même temps)

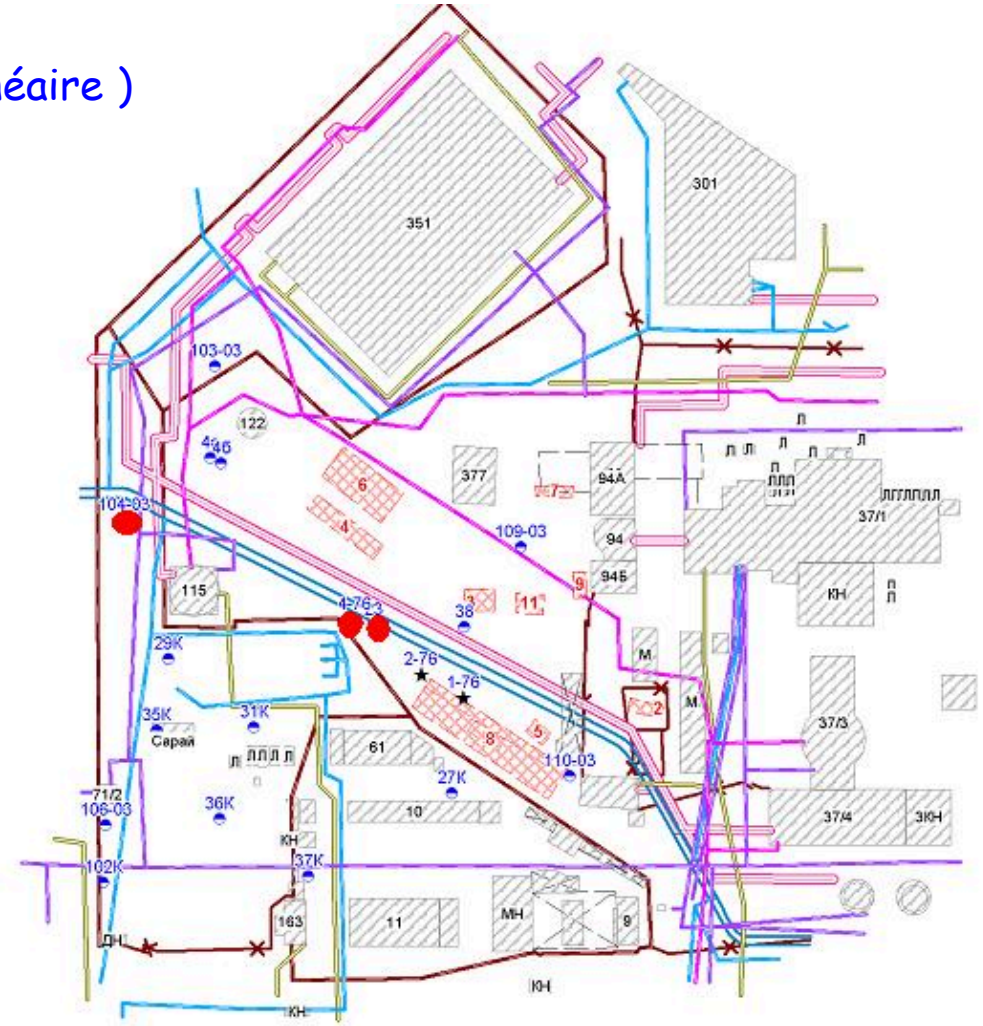


Etape C': Analyse de sensibilité

- Sorties avec de grands R^2 (relation linéaire)
  - p23 (R^2 = 0,78) p104 (R^2=0,68)
  - p4-76 (R^2 = 0,71)

- Sorties avec de grands R^{2*} (relation monotone)
  - 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)

- Entrées les + influentes
  - Distribution coefficient, layer 2
  - Distribution coefficient, layer 1
  - Infiltration intensity
  - permeability, layer 2



PROBLEME: 14 sorties non monotones

Indices de sensibilité pour la sortie « Piézomètre p104 »

Ajustement d'un métamodèle PG : R^2 = 93% (régression linéaire R^2 = 68%)

Analyse de sensibilité basée sur le PG

Estimation des indices de Sobol + construction des intervalles de prédiction

(en %)	SRC_i^2 (régression linéaire)	S_i (PG prédicteur)	$\mu_i = E_\Omega[\widetilde{S}_i]$ (PG global)	IC- 90% $(\widetilde{S}_i)$ (PG global)
per1	2	8	8	[5 ; 11]
kd1	52	76	69	[56 ; 83]
I3	13	15	13	[10 ; 17]

