Annexe A

Présentation théorique des différentes méthodes d'estimation de la géostatistique linéaire

Méthodes d'estimation de la géostatistique linéaire

La géostatistique propose différentes techniques d'estimation locale. L'utilisateur est guidé dans le choix d'une méthode par le type de données disponibles et par les caractéristiques de ces données révélées dans l'étape d'analyse.

Dans une première partie, nous présentons les méthodes du krigeage ordinaire et du krigeage intrinsèque qui ne requièrent que les données de concentration.

Dans une seconde partie, nous montrons comment introduire des informations supplémentaires dans le modèle géostatistique.

1. ESTIMATION MONOVARIABLE

2.1 INTRODUCTION

Les données de concentration peuvent provenir de stations de mesure fixes, de moyens mobiles ou de tubes à échantillonnage passif. Elles sont décrites plus précisément au chapitre 4. Les AASQA réalisent presque exclusivement leurs cartographies à partir des données de tubes à diffusion. En effet, seul ce mode d'échantillonnage permet de collecter à des coûts acceptables un grand nombre de données dans l'espace, afin d'évaluer la qualité de l'air d'une agglomération, d'un département ou d'une région entière. L'information fournie par ces cartographies reste cependant limitée à la durée de l'échantillonnage, soit deux à quinze semaines environ.

Les données d'analyseurs (stations fixes et moyens mobiles) n'en constituent pas moins une aide précieuse pour exploiter les données de tubes et quantifier l'incertitude de ces dernières. Des techniques de calcul de l'incertitude, étudiées notamment par AIR NORMAND et fondées sur la norme NF/ISO 13752 ou, en ce qui concerne l'ozone, sur la norme NF ENV 13005, ont été appliquées par des AASQA à l'occasion de campagnes de mesure (ATMO Picardie et al., 2000, ASCOPARG, 2001). Cette incertitude, exprimée comme une variance de l'erreur de mesure, aide à ajuster le variogramme à l'origine et peut être incorporée dans la modélisation géostatistique.

2.2 KRIGEAGE ORDINAIRE

On suppose que la variable régionalisée peut être modélisée par une **fonction aléatoire stationnaire d'ordre 2 ou strictement intrinsèque** et l'on se place dans le cas le plus probable où la moyenne de cette fonction est inconnue dans le champ.

Principe du krigeage

Le krigeage ordinaire à moyenne inconnue a pour but de fournir une estimation locale **non biaisée la plus précise possible** de la variable régionalisée à l'aide d'une **combinaison linéaire pondérée des données expérimentales**.

Dans tout ce qui suit, Z désigne la variable régionalisée étudiée, Z^* son estimateur par krigeage, et Z_i les valeurs prises par Z aux points de données s_i.

γ est le modèle de variogramme ajusté pour Z.

Soit donc $Z^* = a + \sum_{i=1}^{n} \lambda_i Z_i$ un estimateur linéaire de la variable.

L'erreur d'estimation s'écrit : $\varepsilon = Z^* - Z$

Pour que cet estimateur réponde aux exigences du krigeage, le paramètre *a* et les pondérateurs λ_i doivent respecter les contraintes suivantes :

- L'erreur d'estimation est une combinaison linéaire autorisée, c'est-à-dire que son espérance et sa variance existent. Cela implique que a=0 et que $\sum_{i=1}^{n} \lambda_i = 1$. Cette dernière relation assure également **l'absence de biais**.

- L'estimateur est le plus précis possible au sens où **la variance de l'erreur d'estimation est minimale**, soit :

  $$\frac{\partial}{\partial \lambda_i} \mathrm{Var}(Z^* - Z) = 0 \quad \forall i \quad \text{avec } \mathrm{Var}(Z - Z^*) = 2\sum_i \lambda_i \gamma_{xi} - \sum_{i,j} \lambda_i \lambda_j \gamma_{ij}$$

  γ_{ij} est la valeur du variogramme entre les points s_i et s_j.
  γ_{ix} est la valeur du variogramme entre les points s_i et le point cible x.

Les poids λ_i qui remplissent ces conditions sont solutions d'un système matriciel faisant intervenir γ :

$$\begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1N} & 1 \\ \vdots & & \vdots & \vdots \\ \gamma_{N1} & \cdots & \gamma_{NN} & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_N \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma_{1x} \\ \vdots \\ \gamma_{Nx} \\ 1 \end{bmatrix}$$

Le paramètre μ est un multiplicateur de Lagrange qui intervient dans la résolution du système de krigeage et dans le calcul de la variance de l'erreur.

