REVUE DE STATISTIQUE APPLIQUÉE

Y. ELKETTANI

Deux méthodes complémentaires au krigeage ordinaire, pour l'estimation d'un processus spatial stationnaire de moyenne inconnue : la régression PLS et la pondération par la covariance

Revue de statistique appliquée, tome 53, n° 4 (2005), p. 31-47

<http://www.numdam.org/item?id=RSA_2005__53_4_31_0>

Rev. Statistique Appliquée, 2005, **LIII** (4), 31-47

DEUX MÉTHODES COMPLÉMENTAIRES AU KRIGEAGE ORDINAIRE, POUR L'ESTIMATION D'UN PROCESSUS SPATIAL STATIONNAIRE DE MOYENNE INCONNUE : LA RÉGRESSION PLS ET LA PONDÉRATION PAR LA COVARIANCE

Y. ELKETTANI

Faculté des Sciences, Département de mathématiques
Laboratoire d'analyse convexe et variationnelle systèmes dynamiques
et processus stochastiques
BP133, Kénitra, Maroc
Email : youssfielkettani@univ-ibntofail.ac.ma

RÉSUMÉ

L'application de la régression PLS à un champ spatial stationnaire de moyenne connue conduit à des coefficients de pondération positifs et facilement interprétables. Nous proposons une généralisation au cas le plus fréquent où l'espérance du champ spatial est inconnue. La problématique du biais conditionnel, principale critique faite au krigeage dans ce cas, est posée et un estimateur sans biais conditionnel est proposé. Par ailleurs la maximisation de la covariance entre le point à estimer et une combinaison linéaire des observations a également conduit à un autre estimateur qui est basé sur la pondération par la covariance. Les résultats obtenus par les deux nouveaux estimateurs sont comparés à ceux des estimateurs classiques.

Mots-clés : *krigeage ordinaire, biais conditionnel, régression PLS, inverse de la distance*

ABSTRACT

The PLS regression applied to predict spatial stationary random variables with known mean have given positive and meaningful coefficients. In this paper we study the most useful case of unknown mean. The best linear unbiased estimation in that case is the ordinary kriging, but it has a weak point : his conditional bias is not zero. We generalize the use of the PLS regression to the case of unknown mean, leading to a conditional unbiased estimate. The maximization of the covariance between the point to estimate and a linear combination of the observations leads to an other predictor, the covariance weights. Results obtained by these two new predictors are compared to the useful estimators.

Keywords : *PLS regression, ordinary kriging, conditional unbiasedness, inverse distance estimation*

1. Introduction

Parmi les méthodes linéaires d'estimation spatiale on peut citer la méthode des polygones qui consiste à effectuer un découpage de l'espace au voisinage d'un point M_0 en tranches de formes polygonales, contenant chacune une observation, et à affecter à chaque polygone un poids proportionnel à sa dimension. On peut aussi citer la méthode de l'inverse de la distance qui affecte à chaque observation un poids inversement proportionnel à la distance entre l'observation et le point à estimer. Il y'a également le krigeage qui est l'estimateur linéaire sans biais d'erreur quadratique minimale. Sous l'hypothèse de stationnarité on distingue le krigeage simple dans le cas où la moyenne m est connue, du krigeage ordinaire dans le cas où m est inconnue. On peut se référer à l'ouvrage classique de Cressie (1991) ou encore à celui de Chilès et Delfiner (1999), très complet sur la théorie de l'estimation spatiale. La régression PLS (partial least squares) introduite par Wold, Albano et al (1983) et dont les propriétés mathématiques sont développées dans Tenenhaus (1998), est bien adaptée au traitement des données présentant une multicolinéarité. Son application à des données spatiales et sa comparaison avec le krigeage simple et avec l'estimation par inverse de la distance sont présentées dans Elkettani (2001).

Cet article est consacré au cas où la moyenne m est inconnue. L'estimateur linéaire, sans biais, et d'erreur quadratique minimale est dans ce cas le krigeage ordinaire. Toutefois, ce dernier présente un biais conditionnel non nul. Par conséquent, comme le soulignent Rivoirard (1984), Ajerame (1997), ou encore Yates S.R. et Yates M.V. (1998), le krigeage ordinaire n'est pas indiqué dans des problèmes de sélection où il faut décider, sur la base des observations, si la variable aléatoire dépasse ou non un seuil ou valeur de référence. Après un rappel du krigeage ordinaire, nous présentons deux nouveaux algorithmes d'estimation linéaire basés sur la maximisation de la covariance entre une combinaison linéaire des observations et le point à estimer. Le premier estimateur est obtenu par application, au cas où m est inconnue, de la régression PLS légèrement modifiée afin d'obtenir la propriété de non biais conditionnel dans le cas d'un processus spatial gaussien. Et le second est basé sur la pondération par la fonction de covariance. Puis nous comparerons à travers des simulations et sur des données réelles, les résultats obtenus par les nouveaux estimateurs à ceux des estimateurs spatiaux classiques : krigeage simple, krigeage ordinaire et inverse de la distance, ainsi qu'à l'estimateur PLS quand m est connue.

Estimation spatiale linéaire

Soit un domaine D de R^p muni d'un espace de probabilité et sur lequel est étudiée une fonction aléatoire $\Phi : D\text{x}\Omega \rightarrow R$. Pour $M \in D$, on définit la variable aléatoire $f(M) : \omega \longmapsto \Phi(M, \omega), \omega \in \Omega$. Par ailleurs, pour une réalisation ω_0 donnée, supposons que soit observée la fonction aléatoire Φ en n points $M_1, ..., M_n$ de D; l'ensemble $\{\Phi(M_i, \omega_0), i = 1, n\}$ est considéré comme la réalisation unique de la fonction aléatoire Φ. Notons par F le processus spatial $F = \{f(M), M \in D\}$

Notre objet est de prédire la variable aléatoire $F_0 = f(M_0)$ en un point M_0 non observé de D à partir des variables aléatoires observées $f_i = f(M_i)$, $i = 1, n$; $M_i \in D$ dans un voisinage de M_0. Notons par f le vecteur colonne $(f_1...f_n)'$ où $'$ désigne l'opérateur de transposition matricielle. Le processus F est, dans tout ce qui suit, considéré stationnaire, ergodique, d'espérance $m \in R$ et de

fonction de covariance $c(h)$ prédéterminée et présentant un palier $c(0)$ égal à la variance ponctuelle $\sigma^2 = var(F(M))$. L'expression de l'estimateur de F_0, étant une fonctionnelle des vecteurs distance h_i entre M et les observations M_i, elle permet d'obtenir une estimation pour tout point non observé dans le voisinage. Notons par $C = (c_1, ..., c_n)^{'}$ le vecteur colonne où $c_i = cov(F_0, f_i)$, et par K la matrice $n \times n$ dont l'élément (i, j) est $k_{ij} = cov(f_i, f_j)$.