- Le métamodèle PG permet d'avoir une estimation + fiable des indices de sensibilité qu'en utilisant la régression linéaire (SRC ou SRRC)
- L'obtention d'IC avec le PG permet d'intégrer l'erreur résiduelle due au remplacement du code par le métamodèle

Cas 1 - Sorties scalaires – Résultats de l'analyse de sensibilité (2/2)

Entrées les + influentes

- **Groupe 1 : kd1**

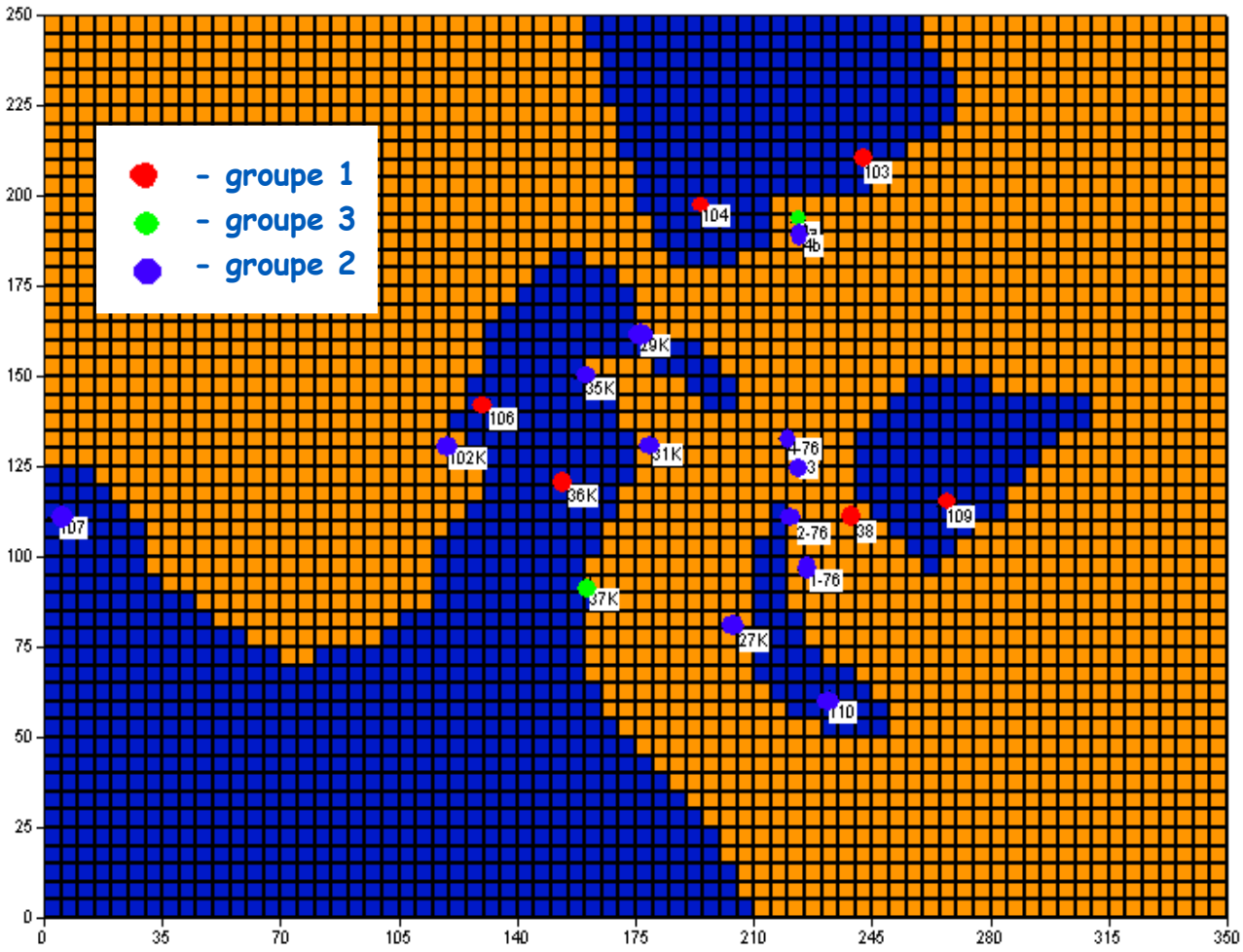
(coef. de partage de la couche 1)

- **Groupe 2 : kd2**

(coef. de partage de la couche 2)