Cette variance de l'erreur d'estimation ou variance de krigeage est donnée par la relation :

$$\sigma_K^{\,2} = \sum_i \lambda_i \gamma_{ix} + \mu$$

Elle ne dépend que du modèle de variogramme et de la configuration relative du point ou du bloc à estimer et des données expérimentales.

Krigeage de blocs :

On peut rechercher une estimation ponctuelle de Z ou une estimation moyenne sur de petits volumes (ou surfaces) contigus appelés blocs. Notons alors Z_v^*, l'estimateur de Z dans le bloc v :

$$Z_v^{\,*} = \sum_{i=1}^{n} \lambda_i Z_i$$

Le krigeage de blocs obéit aux mêmes conditions qu'un krigeage ponctuel. Les poids λ_i sont solutions du système de krigeage :

$$\begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1N} & 1 \\ \vdots & & \vdots & \vdots \\ \gamma_{N1} & \cdots & \gamma_{NN} & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_N \\ \mu \end{bmatrix} = \begin{bmatrix} \bar{\gamma}(s_1, v) \\ \vdots \\ \bar{\gamma}(s_N, v) \\ 1 \end{bmatrix}$$

et la variance de krigeage a pour expression :

$$\sigma_{Kv}{}^2 = \sum_i \lambda_i \bar{\gamma}(s_i, v) - \bar{\gamma}(v, v) + \mu$$

$\bar{\gamma}(s_i, v)$: moyenne du variogramme entre s_i et le volume v.

$\bar{\gamma}(v, v)$: variogramme moyen sur le volume v.

Propriétés du krigeage :

Le krigeage possède deux propriétés importantes :

- C'est un **interpolateur exact**, c'est-à-dire que les valeurs estimées aux points de mesure sont égales aux valeurs expérimentales.
- La valeur moyenne estimée dans un bloc est égale à la moyenne des estimations à l'intérieur de ce bloc.

D'autre part, il effectue un **lissage**. Cela signifie que les estimations sont moins variables que les concentrations que l'on cherche à estimer (leur distribution est plus resserrée autour de la moyenne).

2.3 KRIGEAGE INTRINSEQUE GENERALISE

Les **fonctions aléatoires intrinsèques d'ordre k** offrent un cadre de travail pour aborder les problèmes non stationnaires. Elles permettent en effet de recourir à une gamme plus large de modèles structuraux, les modèles intrinsèques généralisés, et de se ramener ainsi à des propriétés de stationnarité. Leur usage est cependant limité par la difficulté d'interprétation de ces modèles.

Avant d'aborder le krigeage intrinsèque, quelques définitions s'imposent :

- une *combinaison linéaire autorisée d'ordre k* (notée CLA-k) est une combinaison linéaire λ, c'est-à-dire un système de poids affectés à des points, qui vérifie :
  $\forall\, l \le k, \quad \sum_i \lambda_i f_i^l = 0$
  où les fonctions f^l sont des monômes et f^l_i les valeurs de ces monômes aux points de donnée.

- Une variable régionalisée Z est une *représentation d'une fonction aléatoire intrinsèque d'ordre k* (FAI-k) $\tilde{Z}$ si et seulement si pour toute CLA-k λ, $\tilde{Z}(\lambda) = \sum_i \lambda_i Z(x_i)$
- K(h) est une covariance généralisée pour la FAI-k si et seulement si pour toute CLA-k λ,

$$\begin{cases} Var[\tilde{Z}(\lambda)] = \sum \lambda_i K_{ij} \lambda_j \\ K(h) = K(-h) \end{cases}$$

Les FAI-k sont des objets peu aisés à appréhender mais elles ouvrent la voie à des outils de modélisation utiles en cas de non stationnarité. Les logiciels spécialisés tel ISATIS facilitent cependant leur mise en œuvre.