Le biais conditionnel

La principale critique faite au krigeage ordinaire est qu'il présente un biais conditionnel, ce qui le rend non indiqué dans des problèmes de prédiction des variables aléatoires dépassant un seuil critique. Supposons qu'on veuille décider, sur la base de l'estimation f^*, si la variable non observée F_0 dépasse ou non un seuil critique f^c. Cette situation se produit par exemple dans le domaine minier où il est économiquement intéressant de n'exploiter que les blocs dont la teneur en minerai dépasse un seuil de coupure f^c. On retrouve également cette même siltuation dans la pêche, l'agriculture ou l'environnement.

Dans une telle situation il est très appréciable d'avoir l'égalité entre l'espérance conditionnelle $E(F_0 \mid f^* \geqslant f^c)$ qui représente la moyenne espérée après sélection, et l'expression conditionnelle connue qui est $E(f^* \mid f^* \geqslant f^c)$. On peut voir par exemple dans Chilès et Delfiner (chapitre 6), que cette égalité particulièrement recherchée est vérifiée pour un estimateur f^*, dit sans biais conditionnel, c'est-à-dire qui vérifie :

$$E(F_0 \mid f^*) = f^*$$

En effet : Soit $\mathbf{1}_A$ la fonction indicatrice d'un ensemble A, on a

$$E(F_0 \mathbf{1}_{\{f^* \geqslant f^c\}}) = E\{E(F_0 \mathbf{1}_{\{f^* \geqslant f^c\}}) \mid f^\star\}$$

qui s'écrit aussi car $\mathbf{1}_{\{f^* \geqslant f^c\}}$ est f^*-mesurable : $E\{\mathbf{1}_{\{f^* \geqslant f^c\}} E(F_0 \mid f^\star)\}$. Considérons la fonction $h(f^\star) = E(F_0 \mid f^\star)$, que nous supposons croissante (ce qui signifie que la valeur moyenne croît avec la valeur estimée); alors on a l'égalité

$$E(F_0 \mathbf{1}_{\{f^* \geqslant f^c\}}) = E(h(f^\star) \mathbf{1}_{\{f^\star \geqslant f^c\}}) = E\{h(f^\star) \mathbf{1}_{\{h(f^\star) \geqslant h(f^c)\}}\}.$$

Enfin pour un estimateur sans biais conditionnel h est égale à l'identité, et on a $E(F_0 \mathbf{1}_{\{f^* \geqslant f^c\}}) = E(f^* \mathbf{1}_{\{f^* \geqslant f^c\}})$. D'où l'égalité recherchée :

$$E(F_0 \mid f^* > f^c) = E(f^* \mid f^* \geqslant f^c)$$

Par ailleurs la variance d'estimation $\phi^2(f^*)$ d'un estimateur sans biais $f^\star$ se décompose en deux termes :

$$\phi^2(f^*) = E\left[(F_0 - f^*)^2\right] = E\left[(E(F_0 \mid f^*) - f^*)^2\right] + E\left[(F_0 - E(F_0 \mid f^*))^2\right]$$

Et pour un estimateur sans biais conditionnel, le 1^{er} terme est nul et la variance d'estimation se ramène à la distance dans $L^2(\Omega)$ entre la variable à estimer F_0 et le sous espace engendré par les estimations $vect\{f_1^*, ..., f_n^*\}$.

Dans le cas où f^* est sans biais conditionnel, la régression de F_0 en f^* est lineaire avec une pente :

$$p = cov\left(F_0, f^*\right) / var\left(f^*\right) \tag{1}$$

égale à 1. Cette relation qui est nécessaire pour que f^* soit sans biais conditionnel, est suffisante si F est un processus spatial gaussien, c'est-à-dire, si tout k-uplet $(f_{k(1)}, ..., f_{k(k)})$ suit une distribution multigaussienne de dimension k (voir par exemple : Michael L. Stein, chapitre 3). La relation (1) s'écrit aussi

$$var(f^\star) = cov(F_0, f^\star) \tag{2}$$

Et la variance d'estimation d'un estimateur sans biais conditionnel s'écrit :

$$\phi^2(f^*) = \sigma^2 - var(f^*). \tag{3}$$

Cette dernière relation est vérifiée pour tout prédicteur obtenu par projection orthogonale sur le sous espace vectoriel engendré par les estimations, comme par exemple le krigeage simple ou encore l'estimation spatiale par régression PLS. Mais elle n'est pas vérifiée pour le krigeage ordinaire.

Avant d'introduire les nouveaux estimateurs basés sur la maximisation de la covariance entre F_0 et une combinaison linéaire des observations $\lambda' f$, nous commençons par rappeler les résultats obtenus, par la même approche, dans le cas où la moyenne m est connue.

2. Cas où la moyenne m est connue (Elkettani 2001)

Considérons les variables $z_i = \frac{1}{\sigma}(f_i - m), i = 1, ..., n$ et $Z_0 = \frac{1}{\sigma}(F_0 - m)$. Et soit $z_0 = \lambda^{'} z$ un estimateur linéaire de Z_0, où $\lambda = (\lambda_i)_{i=1,n} \in R^n$ est le vecteur des coefficients d'estimation, et $z = (z)_{i=1,...,n}$ le vecteur des z_i.

