Modélisation des codes de calcul dans le cadre des processus gaussiens

Amandine Marrel

Laboratoire de Modélisation des Transferts dans l'Environnement
CEA Cadarache

Introduction (1)

Fiabilité et calcul d'incertitudes :

- **Processus de modélisation**
  - Phénomène réel représenté par des équations déterministes
    Ex : Équations de chimie transport
  - Modèle mathématique
  - Implémentation : obtention d'un code de calcul
    Ex : Calcul du transport de polluant en milieu poreux, calcul d'impacts environnementaux…

→ **Différentes sources d'incertitudes**

- **Analyse d'incertitudes et de sensibilité**
  - Étude du comportement de la réponse du modèle par rapport aux incertitudes sur les variables d'entrée
  - Identification des variables les plus influentes

Introduction (2)

Problème :

- Modèle souvent complexe
- Grand nombre de paramètres d'entrée
- Études d'incertitudes et de sensibilité

nécessitent un grand nombre de simulations → **Coûteux en temps de calcul**

Solution : Construction d'une surface de réponse

- **Remplacer le code de calcul par une fonction plus simple dans le domaine de variation des paramètres influents**
- Caractéristiques de cette fonction :
  - § être **représentative du code**
  - § avoir de **bonnes capacités de prédiction**
  - § nécessiter un **temps de calcul négligeable**

Introduction (3)

Quelle surface de réponse ?

Polynômes, splines, GLM, réseaux de neurones, arbres de régression…?

- Limites des modèles de régression classiquement utilisés
- Difficulté d'interprétation
- Nécessité de disposer d'un prédicteur rapide à évaluer pour réaliser des analyses de sensibilité
- Codes de calcul déterministes : intérêt d'une modélisation statistique permettant d'interpoler exactement les données

↓

Candidat intéressant :
les processus stochastiques gaussiens (krigeage)

Étude des capacités de prédiction des processus gaussiens :

- Mise au point d'une méthodologie
- Application : code de transport de polluants dans les sols

Plan

1. **Théorie des processus gaussiens**
   - Modèle général
   - BLUP
   - Estimation des paramètres
2. **Mise en œuvre**
   - Estimation, prédiction et validation
   - Sélection de variables
   - Algorithme général
3. **Tests sur les données Marthe**
   - Présentation des données
   - Modélisation
   - Résultats et analyse

Conclusion et perspectives

Modèle général (1)

Processus stochastique gaussien :

$Y(X) = g(X) + Z(X)$

Partie régression — Partie stochastique

Z processus stochastique tel que :

- $E[Z(x)] = 0$
- $Cov(Z(x), Z(y)) = \sigma^2 R(x,y)$

où σ^2 est la variance et R la fonction de corrélation

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

Partie régression :

- Choix d'un modèle paramétrique $g(X) = \beta.F(X)$, β vecteur des paramètres
  Ex pour F : Polynôme de degré 0, 1 ou 2, base de Fourier…
- Importance limitée ?

- Notations :
  - Sortie Y (ou réponse du code de calcul)
  - Variable d'entrée X de dimension d : $X = [X_1, \ldots, X_d]$

Modèle général (2)

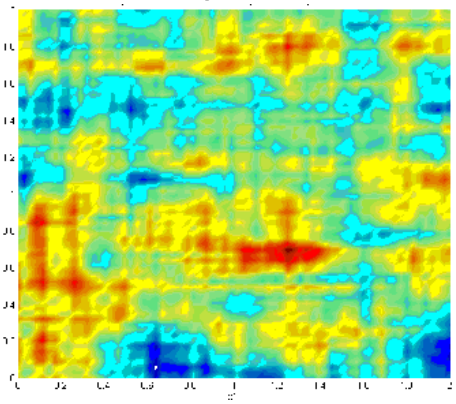
Partie stochastique :