Soit donc Z une représentation d'une FAI-k.
La technique d'estimation employée dans ce contexte est le **krigeage intrinsèque**. Comme dans le krigeage ordinaire, l'estimateur de la concentration en un point x s'écrit :

$$Z^* = \sum_i \lambda_i Z_i$$

Les poids de krigeage λ_i vérifient :

- la condition d'autorisation :
  L'erreur doit être une combinaison linéaire autorisée : $\forall l \le k \quad \sum_i \lambda_i f_i^l - f_x^i = 0$
- la condition d'optimalité :
  La variance d'estimation doit être minimale.
  $Var(\sum_i \lambda_i Z_i - Z(x)) = \sum_{i,j} \lambda_i K_{ij} + K_{xx} - 2\lambda_i K_{ix}$

Les poids λ_i qui remplissent ces conditions sont solutions d'un système matriciel faisant intervenir le modèle de covariance généralisée :

$$\begin{bmatrix} K_{11} \ldots\ldots K_{1N} & f_1^l \\ \vdots \quad\quad \vdots & \vdots \\ K_{N1} \ldots\ldots K_{NN} & f_n^l \\ f_1^l \ldots\ldots\ldots f_n & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_N \\ \mu_l \end{bmatrix} = \begin{bmatrix} K_{1x} \\ \vdots \\ K_{Nx} \\ f_x^l \end{bmatrix}$$

$K_{ij,}$ resp. K_{ix}, est la valeur de la covariance généralisée entre les points s_i et s_j, resp. entre s_i et x.
$\bar{\gamma}(s_i, V)$ est la moyenne du variogramme entre s_i et le volume V.

La variance de krigeage vaut :
$\sigma_K^2 = K_{xx} - \sum_i \lambda_i K_{ix} - \mu_l f_x^l$

2. ESTIMATION MULTIVARIABLE[1]

2.4 INTRODUCTION

Cette deuxième partie décrit la façon d'estimer simultanément plusieurs variables en tenant compte des liens structuraux qui existent entre ces dernières (ex : concentrations de l'été et de l'hiver ou concentrations de NO_2 et d'O_3) ou d'introduire une information complémentaire grâce à des variables *auxiliaires*.

- Pour la première application, on peut recourir à la technique multivariable du ***cokrigeage ordinaire,*** ce qui nécessite une modélisation multivariée.
- Pour la seconde application, on peut également faire appel au **cokrigeage** ou exploiter la méthode non stationnaire du **krigeage avec dérive externe**.

Les propriétés de stationnarité de la variable régionalisée étudiée, les corrélations entre variables et la façon dont se présentent les variables auxiliaires (sur une grille fine ou sur un ensemble plus restreint de points) guideront le modélisateur dans le choix de l'une de ces méthodes.

Ainsi, suivant la disponibilité des données de variables auxiliaires, on distingue les situations suivantes :

a) Ces données sont disponibles aux mêmes points de mesure que le polluant étudié (cas *homotopique*).

b) Ces données sont disponibles en un plus grand nombre de points, sans être toutefois connues de manière très dense : sortie d'un modèle de dispersion selon un maillage relativement lâche (AIR Pays de la Loire, *in* Table Ronde des Utilisateurs d'Isatis, 2002), données d'autres polluants issues d'un plus grand nombre de stations…On parle d'*hétérotopie* –variable principale et variables auxiliaires sont connues sur des ensembles disjoints de points- ou *d'hétérotopique partielle* – certains points sont communs aux différentes variables-,

c) Les données auxiliaires sont disponibles en tout point d'une grille relativement fine à l'échelle du domaine (cas *hétérotopique dense*).

Dans les situations a) et b), la prise en compte des données auxiliaires peut se faire par un ***cokrigeage ordinaire***.
Dans la situation c), deux techniques permettent l'introduction de données hétérotopiques denses dans le krigeage (Wackernagel, 2001) :

- L'une, celle du cokrigeage, prend en considération la structure de covariance croisée. Toutefois les difficultés numériques induites par la grande densité de données secondaires imposent de réduire la taille de voisinage. On effectue alors un ***cokrigeage colocalisé*** **ou** ***multicolocalisé*** (Wackernagel et al., 2002, Rivoirard, 2001).
- L'autre, celle du ***krigeage avec dérive externe*** **(KDE)**, consiste à se placer dans le cadre de la géostatistique non stationnaire et à considérer la variable auxiliaire comme une dérive externe.

Les paragraphes suivants décrivent plus en détail les différentes méthodes évoquées ci-dessus.