2.1. L'estimation par krigeage simple

La méthode du krigeage stationnaire à moyenne connue, dit krigeage simple (ks), consiste à chercher le vecteur des coefficients $\lambda = (\lambda_i)_{i=1,n}$, tel que la combinaison linéaire $\lambda^{'} f$ minimise l'erreur quadratique $E = E\left(Z_0 - \lambda^{'} z\right)^2$. La résolution de cette équation donne $\lambda = K^{-1}C$ que nous noterons par la suite λ'_{ks}. L'estimateur de Z_0 est $z_{ks} = z'K^{-1}C$ et F_0 est estimé sans biais par $f_{ks} = \sigma z_{ks} + m$ dont l'erreur quadratique est :

$$\phi^2(f_{ks}) = E(F_0 - f_{ks})^2 = \sigma^2 - 2\lambda'_{ks}C + \lambda'_{ks}K\lambda_{ks} = \sigma^2 - C'K^{-1}C$$

Or la variance du krigeage simple est $var(f_{ks}) = \lambda'_{ks}K\lambda_{ks} = C'K^{-1}C$ on peut alors écrire, comme dans (3) :

$$\phi^2(f_{ks}) = \sigma^2 - var(f_{ks})$$

2.2. Estimation spatiale basée sur la régression PLS

Il s'agit de chercher la combinaison linéaire $t = \lambda' z$, sous la contrainte $\|\lambda\| = 1$ telle que $cov\,(Z_0, t)$ soit maximale. Ceci revient à rechercher la première composante de la régression PLS de Z_0 sur

$$z = \{z_i = z\,(M_i)\,, i = 1...n, M_i \in D\}$$

Nous n'avons qu'une seule composante PLS t, car le tableau des régresseurs z est de dimension $(1 \times n)$ que nous déterminons en utilisant la structure de covariance existant entre les données. Soit μ un multiplicateur de Lagrange, en maximisant l'expression $E = \frac{1}{\sigma^2}\lambda' C - \mu\left(\lambda^{'}\lambda - 1\right)$, on obtient $\lambda = \frac{1}{(C'C)^{\frac{1}{2}}}C$, puis la composante PLS $t = \lambda^{'} z$. Enfin on régresse Z_o sur t pour obtenir l'estimateur PLS z_{pls} donné par : $z_{pls} = \frac{cov\,(z_o, t)}{var\,(t)}t = \frac{C'C}{C'KC}C'z = \lambda'_{pls}z$ en notant λ_{pls} le vecteur des coefficients :

$$\lambda_{pls} = \frac{C'C}{C'KC}C.$$

Enfin Z_0 est estimé par $z_{pls} = \lambda'_{pls}z$ et F_0 est estimé sans biais par $f_{pls} = \sigma z_{pls} + m$. L'erreur quadratique est :

$$\phi^2(f_{pls}) = E(F_0 - f_{pls})^2 = \sigma^2 - \frac{C'CC'C}{C'KC}$$

Or la variance de l'estimateur est $var(f_{pls}) = \lambda'_{pls}K\lambda_{pls} = \frac{C'C}{C'KC}C$ on peut alors écrire comme dans (3) :

$$\phi^2(f_{pls}) = \sigma^2 - var(f_{pls})$$

Les coefficients λ_{pls} sont facilement interprétables et sont positifs quand la fonction de covariance l'est, ce qui est le cas de presque toutes les fonctions de covariance usuelles à l'exception de la fonction sinus cardinal.

3. Cas où la moyenne est inconnue

Dans cette situation l'estimateur fréquemment utilisé est le krigeage ordinaire (ko), mais on trouve aussi l'estimation par inverse de la distance. Nous allons tout d'abord rappeler les équations du krigeage ordinaire (paragraphes 3.1 et 3.2), avant de développer l'estimation par régression PLS dans le cas où m est inconnue (paragraphe 3.3). Par ailleurs, après un bref rappel de l'estimateur par inverse de la distance, nous présenterons une nouvelle méthode d'estimation (paragraphe 3.4) : la pondération par la covariance, qui est aussi basée sur la maximisation de la covariance entre une combinaison linéaire des observations et le point à estimer.

Rappelons tout d'abord que dans le cas m inconnue, pour qu'un estimateur spatial linéaire $f_0 = \lambda' f$ soit sans biais, il doit vérifier la condition $\lambda' 1_n = 1$ où 1_n est le vecteur colonne de R^n dont toutes les composantes sont égales à un. En effet :

$$E(f_0) = \lambda' E f = m\lambda' 1_n$$

3.1. Krigeage de la moyenne

On estime m par une combinaison linéaire des observations $m^* = \lambda'_m f$ qui minimise l'erreur quadratique $E\left(m - m^*\right)^2$ sous la contrainte de non biais $\lambda'_m 1_n = 1$. Soit μ_0 un multiplicateur de Lagrange et 0_n le vecteur nul de R^n, alors λ_m et μ_0 sont solution de l'équation

$$\begin{pmatrix} K & 1_n \\ 1'_n & 0 \end{pmatrix} \begin{pmatrix} \lambda_m \\ \mu_0 \end{pmatrix} = \begin{pmatrix} 0_n \\ 1 \end{pmatrix}$$

La variance de $m^\star$ est

$$var(m^\star) = \lambda'_m K \lambda_m$$

et le vecteur λ_m vérifie

$$K\lambda_m - \lambda'_m K \lambda_m 1_n = 0_n \tag{4}$$

L'expression des coefficients λ_m est donnée par :

$$\lambda_m = \frac{K^{-1} 1_n}{1'_n K^{-1} 1_n}$$

Et la variance d'estimation du krigeage de la moyenne est $var(m^*) = \lambda'_m K \lambda_m = \frac{1}{1'_n K^{-1} 1_n}$

3.2. Le krigeage ordinaire

La méthode du krigeage stationnaire à moyenne inconnue, dit krigeage ordinaire (ko), consiste à chercher le vecteur des coefficients λ_{ko}, tel que la combinaison linéaire $\lambda'_{ko} f$ minimise l'erreur quadratique $E\left(F_0 - \lambda'_{ko} f\right)^2$ sous la contrainte de non biais $\lambda'_{ko} 1_n = 1$.