- **Groupe 3 : i3**

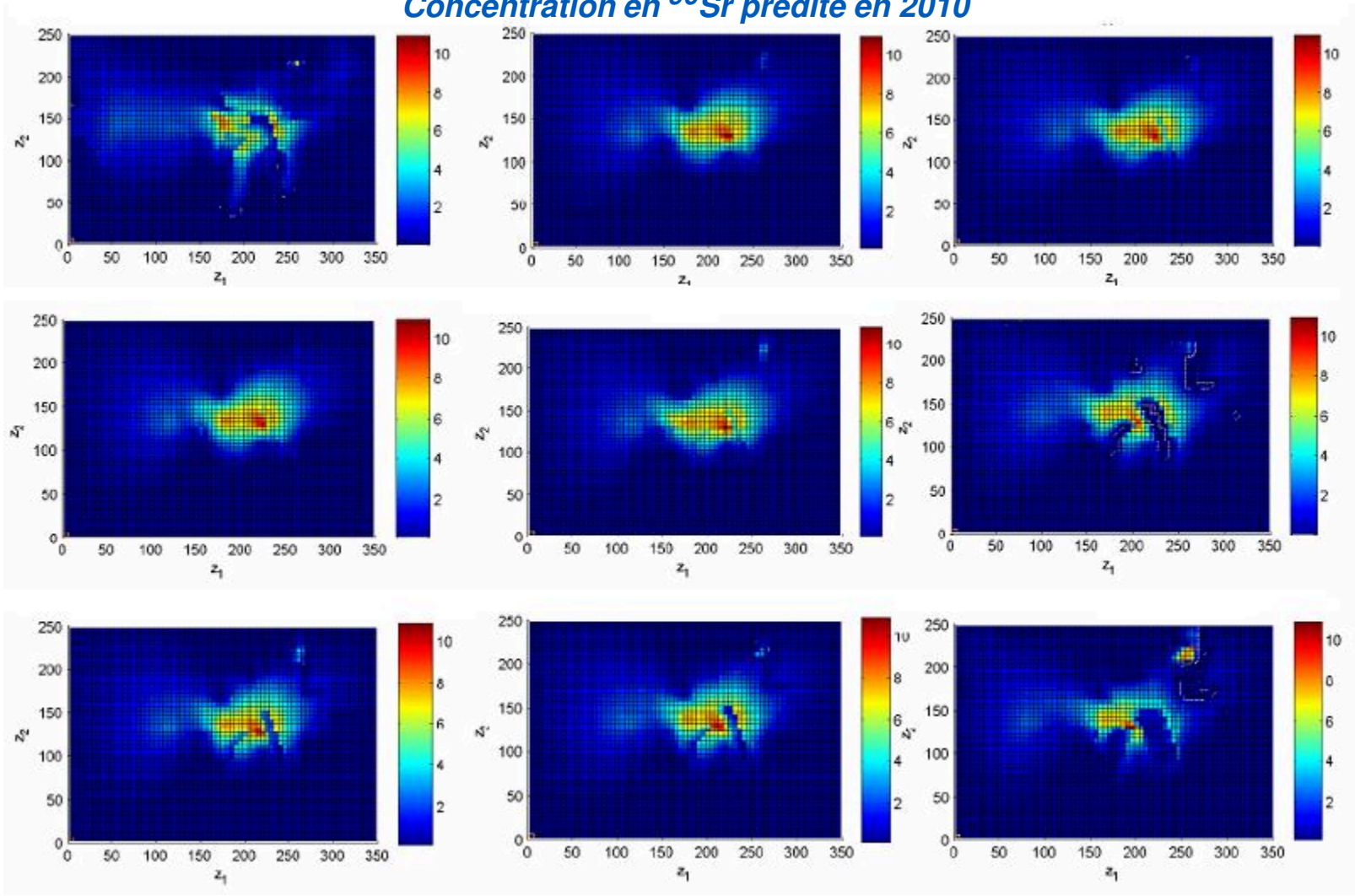
(intensité d'infiltration)



Cas 2 : sortie spatiale

Quelques exemples de cartes obtenues en sortie du calcul

Concentration en ^{90}Sr prédite en 2010



Méthodologie pour sorties fonctionnelles (1/2)

Sortie spatiale de 4096 pixels ⟹ sortie fonctionnelle 2D

Comment traiter une sortie fonctionnelle ?