[1] Dans les méthodes d'estimation multivariable nous englobons les méthodes multivariables au sens strict, *i.e.* les méthodes de cokrigeage, et les méthodes de krigeage avec dérive externe

1.1.1 Le cokrigeage

a) Cokrigeage ordinaire

Le cokrigeage ordinaire repose sur l'hypothèse de stationnarité d'ordre 2 ou de stationnarité intrinsèque **conjointe**.

Dans ce contexte, soient Z^1, Z^2....Z^N, N variables dont on souhaite estimer les valeurs dans le domaine d'étude (par exemple Z^1 est la concentration d'O_3 et Z^2, la concentration de NO_2). Si Z^1 est l'unique variable à estimer, alors Z^2,.... Z^N représentent des variables auxiliaires corrélées à Z^1.

En un point x, l'estimateur par cokrigeage ordinaire de la variable Z^{k0} s'écrit :

$$Z^{k_0 *}(x) = \sum_{k=1}^{N} \sum_{i=1}^{n_k} \lambda_i^k Z_i^k$$

où n_k désigne le nombre de données de la variable Z^k.

Comme pour le krigeage ordinaire, les pondérateurs doivent satisfaire à la condition de non biais et rendre la variance d'estimation minimale. Ces exigences imposent de résoudre le système de cokrigeage suivant (exemple de deux variables):

$$\begin{bmatrix} \gamma_{ij}^{11} & \gamma_{ij}^{12} & 1 & 0 \\ \gamma_{ij}^{12} & \gamma_{ij}^{22} & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_i^1 \\ \lambda_i^2 \\ \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \gamma_{ix}^{11} \\ \gamma_{ix}^{22} \\ 1 \\ 0 \end{bmatrix}$$

$\gamma kk_{ij\_}$: variogramme simple de la variable Z^k entre les points de données s_i et s_j (k=1 ou 2).
$\gamma 12_{ij\_}$: variogramme croisé des variables Z^1 et Z^2 entre les points de données s_i et s_j.

La variance de cokrigeage vaut (en supposant que Z^{k0}= Z^1 est la variable à estimer):

$$\sigma_K^{\ 2} = \sum_{k=1}^{2} \sum_{i=1}^{n_k} \lambda_i^k \gamma_{ix}^{k1} + \mu_1 - \gamma_{xx}^{11}$$

Remarques :
- La variance de cokrigeage est toujours inférieure à la variance de krigeage.
- Le gain d'un cokrigeage homotopique (cas où les données de concentration et de variables auxiliaires sont connues exactement aux mêmes points) est généralement limité.
- Le cokrigeage assure la cohérence des estimations entre les différentes variables.
- En cas de corrélation intrinsèque (i.e. quand les variogrammes simples et croisés sont proportionnels à un même variogramme), le cokrigeage est identique au krigeage.

b) Cas particulier du cokrigeage colocalisé

Le cokrigeage colocalisé s'applique à l'étude d'une variable de concentration Z, dans le cas où les variables auxiliaires sont connues sur une grille dense. Pour éviter les problèmes liés à l'inversion d'une matrice de grande dimension, une simplification du voisinage de cokrigeage est nécessaire.

Dans un cokrigeage colocalisé au sens strict, l'estimation en un point tient compte de la valeur de la variable auxiliaire en ce seul point. Si la variable de concentration se décompose en une fonction linéaire de la variable auxiliaire et en un résidu sans corrélation spatiale, le cokrigeage colocalisé est strictement équivalent au cokrigeage ordinaire, sinon il n'en constitue qu'une approximation.

Dans un cokrigeage multicolocalisé, l'estimation en un point x tient compte des données de la variable auxiliaire en ce point et aux points expérimentaux (pour un exemple d'application, voir Jeannée et al., 2003) :

$$Z^*(x) = \lambda_x Z^1(x) + \sum_i [\lambda_i Z_i + \lambda_i^1 Z_i^1]$$

Z^1 : variable auxiliaire

L'utilisation la plus rigoureuse du cokrigeage multicolocalisé impose que les covariances simple et croisée respectent l'hypothèse d'un modèle à résidu (le variogramme croisé des deux variables doit être proportionnel au variogramme simple de la variable auxiliaire). Si tel n'est pas le cas, le cokrigeage multicolocalisé est là encore une approximation du cokrigeage ordinaire. Lorsqu'il existe un lien structural entre les variables, il n'en constitue pas moins un moyen efficace d'introduire des informations auxiliaires denses.