Soit μ un multiplicateur de lagrange et 0_n le vecteur nul de R^n, alors λ_{ko} et μ sont solution du système d'équations :

$$\begin{pmatrix} K & 1_n \\ 1'_n & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \begin{pmatrix} C \\ 1 \end{pmatrix}$$

La résolution de ces équations conduit à l'expression du vecteur des coefficients $\lambda_{ko} = \lambda_m + \lambda_{ks} - \lambda'_{ks} 1_n \lambda_m$ où λ_{ks} et λ_m désignent les coefficients du krigeage simple et du krigeage de la moyenne, respectivement. Ce qui s'écrit encore :

$$\lambda_{ko} = \lambda_{ks} + w_m \lambda_m$$

où $w_m = \left(1 - C' K^{-1} 1_n\right)$ est appelé poids de la moyenne. Et l'estimateur obtenu est $f_{ko} = m^* + f_{ks} - \lambda'_{ks} 1_n m^*$ qui s'écrit aussi :

$$f_{ko} = m^* + \lambda'_{ks}(f - m^* 1_n)$$

Cette dernière expression montre que le krigeage ordinaire consiste à faire un krigeage simple avec les variables centrées $(f - m^* 1_n)$ puis de rajouter la moyenne estimée m^*. Nous reprenons cette idée dans la construction de l'estimateur par régression PLS dans le cas m inconnue. Enfin le minimum atteint par l'erreur quadratique est

$$\phi^2(f_{k0}) = E(f_{ko} - F_0)^2 = \lambda'_{ko} K \lambda_{ko} + \sigma^2 - 2\lambda'_{ko} C$$

Et on a la relation avec la variance de l'estimateur de la moyenne et l'erreur quadratique du krigeage simple (voir par exemple Rivoirard 1984) :

$$\phi^2(f_{ko}) = \phi^2(f_{ks}) + (w_m)^2 var(m^*)$$

où w_m est le poids de la moyenne. Remarquons que dans ce cas on n'a pas d'expression du type (3) pour le krigeage ordinaire dont le biais conditionnel est non nul.

3.3. Régression PLS dans le cas où la moyenne est inconnue

Nous allons étendre l'estimation spatiale par régression PLS au cas où la moyenne $m = E(F)$ du champ est supposée inconnue. Nous commençons par estimer m par $m^\star$ obtenu par krigeage de la moyenne, avant de procéder à la régression PLS sur la fonction aléatoire centrée.

Considérons la variable $Z_0 = \frac{1}{\sigma}(F_0 - m^\star)$, on cherche le vecteur λ de R^n tel que la variable aléatoire $t = \frac{1}{\sigma}\lambda'(f - m^* 1_n)$ maximise $cov(t, Z_0)$ sous la contraine de normalisation $\lambda'\lambda = 1$. Soit μ un multiplicateur de Lagrange, on maximise l'expression :

$$\begin{aligned} E &= \frac{1}{\sigma^2} cov\left(\lambda'(f - (\lambda'_m f)1_n), F_0 - \lambda'_m f\right) + \mu(\lambda'\lambda - 1) \\ &= \frac{1}{\sigma^2}\left(\lambda' C - \lambda'_m C \lambda' 1_n - \lambda' K \lambda_m + \lambda'_m K \lambda_m \lambda' 1_n\right) + \mu(\lambda'\lambda - 1) \end{aligned}$$

qui se réduit, compte tenu de (4) à : $E = \frac{1}{\sigma^2}\left(\lambda' C - \lambda'_m C \lambda' 1_n\right) + \mu\left(\lambda'\lambda - 1\right)$ L'annulation de la dérivée de E par rapport à λ conduit à l'équation

$$\frac{1}{\sigma^2}\left(C - \lambda'_m C 1_n\right) + 2\mu\lambda = 0_n$$

L'annulation de la dérivée de E par rapport à μ restitue la contrainte de normalité $\lambda'\lambda = 1$. Dans le cas où C n'est pas proportionnel à 1_n (le cas C proportionnel à 1_n est traité dans la remarque 3 ci-dessous), on déduit des deux égalités précédentes, l'expression de λ

$$\lambda = (C - \lambda'_m C 1_n) \,/\, [(C' - \lambda'_m C 1'_n)\,(C - \lambda'_m C 1_n)]^{\frac{1}{2}}$$

Posons $\eta = \lambda - \lambda' 1_n \lambda_m$, un vecteur orthogonal à 1_n ($\eta' 1_n = 0$ car $\lambda'_m 1_n = 1$). Alors on a $t = \frac{1}{\sigma}\eta' f$ et $var(t) = \frac{1}{\sigma^2}\eta' K \eta$ et, d'après l'orthogonalité entre η et 1_n, et compte tenu de (4) :

$$cov(Z_0, t) = \frac{1}{\sigma^2}\eta' C$$

Puis on régresse Z_0 sur t obtenant une estimation de Z_0 donnée par $z_{plsm} = rt$ où $r = cov\,(Z_0, t)\,/var\,(t)$. Enfin l'estimateur de F_0 est $f_{plsm} = \sigma z_{plsm} + m^\star$ qui a pour expression :

$$f_{plsm} = (r(\lambda - \lambda' 1_n \lambda_m) + \lambda_m)'\, f = \lambda'_{plsm} f \tag{5}$$

où $\lambda_{plsm} = (r(\lambda - \lambda' 1_n \lambda_m) + \lambda_m)$ est le vecteur des coefficients.

Remarque 1. – Le prédicteur f_{plsm} est un estimateur sans biais. De plus, pour tout $a \neq 0$ on peut définir un estimateur sans biais de F_0 par : $f_{plsm}(a) = (ar(\lambda - \lambda' 1_n \lambda_m) + \lambda_m)'\, f = \lambda'_{plsm}(a) f$. En effet $\lambda'_m 1_n = 1$ entraîne que $\lambda'_{plsm} 1_n = 1$.

Remarque 2. – Le prédicteur $f'_{plsm}(a)$ est la somme de deux termes non corrélés entre eux :

$$f_{plsm}(a) = (ar\eta + \lambda_m)' f$$

En effet $cov(arf'\eta, f'\lambda_m) = ar\eta' K \lambda_m$ et par (4) on a $\eta' K \lambda_m = \eta' 1_n \lambda'_m K \lambda_m$; or par construction de η on a $\eta' 1_n = 0$, et par conséquent $cov(f'\eta, f'\lambda_m) = 0$. On a donc :

$$var(f_{plsm}(a)) = a^2 r^2 \eta' K \eta + \lambda'_m K \lambda_m \tag{6}$$

et

$$cov(f_{plsm}(a), F_0) = (ar\eta' + \lambda'_m) C \tag{7}$$

Enfin l'erreur quadratique pour cet estimateur est donnée par :

$$\phi^2(f_{plsm}) = \sigma^2 + a^2 r^2 \eta' K \eta + \lambda'_m K \lambda_m - 2(ar\eta + \lambda_m)' C \tag{8}$$

Remarque 3. – Si le vecteur C est proportionnel à 1_n, ce qui est le cas si toutes les observations sont sur un cercle de centre M avec une fonction de covariance

isotrope, $\lambda'_m C 1_n = C$. Dans ce cas, l'annulation de la dérivée de E par rapport à λ conduit à $\lambda = 0$ et par conséquent $f_{plsm} = m^\star$ qui est le même estimateur que celui obtenu par le krigeage ordinaire.