- **Utilisation de la discrétisation complète de la fonction**
  - Construction d'un métamodèle puis analyse de sensibilité en chaque point de discrétisation

⟹ Possible mais peut être très coûteux en fonction du métamodèle

⟹ **Synthèse de l'information ou isolement de l'info principale**

- **Remplacer la fonction par quelques paramètres d'intérêt** (valeur finale, max, moyenne, …)
  - Exploitation réduite, fortement liée à la problématique de départ

- **Décomposition dans une base fonctionnelle** (Fourier, ondelettes,…)

Méthodologie pour sorties fonctionnelles (2/2)

- **Etape 1 : Décomposition spatiale des 300 cartes**
  - Centrage des cartes (moyenne empirique)
  - Décomposition sur une base d'ondelette (Daubechies)
  - Tri des coefficients par valeur moyenne en norme L_2

- **Etape 2 : Modélisation des coefficients en fonction de X**
  - Modélisation des 100 premiers coefficients par métamodèle PG (contrôle de la prédictivité par Q_2)
  - Modélisation des coefficients 101 à 1000 par régression linéaire simple (avec sélection par AIC)

- **Etape 3 : Prédiction pour un nouveau jeu d'entrée x^***

  x^* => prediction des coefficients => reconstitution de la carte

Analyse de sensibilité :

Obtention de cartes spatiales d'indices de Sobol

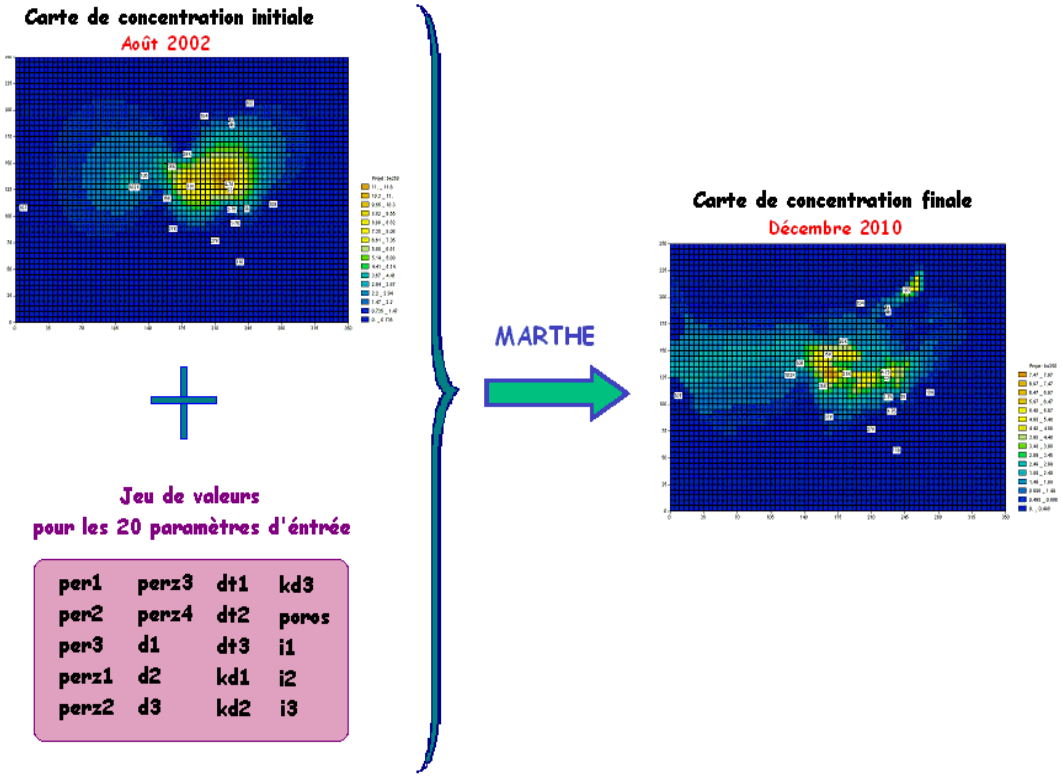
Application

N = 300 simulations

p = 20 entrées

K = 4096 pixels

k = 100 coefficients d'ondelettes modélisés par proc. gaussien



Prédictivité moyenne (métamodèle fonctionnel): Q_2 = 72%

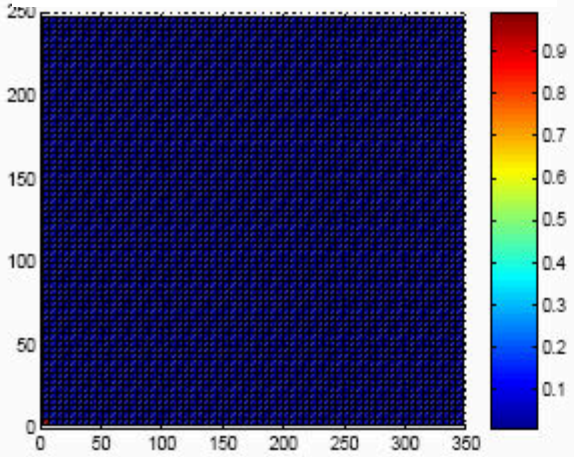
Estimation des cartes d'indices de Sobol du 1er ordre et d'ordre total (22000 appels au métamodèle)