1.1.2 Krigeage avec dérive externe

Cette technique d'estimation conduit à se placer dans le cadre non stationnaire. Elle repose sur la connaissance d'une variable auxiliaire supposée mesurer indirectement le même phénomène que celui qui est étudié mais connue en tout point d'une grille couvrant le domaine d'étude.

Soit Φ(s) une telle variable.

Considérer Φ comme une dérive revient à supposer :
- qu'en moyenne, Z est proportionnelle à Φ, à une constante additive près

$$E[Z(s)] = a\Phi(s)+b \text{ soit } Z(s)= a\Phi(s)+b + R(s)$$

- et que la différence R(s)=[Z(s)- aΦ(s)-b] est une fonction aléatoire.

Comme dans le krigeage ordinaire, l'estimation de Z en un point x s'obtient par une combinaison linéaire des observations :

$$Z^*(x) = \sum_{i=1}^{n} \lambda_i Z_i$$

Les poids de krigeage doivent respecter les contraintes suivantes :

- condition d'autorisation et de non biais : $\sum_{i=1}^{n} \lambda_i = 1$
- hypothèse de dérive externe : $\sum_{i=1}^{n} \lambda_i \Phi(s_i) = \Phi(x)$
- minimisation de la variance d'estimation : $\sum_{1}^{N} \lambda_j C_{ij} - \mu_0 - \mu_1 \Phi(s_i) = C_{ix} \quad \forall i = 1...n$

Le système de krigeage associé s'écrit alors :

$$\begin{bmatrix} K_{11}.....K_{1N} & 1 & \Phi(s_1) \\ \vdots \quad\quad \vdots & \vdots & \vdots \\ K_{N1}.....K_{NN} & 1 & \Phi(s_N) \\ 1.....1 & 0 & 0 \\ \Phi(s_1).....\Phi(s_N) & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_N \\ \mu_0 \\ \mu_1 \end{bmatrix} = \begin{bmatrix} C_{1x} \\ \vdots \\ C_{Nx} \\ 1 \\ \Phi(x) \end{bmatrix}$$

Kij est la covariance associée à la fonction aléatoire Z(s)- aΦ(s)-b .

Ce système peut se généraliser à un ensemble de variables auxiliaires (Φ_1…. Φ_N) corrélées avec la variable de pollution et supposées décrire la tendance du polluant :

$E(Z(s)) = \sum_i a_i \Phi_i(s)$

(Remarque : Isatis permet d'introduire au maximum trois variables en dérive. Si plus de trois variables expliquent la concentration du polluant, une combinaison de ces dernières peut être recherchée.)

Par ailleurs, le cadre stationnaire ne suffit généralement pas à décrire les propriétés de la fonction aléatoire. Aussi est-on amené à se placer dans le contexte de non stationnarité et à définir pour cette fonction aléatoire une covariance intrinsèque généralisée.
Soit f_l (l=1…L) l'ensemble des fonctions de base définissant cette covariance généralisée. Le système d'équations à résoudre est alors :

- condition d'autorisation et de non biais : $\sum_{i=1}^{n} \lambda_i f_l(s_i) = f_l(x) \qquad l = 0...L$
- hypothèse de dérive externe : $\sum_{i=1}^{N} \lambda_i \Phi_k(s_i) = \Phi_k(x) \qquad k = 1...N$
- minimisation de la variance d'estimation : $\sum_{j=1}^{n} \lambda_j K_{ij} - \sum_{0}^{L} \mu_l' f_l(s_i) - \sum_{k=1}^{N} \mu_k \Phi^k(s_i) = K_{ix} \quad \forall i = 1...n$

Remarques :
- Le krigeage avec dérive externe filtre les constantes a_i qui n'apparaissent pas dans le système d'équations.
- Comparé à une regression linéaire multiple suivie d'un krigeage des résidus, il a la propriété d'optimiser localement, à l'intérieur du voisinage de krigeage, la combinaison linéaire des variables explicatives.
- Dans le cas où variable d'intérêt et variable(s) auxiliaire(s) sont peu corrélées entre elles, l'estimation par krigeage avec dérive externe ressemble à la dérive, tandis que l'estimation par cokrigeage s'approche du krigeage des observations.