Un estimateur linéaire sans biais conditionnel pour processus gaussiens : À partir de (6) et de (7), la condition (2) se ramène à l'équation du 2^{nd} ordre

$$a^2 r^2 \eta' K \eta - a r \eta' C + \lambda'_m K \lambda_m - \lambda'_m C = 0. \tag{9}$$

Si le discriminant de cette équation du 2^{nd} degré est positif, alors en prenant pour valeur de a une solution a_1 ou a_2 de (9), nous obtenons un estimateur qui vérifie (2). Sa variance d'estimation se réduit d'après (8) et (9) à

$$\phi^2 = \sigma^2 - a r \eta' C - \lambda'_m C$$

qui est minimale pour $a_0 = \max(a_1, a_2)$.

On a ainsi obtenu pour $a = a_0$, un estimateur particulier qui est sans biais conditionnel dans le cas d'un processus spatial gaussien, et que nous appelons PLS ordinaire (PLSO) par analogie avec le krigeage ordinaire. Par ailleurs en utilisant (9), l'expression $a_0 r \eta' C + \lambda'_m C$ s'écrit $a_0^2 r^2 \eta' K \eta + \lambda'_m K \lambda_m$ qui est, d'après (6), la variance de l'estimateur $f_{plsm}(a_0)$ qu'on notera f_{plso}; et par conséquent la variance d'estimation de ce dernier peut s'écrire :

$$\phi^2(f_{plso}) = \sigma^2 - var(f_{plso})$$

à l'instar de l'expression (3) obtenue précédemment pour les estimateurs f_{ks} et f_{pls}.

Il faut toutefois noter que l'estimateur sans biais conditionnel f_{plso} n'est pas défini si le discriminant de l'équation (9) est strictement négatif. Dans ce cas la régression PLS aboutit à l'estimateur sans biais f_{plsm} donné par (5). Toutefois, sur les 520 estimations PLSO effectuées dans les exemples qui suivent, ce cas ne s'est produit qu'une seule fois (Voir exemples ci-dessous).

3.4. Nouvel estimateur : la pondération par la covariance

En appliquant la démarche de la maximisation de la covariance qui est à la base de diverses méthodes d'analyse des données multidimensionnelles, nous allons aboutir à un estimateur qui a, dans une certaine mesure, le même type d'expression que l'inverse de la distance.

Rappelons tout d'abord que l'estimation par inverse de la distance attribue à chaque observation f_i un coefficient inversement proportionnel à la distance d_i entre F_0 et f_i. Puis on somme à 1 le vecteur des coefficients pour assurer le non biais. L'estimateur par inverse de la distance s'écrit alors $f_{id} = \sum_{i=1,n}(\frac{1}{di} f_i)/\sum \frac{1}{di} = \lambda'_{id} f$, en notant $\lambda_{id} = \left((\frac{1}{di})/\sum_{j=1,n} \frac{1}{d_j} \right)_{i=1,n}$ le vecteur des coefficients.

Cherchons maintenant la combinaison lineaire $t = \lambda^{'} f$ qui maximise $cov\,(F_0, t)$ sous la contrainte $\|\lambda\| = 1$. Soit μ un multiplicateur de Lagrange, on maximise

$$E = cov\,(F_0, t) - \mu\,(\|t\| - 1) = C'\lambda - \mu\left(\lambda^{'}\lambda - 1\right)$$

ce qui conduit à l'équation $C - 2\mu\lambda = 0$ et à la contrainte de normalité $\lambda^{'}\lambda = 1$. Alors on déduit $\lambda = \dfrac{C}{(C'C)^{\frac{1}{2}}}$ et $t = \dfrac{C'f}{(C'C)^{\frac{1}{2}}}$

La régression de F_0 sur t nous donne un estimateur de F_0 combinaison linéaire des observations $f_{pc} = b\lambda' f$. Et en tenant compte de la contrainte de non biais $b\lambda' 1_n = 1$, on obtient l'estimateur de pondération par la covariance :

$$f_{pc} = \frac{C'f}{C'1_n} = \lambda'_{pc} f$$

où le vecteur des coefficients de régression λ_{pc} est donné par :

$$\lambda_{pc} = \frac{C}{C'1_n}$$

Enfin l'erreur quadratique est ici :

$$\phi^2(f_{pc}) = \sigma^2 + \frac{C'KC}{C'1_n 1'_n C} - 2\frac{C'C}{C'1_n}$$

Remarquons que pour des fonctions de covariance decroissantes avec la distance, la pondération par la covariance revient à pondérer de façon inversement proportionnelle à la distance liée au produit scalaire : «covariance des variables aléatoires dans $L^2(\Omega)$», alors que l'estimateur f_{id} utilise la distance euclidienne dans l'espace D. Pour une structure anisotrope, la pondération par la covariance tient compte de l'anisotropie du phénomène étudié contrairement à l'estimateur par inverse de la distance, comme nous le verrons dans l'exemple 1 ci-dessous. Par ailleurs l'expression de f_{pc} permet, à l'instar de l'estimateur PLS dans le cas m connue, d'interpréter facilement le vecteur des coefficients λ_{pc}, et ce dernier est positif chaque fois que la fonction de covariance est positive. Cette propriété est souvent recherchée par les utilisateurs de l'estimation spatiale comme on peut le constater dans Barnes et Johnson (1984) qui ont proposé d'imposer au krigeage une condition de positivité des poids, et des algorithmes dans le même sens sont également développés dans Szidarovsky *et al.* (1987), puis Herzfeld (1989).