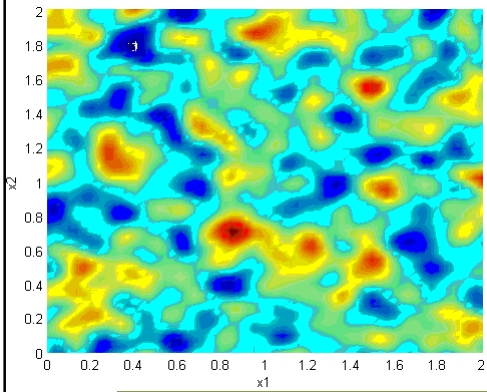
- Hypothèse de normalité du processus
  - ⇒ Entièrement caractérisé par la fonction de covariance
- Hypothèse de stationnarité
  - ⇒ *R(x,y) = R(x-y)*
- Hypothèse d'isotropie :
  - ⇒ *R(x,y) = R(‖ x - y‖)* infondée pour les codes de calcul
- Différentes formes possibles pour la fonction de covariance :
  - Linéaire : $R(x,y) = \max\left\{1 - \sum_{i=1}^{d} \theta_i |x_i - y_i|, 0\right\}$
  - Exponentielle généralisée : $R(x,y) = \exp\left(-\sum_{i=1}^{d} \theta_i |x_i - y_i|^{p_i}\right)$
  - Sphérique
  - ….
- Effet de pépite : $\tilde{R}(x,y) = R(x,y) + \delta$

Modèle général (3)

Exemples de processus gaussiens centrés isotropes 2D

Fonction de covariance exponentielle de paramètre $\theta = 3$ ⇒





⇐ Fonction de covariance gaussienne de paramètre $\theta = 100$

BLUP

Meilleur prédicteur linéaire sans biais (BLUP)

- Formulation : $\hat{Y}(x) = \beta F(x) + r(x)R^{-1}\left[Y - \beta F\right]$
- Interpolateur exact
- Formulation du MSE : $MSE[\hat{Y}(x)] = \sigma^2\left(1 + {}^t u(x)({}^t FR^{-1}F)^{-1}u(x) - {}^t r(x)R^{-1}r(x)\right)$

  avec $u(x) = {}^t FR^{-1}r(x) - f(x)$
- Caractérisé par des paramètres qualitatifs et quantitatifs:
  - Forme de la fonction de covariance
  - Paramètres de régression $\boldsymbol{\beta}$
  - Paramètres de covariance $\boldsymbol{\theta}$
  - Paramètre de variance $\boldsymbol{\sigma}$

  → **Estimation de ces paramètres**

Notations :
- Modèle $Y(X) = \beta F(X) + Z(X)$
- N observations : $X_S = [x_1, \ldots, x_N]$
- Base d'apprentissage : Y_S, X_S, $F = F(X_S)$
- Matrice de corrélation : $R = \left(R(x_i, x_k)\right)_{i,k}$
- Vecteur de corrélation : $r(x) = \left[R(x_1, x), \ldots, R(x_N, x)\right]$

Estimation des paramètres (1)

Méthode du maximum de vraisemblance

- Expression de la log-vraisemblance sous l'hypothèse gaussienne :

$$\ln L(\beta, \theta, \sigma) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\ln(\det R_\theta) - \frac{1}{2}\sigma^2\left[Y - \beta F\right]R_\theta^{-1}\left[Y - \beta F\right]$$

- Estimateur du maximum de vraisemblance

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

- Estimation conjointe de β et σ :

$$\begin{cases} \beta^* = \left[F' R_\theta^{-1} F\right]^{-1} F' R_\theta^{-1} Y \\ \sigma^{2*} = \dfrac{1}{N}\left[Y - \beta^* F\right]R_\theta^{-1}\left[Y - \beta^* F\right] \end{cases}$$

- Estimation du paramètre de corrélation θ :

$$\theta^* = \underset{\theta}{\text{Argmax}} -\frac{1}{2}\left[N\ln\left(\sigma^{2*}\right) + \ln\left|R_\theta\right|\right]$$

- Minimisation numérique de la fonction $\psi(\theta) = \left|R_\theta\right|^{-1/N}\sigma^{2*}$

Estimation des paramètres (2)

Optimisation numérique difficile :

