REVUE DE STATISTIQUE APPLIQUÉE

YOUSSFI ELKETTANI

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, tome 49, nº 2 (2001), p. 69-84

<http://www.numdam.org/item?id=RSA_2001__49_2_69_0>

© Société française de statistique, 2001, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

Rev. Statistique Appliquée, 2001, **XLIX** (2), 69-84

ANALYSE DES REDONDANCES ET RÉGRESSION PLS APPLIQUÉES AUX DONNÉES SPATIALES. COMPARAISON AVEC L'ESTIMATION PAR KRIGEAGE ET PAR INVERSE DE LA DISTANCE

Youssfi Elkettani

Faculté des Sciences, Département de mathématiques
Laboratoire d'analyse convexe et variationnelle
Systèmes dynamiques et processus stochastiques
BP133, Kénitra, Maroc
email : elkettani.y@usa.net

RÉSUMÉ

L'objet de cet article est de présenter une extension de l'analyse des redondances et de la régression PLS, particulièrement adaptée à l'explication de tableaux de données multicolinéaires, au cas de données spatiales. Cette extension qui repose sur la dite structure, conduit pour l'analyse des redondances au même résultat que l'estimation optimale, obtenue par krigeage simple, tandis que la régression PLS fournit des coefficients de régression mieux interprétables qui permettent de procéder aisément au choix des observations les plus explicatives du point à estimer. La régression PLS est donc une méthode complémentaire au krigeage, et mieux adaptée aux données spatiales que l'estimation par inverse de la distance qui ne fait pas intervenir la structure de covariance du phénomène.

Mots-clés : *Régression PLS, analyse des redondances, inverse de la distance, krigeage simple, poids de krigeage, poids négatifs.*

ABSTRACT

The aim of this article is the extension of the redundancy analysis and the partial least square regression, PLS, particularly adapted for the explanation of the data tables with multicolinearities, to the spatial data. This application, based on the covariance structure of the spatial data, gives for the redundancy analysis, exactly the same results as the simple kriging which is the optimal estimation, while the PLS regression gives coefficients more meaningful and then useful for the choice of the more explicative observations for the point to estimate. Then the PLS regression is an interesting tool for spatial data analysis, completing the kriging, and is more adapted than the inverse distance method which does not use the covariance structure.

Keywords : *PLS regression, redundancy analysis, simple kriging, inverse distance estimation, kriging weights, negative weights.*

1. Introduction : Les méthodes explicatives

Les méthodes d'analyse de données permettent d'étudier les liaisons entre deux ensembles de variables Y et X observés sur les mêmes individus. On distingue :

– d'une part les méthodes descriptives des associations pouvant exister entre Y et X, les deux ensembles de variables jouant alors un rôle symétrique,

– et d'autre part, les méthodes explicatives qui sont utilisées dans les cas où Y constitue un ensemble de variables réponses qu'il s'agit d'expliquer à partir des variables prédicteurs X. On trouve dans ce volet toutes les méthodes de type régression, et dont l'application a touché tous les domaines et presque tous les types de données, y compris les données spatiales. On peut voir Tomassone, Lesquoy et Millier (1983) pour la régression linéaire, ou Antoniadis *et al* (1992) pour la non linéaire. Une généralisation de la régression aux distributions non gaussiennes, et plus particulièrement aux données catégorielles a donné lieu aux modèles linéaires généralisés, voir Nelder et Mc Cullagh (1983, 1989 2nde édition), Pregibon (1979) qui a introduit des méthodes analytiques de choix et de validation de ces modèles, ou Elkettani (1984) pour l'utilisation de la quasivraisemblance dans la modélisation de courbes de dosages, puis Fahmeir et Tutz(1991) pour un panorama des applications y compris les données correlées, et enfin Antoniadis *et al* (1992) pour une extension aux fonctions de prédiction non linéaires. Cependant dans le cas où les colonnes de X sont multicolinéaires, la régression multiple présente des difficultés d'interprétation de l'équation de régression $Y = X \star \beta$, car les composantes du vecteur β sont non représentatives de la corrélation existant entre Y et les colonnes de X. On trouve alors d'autres méthodes explicatives en analyse des données, comme l'analyse des redondances développée par Van den Wollenberg (1977) et la régression PLS introduite par Wold, Albano *et al* (1983) et dont il faut se référer à Tenenhaus, Gauchy et Menardo (1995) puis Tenenhaus (1998) pour d'amples développements sur ses propriétés mathématiques et ses applications. Ces deux dernières méthodes et plus particulièrement la régression PLS sont bien adaptées au traitement des données présentant une multicolinéarité, et il est alors intéressant de se demander quel pourrait être l'apport de ces dernières méthodes dans le traitement des données spatiales qui présentent une corrélation structurée.