→ 20 cartes d'indices de sensibilité

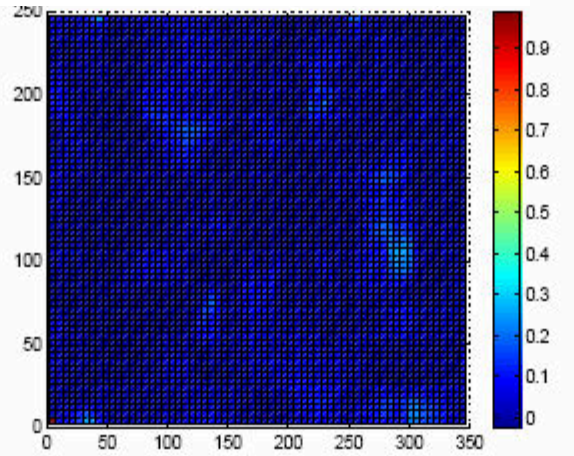
Cas 2 - Sortie spatiale – Résultats de l'analyse de sensibilité (1/3)

Cartes d'indices de Sobol du 1er ordre pour 6 entrées

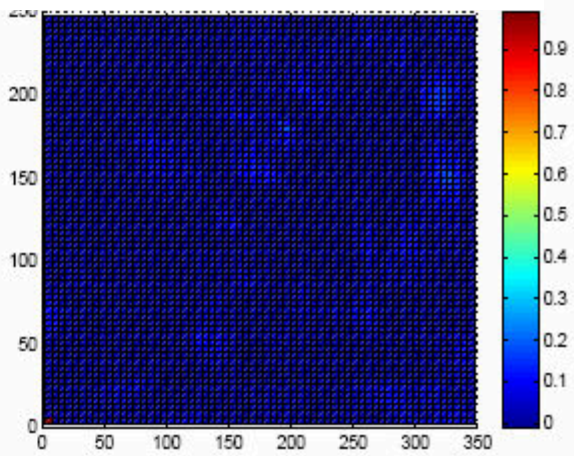
Perméabilité couche 1



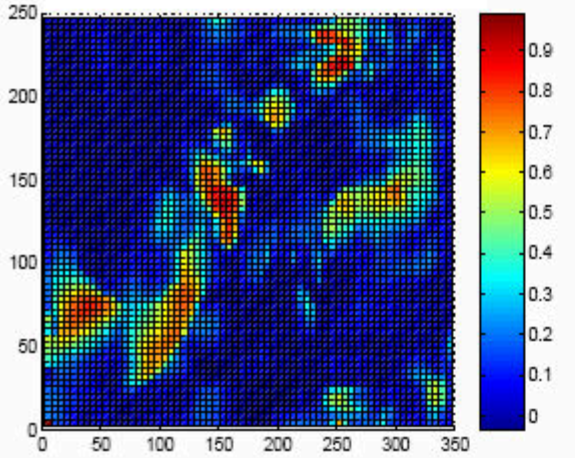
Perméabilité couche 2