- Problèmes de conditionnement de la matrice de corrélation R
  ⟹ Ajout dans la fonction de covariance d'un effet de pépite
- Fonction $\Psi(\theta)$ coûteuse
- Problèmes de minima locaux
- Grand nombre de paramètres (dimension du vecteur des paramètres d'entrées > 20)
  ⟹ Utilisation d'un algorithme d'optimisation itératif

Plan

Estimation, prédiction et validation

Estimation des paramètres par minimisation de $\Psi(\theta)$:

- **Algorithme stochastique basé sur la toolbox matlab DACEFIT**

  ⇒ basé sur la méthode de Hookes & Jeeves

Implémentation rapide du prédicteur au point x (EBLUP):

$$\begin{cases} \hat{Y}(x) = \hat{\beta}F(x) + {}^{t}r_{\hat{\theta}}(x)\hat{\gamma} \\ \text{avec } \hat{\gamma} = R_{\hat{\theta}}^{-1}\left[Y - \hat{\beta}F\right] \end{cases}$$

⇒
- $\hat{\beta}$ et $\hat{\gamma}$ indépendants de x
- Seulement $F(x)$ et ${}^{t}r_{\hat{\theta}}(x)$ à recalculer

Validation du modèle :

- Critères :
  - Analyse graphique des résidus
  - Coefficient de détermination R^2

$$R^2(Y,\hat{Y}) = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(\bar{Y} - Y_i)^2}$$

- Méthodes :
  - Validation croisée :

    ⇒ Estimation des paramètres sur la base d'apprentissage

    ⇒ Calcul du critère (R^2 par ex.) sur la base de test

Sélection de variables

Nombre de variables d'entrée important (> 10)

- Procédure d'estimation difficile (grand nombre de paramètres)
- Compromis « biais-variance » : modèle simple/modèle complexe

Mise en place d'une procédure de sélection de variables

- Sélection ordonnée des variables

  ex : variables d'entrée classées par coefficient de corrélation décroissant

- Sélection des variables de la partie régression :

  ⇒ Minimisation du critère AICC

$$AICC = -2\ln L\left(\hat{\beta},\hat{\theta},\hat{\sigma}\right) + 2N\frac{m_1 + m_2 + 1}{N - m_1 - m_2 - 2}$$

  - m_1 : nombre de variables explicatives dans la régression
  - m_2 : nombre de variables explicatives dans la covariance

- Sélection de variables de la fonction de covariance

  ⇒ Maximisation du R^2 calculé par validation croisée

Algorithme général

Procédure de modélisation par les processus gaussiens

- Etape 1 : Tri des variables d'entrée
- Etape 2 : Boucle sur les variables dans la fonction de covariance
  - Etape 2.1 : Boucle sur les variables dans la régression
    - Estimation des paramètres β, θ, σ
    - Calcul du critère AICC
  - Etape 2.1 : Sélection du Modèle de régression optimal
    → AICC minimal
  - Etape 2.2 : Calcul du R^2 par validation croisée
- Etape 3 : Sélection du modèle de covariance optimal
  → R^2 maximal

Plan

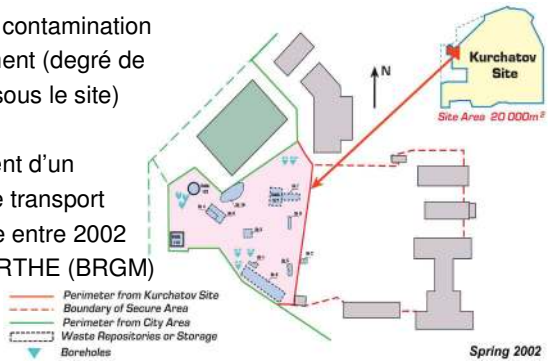
Présentation des données Marthe (1)

§ Issues d'une collaboration
Institut Kurchatov, CEA/DTN, CEA/DER

§ Site de stockage des déchets radioactifs sur le site de l'institut Kurchatov à Moscou

§ Estimation de l'impact de la contamination radioactive sur l'environnement (degré de contamination de la nappe sous le site)

§ Modélisation : développement d'un scénario d'écoulement et de transport généralisé de ^{90}Sr sur le site entre 2002 et 2010, avec le logiciel MARTHE (BRGM)