4. Applications

Nous allons appliquer les estimateurs étudiés à deux jeux de données réelles, qui figurent parmi ceux qui illustrent le logiciel Variowin de Pannatier (1996). Les deux variables étudiées représentent le taux de plomb et le taux de cadmium dans

le sol en ppm; les mesures ont été effectuées en 60 sites dans le plan horizontal. La figure 1 présente la localisation des 60 points d'observation. Le taux de cadmium est isotrope alors que le taux de plomb est anisotrope. Enfin nous appliqueront les méthodes d'estimation étudiées à une variable spatiale simulée à l'aide du programme de simulations séquentielles disponible dans le logiciel GSLIB de Deutsch et Journel (1992).

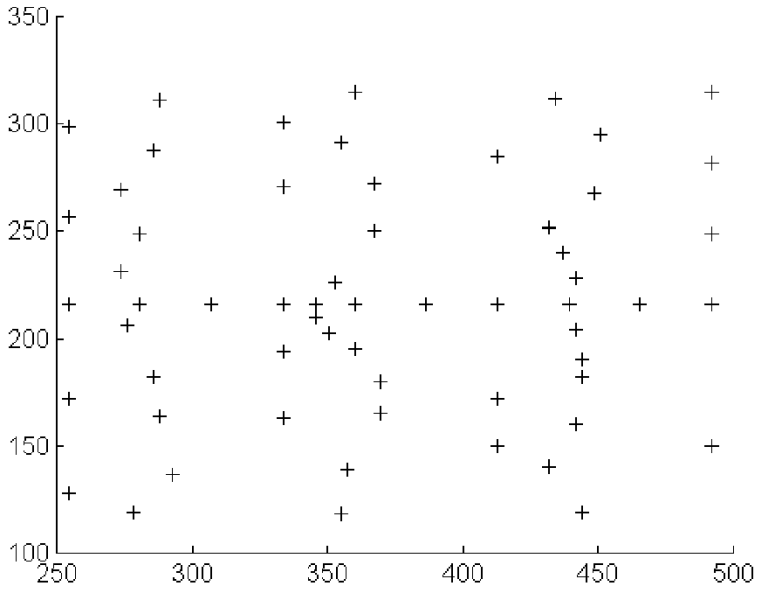

FIGURE 1
Localisation des 60 points d'observation

Pour chaque exemple nous présenterons pour les six méthodes présentées (à savoir le krigeage simpe, l'estimateur PLS pour m connue, l'inverse de la distance, la pondération par la covariance, le krigeage ordinaire et l'estimateur PLSO pour m inconnue) : une estimation ponctuelle, ainsi que les résultats moyens obtenus par validation croisée, et enfin nous apprécierons le biais conditionnel du krigeage ordinaire et de l'estimateur PLSO en évaluant empiriquement la pente p donnée par (1). Toutefois lors de la validation croisée dans l'exemple 1, l'estimation PLSO du taux de plomb a conduit, pour le point $M_0(x_0 = 492, y_0 = 150)$, à un discriminant de l'équation (9) négatif. f_{plso} a donc été remplacée pour ce point relativement éloigné des autres par l'estimation f_{plsm} donnée par (5).

4.1. Exemple 1

La variable étudiée est le taux de plomb. Il varie entre 1 et 302.5 ppm. La fonction $c(h)$ ajustée aux données est une covariance sphérique anisotrope dont l'axe de plus grande variabilité, de portée $a = 93.8$, est de direction nord-est faisant un

angle de 43.4 avec l'axe ouest-est. Avec un rapport d'anisotropie égal à 0.47, et une variance $\sigma^2 = 2893.76$ l'expression de la covariance est donnée par :

$$c(h) = \sigma^2(1 - sph\,(h))$$

avec : $sph\,(h) = \left(\frac{3}{2}(h/a) - \frac{1}{2}\,(h/a)^3\right)\chi_{\{0<h<a\}}$ où h est la distance exprimée dans le repère obtenu par la transformation qui associe à l'ellipse anisotrope, le cercle de même centre et de rayon égale à la plus grande portée (*cf.* par exemple Panatier 1996, ou Isaaks et Srivastava 1989). Nous nous proposons d'estimer la fonction aléatoire au point $M_0(x_0 = 444, y_0 = 182)$. Les valeurs obtenues par les six estimateurs étudiés sont très éloignées de la valeur observée $v_0 = 118$, mais plus proches de la valeur moyenne du processus $m = 46$. Aussi, il semble intéressant d'évaluer, par validation croisée, le biais expérimental moyen des estimateurs.

À cet effet nous avons procédé à l'estimation une par une de toutes les valeurs observées en retirant à chaque fois de l'ensemble des observations le point à estimer. Le biais moyen et l'erreur quadratique moyenne obtenus sont alors rapportés respectivement à l'écart-type σ et à la variance σ^2. Le biais standardisé moyen b et l'erreur quadratique standardisée moyenne e ainsi obtenus sont donnés par :

$$b(f^*) = \frac{1}{n}\sum_{i=1}^{n}(f_i - f^*{}_{(-i)})/\sigma \tag{10}$$

où $f^*_{(-i)}$ est l'estimateur f^* de f_i obtenu à partir de toutes les observations excepté f_i. Et

$$e(f^*) = \frac{1}{n}\sum_{i=1}^{n}(\phi^2(f^*_{(-i)}))/\sigma^2. \tag{11}$$

Les biais standardisés moyens obtenus sont tous de l'ordre de 10^{-2}. Le tableau 1 présente l'estimation ponctuelle $f^*(M_0)$ et son erreur quadratique $\phi^2(f^*)$, le biais et l'erreur quadratique standardisés moyens b et e pour les six méthodes d'estimation étudiées :

TABLEAU 1
Résultats obtenus pour le taux de plomb

	$f^*(M_0)$	$\phi^2(f^*)$	$b(f^*)$	$e(f^*)$
KS	48.6348	352.8	0.017	0.251
PLS	33.8699	881.6	0.067	0.458
Inv. dist.	45.729	1332.9	0.026	0.624
Pond. covar.	36.4651	973.7	0.041	0.483
KO	48.6345	352.8	0.018	0.252
PLSO	30.4546	908.2	0.025	0.426

Dans cet exemple anisotrope, l'erreur quadratique standardisée moyenne de la pondération par la covariance est moins élevée que celle de l'inverse de la distance. Les krigages simple et ordinaire présentent les meilleurs erreurs quadratiques moyens. Par contre l'estimation PLSO a corrigé le biais conditionnel obtenu par le ko. En effet, la pente de la régression des observations sur les estimations par krigeage ordinaire vaut $p_{ko} = 0.75$, contre $p_{plso} = 0.88$ pour l'estimateur PLSO.