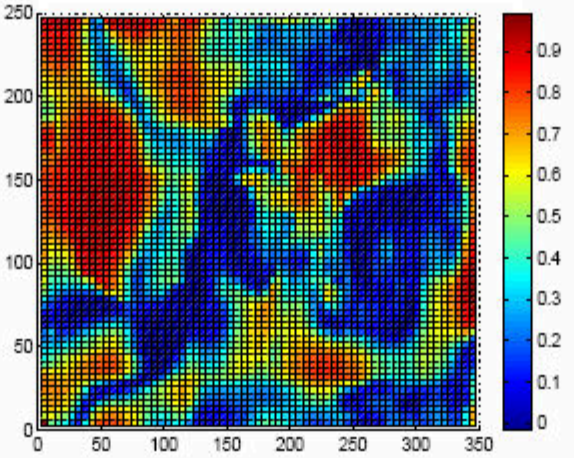
Perméabilité couche 3



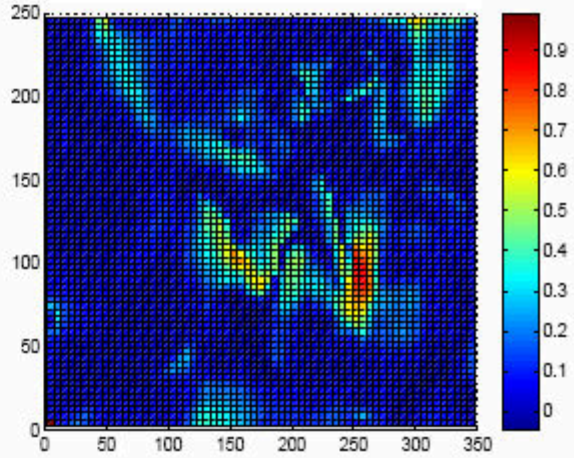
Kd couche 1



Kd couche 2



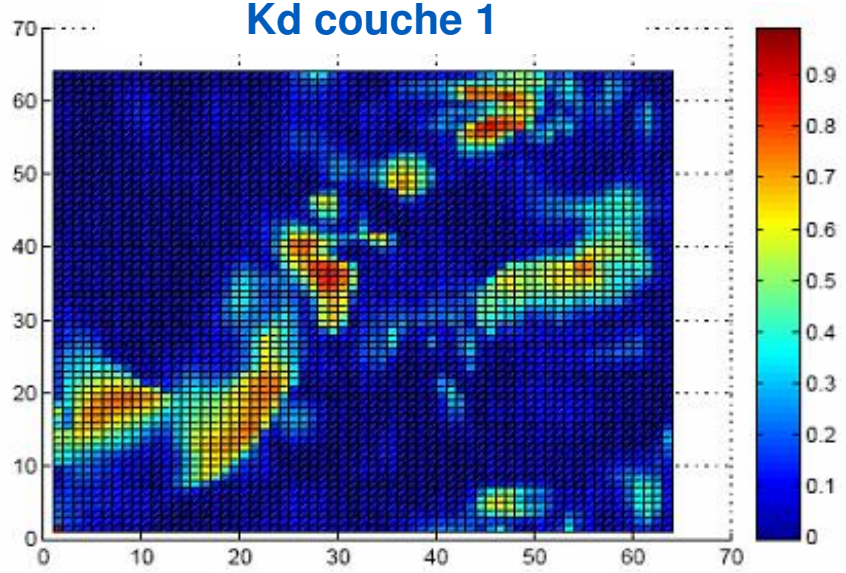
Infiltration forte (rouge)



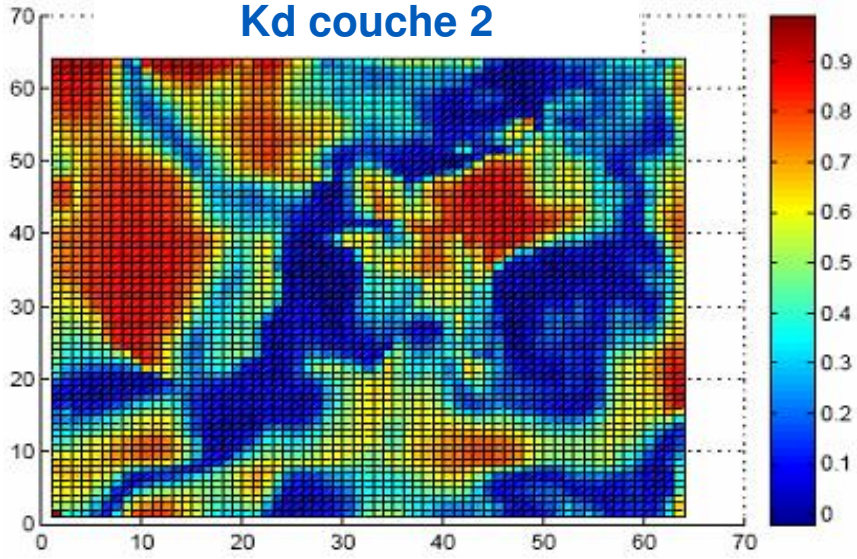
Cas 2 - Sortie spatiale – Résultats de l'analyse de sensibilité (2/3)