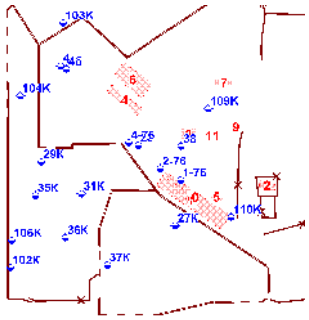


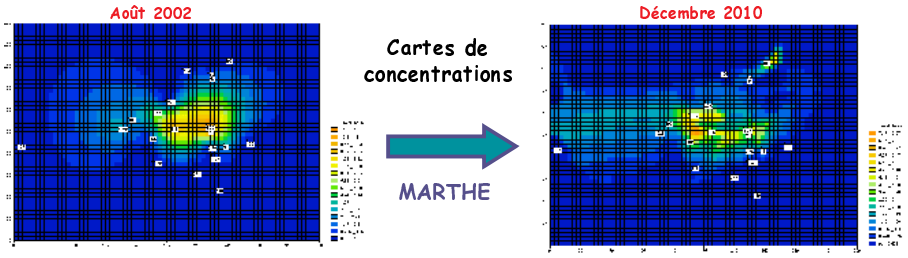
Présentation des données Marthe (2)

20 paramètres d'entrée :
perméabilité, dispersivités longitudinales et transversales, coefficients de partage volumique, porosité, intensités d'infiltration, …

20 variables de sortie :
concentrations aux 20 piézomètres calculées par le logiciel MARTHE en fin d'année 2010

300 simulations (LHS)





Modélisation

Comparaison de trois modèles :

- **Régression linéaire simple**
- **Boosting**
  - Agrégation d'arbres de régression
- **Processus gaussiens**
  - Fonction de régression : polynôme de degré 1
  - Fonction de covariance exponentielle généralisée

Validation :

- Base d'apprentissage de 250 données
- Base de test de 50 données
- Validation croisée à 6 échantillons

Quelques résultats…

Sortie	Régression linéaire R^2	Boosting R^2	Processus gaussien R^2
1	0.31	0.59	0.84
2	0.48	0,64	0.76
3	0.10	0.4	0.3
9	0.19	0,57	0.83
10	0.74	0,94	0.94
12	0.55	0,82	0.81
16	0.59	0,64	0.90
18	0.67	0,96	0.96
19	0.16	0.17	0.09

Performances supérieures à la régression

Capables de concurrencer et de dépasser le boosting

Inadapté à certaines sorties (où le boosting échoue aussi)

Plan

1. **Théorie des processus gaussiens**
   - Modèle général
   - BLUP
   - Estimation des paramètres
2. **Mise en œuvre**
   - Estimation, prédiction et validation
   - Sélection de variables
   - Algorithme général
3. **Tests sur les données Marthe**
   - Présentation des données
   - Modélisation
   - Résultats et analyse

Conclusion et perspectives

Conclusion

Points forts de la modélisation par des processus gaussiens :

- Modèle souple et performant
- Interpolateur exact
- Interprétation possible :
  - ⇒ Partie régression : tendance, comportement général
  - ⇒ Fonction de covariance : traduction des hétérogénéités locales
- Calcul rapide de prédicteur ⇒ Analyse de sensibilité facilitée
- Cadre gaussien : inférence sur les paramètres & estimation du MSE

Points faibles :

- Peu robuste aux valeurs extrêmes (dépendances spatiales)
- Difficultés de mise en œuvre
- Choix de la fonction de covariance a priori ?
- Choix de la fonction de régression a priori ?

Perspectives

Développement du modèle des processus gaussiens

- Extension à d'autres fonctions de régression et de covariance
- Amélioration de la sélection de variables
- Cokrigeage
- Inférence sur les paramètres
- Plan d'échantillonnage

Extension au cadre spatio-temporel

- Processus temporel en sortie
- Corrélation spatiale des sorties
- Champ aléatoire en entrée

Objectif final : Intégration des processus gaussiens dans la procédure de modélisation du transfert de polluants dans les sols