4.2. Exemple 2

La variable étudiée est le taux de cadmium. Il varie entre 0 et 17.5 ppm. La fonction $c(h)$ ajustée aux données est une covariance exponentielle isotrope définie par : $c\,(h) = 13.25ex(-(h/b))$ où h est la distance euclidienne dans le plan horizontal, et $b = 142.89$ un paramètre de la portée. La variance est $\sigma^2 = 16.3$ et pour les estimateurs ks et PLS, nous considérons la valeur moyenne du processus $m = 6.65$. Nous nous proposons d'estimer la variable aléatoire au même point $M_0(x_0 = 444, y_0 = 182)$ que dans l'exemple 1. La valeur observée pour le taux de cadmium en M_0 est $v_0 = 14.5$. Le tableau 2 présente l'estimation ponctuelle $f^*(M_0)$ et son erreur quadratique $\phi^2(f^*)$, ainsi que le biais et l'erreur quadratique standardisés moyens b et e définis par 10 et 11 pour les six méthodes d'estimation étudiées :

TABLEAU 2
Résultats obtenus pour le taux de cadmium

	$f^*(M_0)$	$\phi^2(f^*)$	$b(f^*)$	$e(f^*)$
KS	10.6688	6.0577	0.028	0.424
PLS	8.3746	9.4909	0.100	0.631
Inv. dist.	9.2810	8.0992	0.095	0.591
Pond. covar.	8.1779	9.5911	0.071	0.637
KO	10.6628	6.0578	0.019	0.425
PLSO	8.3372	7.4063	0.021	0.503

Dans cet exemple isotrope, on voit que l'erreur quadratique standardisée moyenne de la pondération par la covariance, du même ordre que celle de la PLS, n'a pas été meilleure que celle de l'inverse de la distance, comme c'était le cas dans l'exemple anisotrope précédent. Les krigages simple et ordinaire présentent les meilleurs erreurs quadratiques, toutefois l'estimation PLSO a corrigé le biais conditionnel obtenu par le ko. En effet, la pente de la régression des observations sur les estimations par krigeage ordinaire vaut $p_{ko} = 0.94$, contre $p_{plso} = 0.99$ pour l'estimateur PLSO.

4.3. Simulations

La simulation d'un nombre N de réalisations de la fonction aléatoire F peut se faire sur une grille aussi fine qu'on le désire du domaine D, donnant ainsi N images possibles de l'incertain spatial. Les réalisations obtenues doivent reproduire les paramètres de la fonction aléatoire et partculièrement les moments d'ordre un et deux. Les simulations sont en plus dites conditionnelles à la variable observée si elles prennent les mêmes valeurs qu'elle aux points d'observation. Lantuejoul (1994) établit une comparaison des méthodes de simulation spatiale les plus courantes, et on peut également se référer à Chilès et Delfiner (1999, chapitre 7) pour une présentation complète du sujet. Nous utilisons ici l'algorithme de simulations séquentielles qui est toujours appliquable, et dont le principe est le suivant : Considérons une variable aléatoire vectorielle $F = (F_1, ..., F_N)'$ pour laquelle une réalisation du sous vecteur $(F_1, ..., F_n)$ est connue et est égale à $(f_1, ..., f_n), (0 \leqslant n \leqslant N)$. La distribution de $(F_1, ..., F_N)$ peut être factorisée, sous la forme du produit de la distribution $\Pr\{f_1 \leqslant F_1 < f_1 + df_1, ..., f_n \leqslant F_n < f_n + df_n\}$, et de la loi conditionnelle à $F_i = f_i, i = 1...n$, : $\Pr\{f_{n+1} \leqslant F_{n+1} < f_{n+1} + df_{n+1}, ..., f_N \leqslant F_N < f_N + df_N \mid f_1, ..., f_n\}$. Or ce dernier terme peut s'écrire, à la suite d'une succession d'applications de la formule de bayes :

$$\Pr\{f_{n+1} \leqslant F_{n+1} < f_{n+1} + df_{n+1}, ..., f_N \leqslant F_N < f_N + df_N \mid f_1, ..., f_n\} =$$

$$\Pr\{f_{n+1} \leqslant F_{n+1} < f_{n+1} + df_{n+1} \mid f_1, ..., f_n\} \times$$

$$\Pr\{f_{n+2} \leqslant F_{n+2} < f_{n+2} + df_{n+2} \mid f_1, ..., f_n, f_{n+1}\} \times ... \times$$

$$\Pr\{f_N \leqslant F_N < f_N + df_N \mid f_1, ..., f_n, f_{n+1}, ..., f_{N-1}\}$$

ce qui induit la procédure générale de simulation séquentielle, conditionnelle si n est positif, et non conditionnelle si n est nul, qui consiste à simuler les variables une à la fois, en rajoutant à chaque fois la dernière simulation dans l'ensemble de conditionnement. Les probabilités conditionnelles peuvent être facilement calculées dans le cas d'un vecteur aléatoire gaussien. Si la fonction aléatoire n'est pas gaussienne, on commence par lui appliquer la transformation non linéaire qui consiste à faire coïncider sa fonction de répartition empirique $H_F(f)$ =Prob($F \leqslant f$) avec celle d'une variable Y de loi normale centrée réduite qu'on note $G_Y(y)$. On pose donc l'égalité $H_F(f_p) = G_Y(y_p) = p, \forall p \in [0, 1]$. Et on obtient la suite de n valeurs de loi normale $y_l = G_Y^{-1}(\frac{l}{n}), l = 1, ..., n$ qui sont les images des n-tiles de H_F. Les simulations sont alors faites conditionellement aux y_l, puis on retourne à la loi H_F par transformation inverse.