Cartes d'indices de Sobol pour les entrées kd1, kd2 et i3

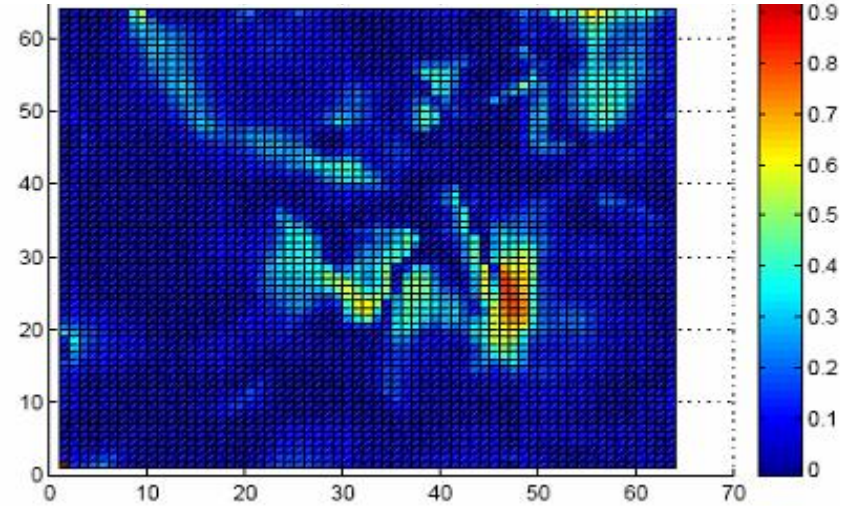
Kd couche 1

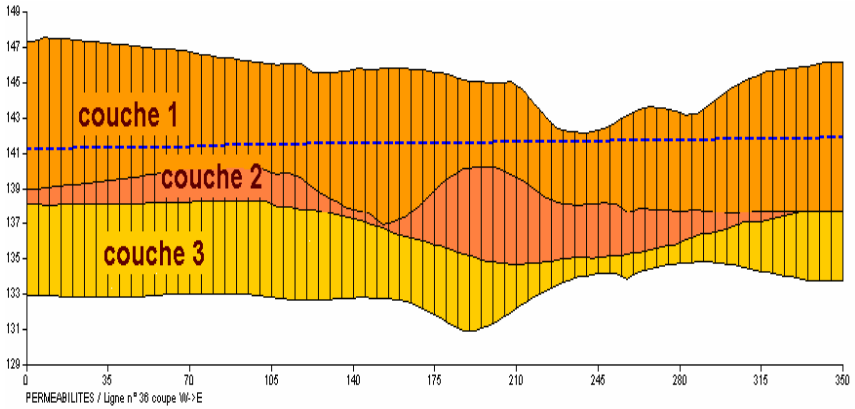


Kd couche 2



Infiltration forte *(canalisations rouges)*

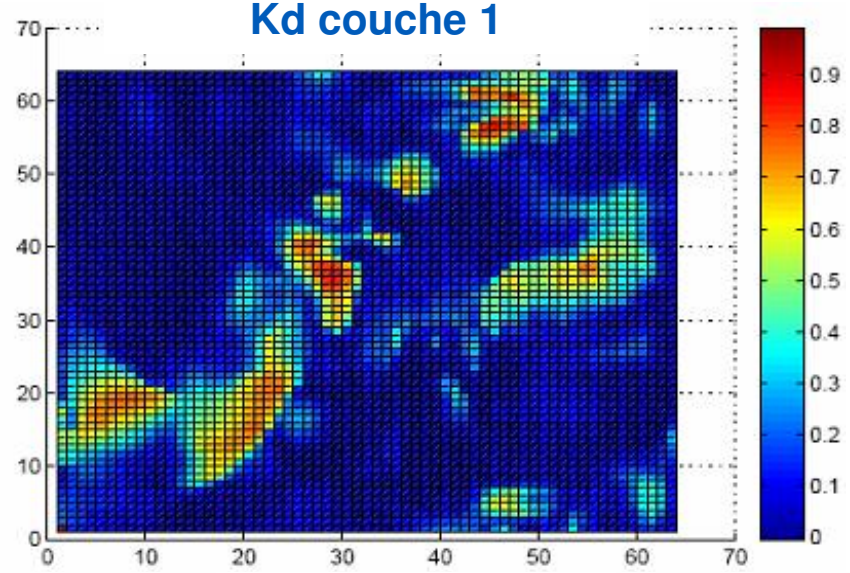




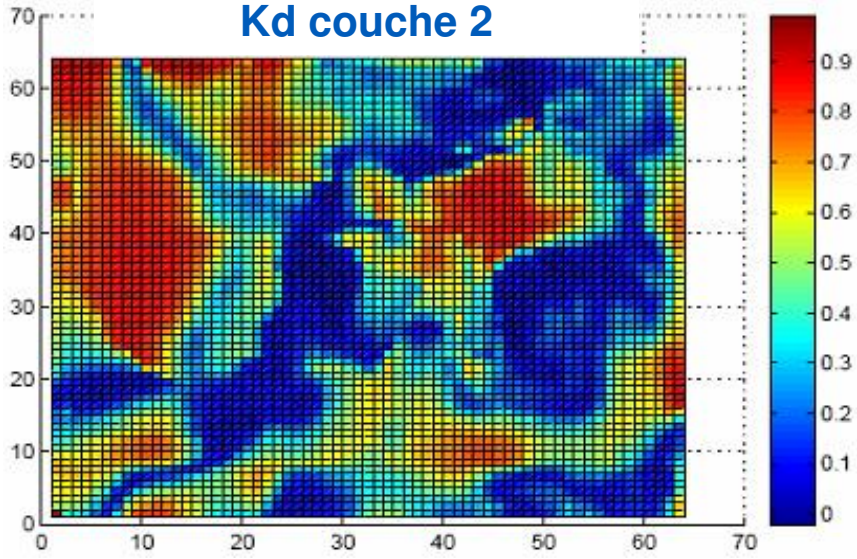
Cas 2 - Sortie spatiale – Résultats de l'analyse de sensibilité (3/3)