4.3.1 Données simulées

Nous allons appliquer les méthodes d'estimation étudiées à une variable régionalisée, que nous allons simuler, de façon non conditionnelle, sur une grille carrée ayant $20 * 20 = 400$ noeuds, sur le domaine $D = [0, 80] \star [0, 80] \subset R^2$. La variable simulée est la réalisation d'une fonction aléatoire gaussienne, centrée, stationnaire et isotrope, de fonction de covariance définie par la combinaison d'une

sphérique et d'une exponentielle :

$$c\,(h) = 0.05(1 - \left(\frac{3}{2}\frac{h}{a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right) 1_{\{0<h<a\}}) + 0.55\,\exp(-(h/b))$$

où h étant la distance euclidienne dans le plan, $a = 25.08$ et $b = 17.67$ sont les paramètres de portée respectivement de la sphérique et de l'exponentielle. La variance est $\sigma^2 = 1.2$, et pour les estimateurs ks et PLS, nous considérons la valeur moyenne du processus $m = 0$. Nous nous proposons d'estimer la fonction aléatoire au point $M_0(x_0 = 30, y_0 = 42)$. La valeur simulée en M_0 est $s_0 = -0.1596$. Le tableau 3 présente les résultats obtenus pour les six méthodes d'estimation étudiées (l'estimation ponctuelle $f^*(M_0)$ et son erreur quadratique $\phi^2(f^*)$, ainsi que le biais et l'erreur quadratique standardisés moyens b et e définis par 10 et 11) :

TABLEAU 3
Résultats obtenus pour les données simulées

	$f^*(M_0)$	$\phi^2(f^*)$	$b(f^*)$	$e(f^*)$
KS	-0.6452	0.7564	0.008	0.640
PLS	-0.2086	0.9331	0.021	0.781
Inv. dist.	-0.1647	0.9911	0.011	0.835
Pond. covar.	-0.1396	0.9623	0.015	0.799
KO	-0.6495	0.7564	0.001	0.640
PLSO	-0.1922	0.8735	0.053	0.748

L'examen de la dernière colonne du tableau 3 montre que pour les données simulées, et à l'instar des deux exemples précédents, l'erreur quadratique standardisée moyenne de l'estimation PLSO est meilleure que toutes les autres, à l'exception de celles des krigeages. Par ailleurs, le biais conditionnel du krigeage ordinaire s'est avéré ici asez important, puisque la pente de la régression des observations sur les estimations par krigeage ordinaire vaut $p_{ko} = 0.78$, alors que pour l'estimateur PLSO, la pente des valeurs observées sur les valeurs estimées est $p_{plso} = 1.03$, qui est ici encore proche de 1.

5. Conclusions

Le cas où la moyenne est inconnue est le plus fréquent dans la pratique. Le krigeage ordinaire qui est dans ce cas l'estimateur linéaire sans biais, et optimal au sens des moindres carrés, présente un biais conditionnel qui le rend inutilisable dans les situations d'estimations basées sur la sélection des variables dépassant un seuil donné.

Après un krigeage de la moyenne, la régression PLS a été appliquée avec une variante, donnant un estimateur linéaire sans biais conditionnel dans le cas m inconnue. Par ailleurs, la maximisation de la covariance entre la variable aléatoire à estimer et une combinaison linéaire des observations a conduit à l'estimateur de pondération par la covariance qui remplace la distance euclidienne dans l'expression de f_{id}, par la distance associée au produit scalaire défini par la covariance entre variables aléatoires dans l'espace $L^2(\Omega)$, tenant compte ainsi de la structure probabiliste du phénomène étudié. Par ailleurs, comme pour la régression PLS dans le cas m connu, la pondération par la covariance a des coefficients proportionnels et du même signe que la covariance.

Dans l'exemple anisotrope ci-dessus, la pondération par la covariance a amélioré les résultats de l'estimation par inverse de la distance. Et dans les trois situations étudiées l'erreur quadratique standardisée moyenne de f_{plso} est moins élevée que celles de tous les estimateurs autres que les krigeages, mais ceci n'est pas vrai pour les erreurs quadratiques ponctuelles (exemple 1). Enfin, dans les deux exemples étudiés ainsi que sur les données simulées, l'estimateur PLSO a donné des pentes de régression très proches de 1, ce qui confirme son caractère sans biais conditionnel, contrairement au krigeage ordinaire.

Références

AJERAME (1997), Géostatistique appliquée à la quantification du risque, thèse de doctorat à la faculté des sciences agronomiques de Louvain.

BARNES et JOHNSON (1984), Positive kriging, 2nd NATO A.S.I. «Geostatistics for natural resources characterization» Part 2, D, Reidel Publ. Co. Dordrecht, Netherlands.

CHILÈS et DELFINER (1999), *Geostatistics modeling spatial uncertainty*, John Wiley & Sons, Inc.

CRESSIE (1991), *Statistics for spatial data*, John Wiley & sons, Inc.

DEUTSCH et JOURNEL (1992), *GSLIB : Geostatistical Software Library*, Oxford University Press, New York.

ELKETTANI (2001), Analyse des redondances et régression PLS appliquées aux données spatiales. Comparaison avec l'estimation par krigeage et par inverse de la distance, *Revue de Statistique Appliquée*, 49 n° 2, pp 67-82.

HERZFELD (1989), A note on programs performing kriging with non negative weights, *Matematical Geology*, 21, pp. 391-393.

ISAAKS et SRIVASTAVA (1989), *An introduction to applied geostatistics*, Oxford University Press.

LANTUEJOUL (1994), Non conditional simulation of stationary isotropic multigaussian random functions, M armstrong and P.A. Dowd (eds) : Geostatistical Simulations, 147-177, Kluwer Academic Publishers.

MICHAEL L. STEIN (1998), *Interpolation of Spatial Data, some theory for kriging*, Springer Series in Statistics

PANNATIER (1996), *Data Analysis in 2D*, Springer-Verlag.

RIVOIRARD (1984), Le comportement des poids de krigeage, thèse de Docteur-Ingénieur : Ecole des mines de Paris.

SZIDAROVSKY *et al. (1987)*, Kriging without negative weights, *Mathematical Geology*, 19, pp 549-559.

TENENNHAUS (1998), *La régression PLS, théorie et pratique*, éditions TECHNIP.

WOLD, ALBANO *et al. (1983)*, Pattern recognition : Finding and using regularities in multivariate data; Proc IUFOST conf. «Food research and data analysis», Martens J. ed, Applied sciences publications. London.

YATES S.R., YATES M.V. (1998), Disjunctive kriging as an Approach to Management Decision Making, *Soil Sci. Soc. Am. J.*, 52, pp. 1554-1558