Cartes d'indices de Sobol pour les entrées kd1, kd2 et i3

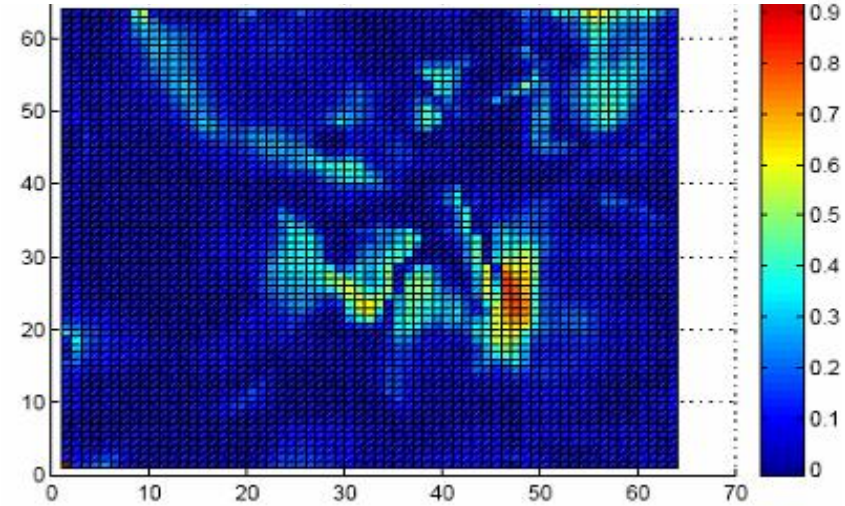
Kd couche 1

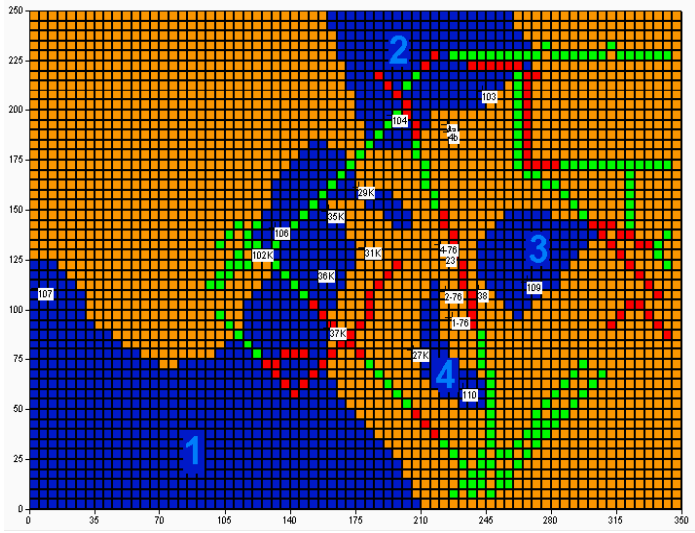


Kd couche 2



Infiltration forte *(canalisations rouges)*





Conclusions principales de l'étude du site de Kurchatov

1. Les prédictions de la contamination au bord du modèle ne dépassent pas le seuil réglementaire (jusqu'en 2010)

2. Les coefficients de partage sont les paramètres les plus influents et interagissent peu avec les autres paramètres
   Réduction de leur incertitudes ⇒ *réduction de l'incertitude de prédiction*

3. L'influence de la forme de la 2ème couche est forte, on l'étudie en faisant varier la forme par simulation géostatistique
   ⇒ Analyse de sensibilité avec champs aléatoires en entrée

4. Certaines mesures sont en dehors de la plage de variation de calculs.

Grâce au métamodèle, on a pu montrer que ça n'est pas du à une mauvaise modélisation des incertitudes des entrées.

⇒ *revenir sur la modélisation hydrogéologique*
⇒ *et/ou prendre en compte l'incertitude sur la carte initiale des concentrations*

Crédits & Bibliographie

- Formation « Démarche Incertitudes », IMdR-LNE
- Chilès & Delfiner, Geostatistics, Wiley, 1999
- Fang, Li & Sudjianto, D*esign 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
- Volkova et al., *Stoch. Environ. Res. Risk Assess.,* 2008

Ce cours est disponible sur : http://www.gdr-mascotnum.fr/doku.php?id=iooss1#academic