Dans le cas d'un champ spatial, la structure des données est bien différente de la situation précédente puisqu'on étudie des phénomènes plutôt caractérisés par leur loi de répartition dans l'espace. L'ensemble des observations est alors considéré comme une réalisation unique d'une fonction aléatoire sur laquelle on émet éventuellement des hypothèses de stationnarité et d'ergodicité qui permettent de faire de l'inférence statistique. L'estimation par krigeage en statistique spatiale se fait par minimisation de l'erreur quadratique, et est basée sur les principes de la régression multiple. Sous l'hypothèse de stationnarité d'ordre 2, on distingue des méthodes linéaires comme le krigeage simple où l'espérance de la fonction aléatoire est supposée une constante connue et le krigeage ordinaire où l'espérance est supposée une constante mais inconnue; on peut voir à ce sujet une abondante littérature, comme par exemple Cressie (1991), ou Wackernagel (1995) ou encore Isaaks et Srivastava (1989). Enfin, dans les cas non stationnaires on trouve le krigeage universel. Et dans le cas non linéaire, le krigeage disjonctif est développé par Rivoirard(1994).

Dans cet article, après un bref rappel de l'analyse des redondances et de la régression PLS, ainsi que des élements de base qui définissent l'estimation spatiale par krigeage simple et par inverse de la distance, nous montrerons que l'analyse des redondances conduit à la même estimation que le krigeage simple, et que la régression PLS, est mieux adaptée que l'estimation par inverse de la distance, et qu'elle donne des coefficients de régression mieux interprétables. Notons enfin que nous préfèrerons employer ci-dessous, pour les différentes méthodes le terme coefficient d'estimation qui correspond plus à notre présente préoccupation plutôt que coefficient de régression.

2. Rappel sur l'analyse des redondances et la régression PLS

Outre la régression multiple qui est la méthode explicative la plus ancienne et la plus utilisée, et dont on peut trouver beaucoup de détails dans Tomassonne, Lesquoy et Millier (1983), nous rappelerons les principes de l'analyse des redondances, suivant la même démarche que Tenenhaus (1998); et de la régression PLS en reprenant la présentation de Tenenhaus, Gauchy et Ménardo (1995).

Dans toute cette partie, nous notons G une matrice $n \times q$ dont les colonnes g_1 , ..., g_q, sont des variables appartenant à R^n qui représentent les réponses, et E une matrice $n \times p$ formée de p variables prédicteurs $e_1, ..., e_p$. Soit $R_{11} = (k_{ij})\,, i, j = 1, p$ la matrice de covariance de E où l'élément $k_{i,j} = \text{cov}\,(e_i, e_j)$, et $R_{12} = \text{cov}\,(E, G)$ la matrice $p \times q$ des covariances entre E et G, et $R_{21} = R_{12}^{'}$.

2.1. L'analyse des redondances

Posons $s \leqslant p$, avec p le rang de E. On cherche s vecteurs t_1 , ..., t_s, de variance 1 et non corrélés, combinaisons linéaires des colonnes de E, maximisant le critère : $\sum_{h=1,\ldots,s} \sum_{k=1,\ldots,q} \text{cov}^2\,(g_k, t_h)$, $\text{cov}\,(g_k, t_h)$ désignant la covariance entre g_k et t_h.

On obtient ainsi des composantes $t_h = Ea_h$ expliquant au mieux l'ensemble des variables $g_k, k = 1, q$. Il faut remarquer, voir Tenenhaus(1998), que les a_h sont vecteurs propres de la matrice $R_{11}^{-1} R_{12} R_{21}$ tandis que les vecteurs $v_h = R_{11}^{\frac{1}{2}} a_h$, sont les vecteurs propres normés de la matrice symétrique $R_{11}^{-\frac{1}{2}} R_{12} R_{21} R_{11}^{-\frac{1}{2}}$.

Des mesures de redondance ont été proposées par plusieurs auteurs, on peut voir à ce sujet Cléroux, Lazraq et Lepage (1994). Ces mesures ont été groupées en deux classes par Cramer et Nicewander (1979) : mesures de redondances qui visent à prédire un ensemble de variables par un autre et mesures d'association qui généralisent le concept de coefficient de corrélation.

Notons enfin que l'analyse des redondances a été utilisée par Muller (1981) pour traiter le cas où les prédicteurs sont fortement corrélés entre eux. A cet effet, il a procédé à des régressions sur les composantes issues de l'analyse des redondances. Nous reprendrons la démarche de Muller sur les données spatiales.

2.2. La régression PLS (partial least square)

L'algorithme de régression PLS a été proposé par Wold, Albano *et al* (1983). C'est un algorithme itératif de décomposition orthogonale de l'espace engendré par E et G, faite par une succession de régressions effectuées sur les résidus.

Etape 0 : On considère E_0 et G_0 les tableaux centrés réduits obtenus à partir de E et G respectivement.

Etape 1 : On cherche $t_1 = E_0 a_1$ et $u_1 = G_0 b_1$, les deux variables qui maximisent $\text{cov}\,(u_1, t_1)$ sous la contrainte $\|a_1\| = \|b_1\| = 1$. On régresse ensuite les deux tableaux E_0 et G_0 sur la variable t_1 obtenant les résidus E_1 et G_1 respectivement; d'où les équations : $E_0 = t_1 p_1' + E_1$ et $G_0 = t_1 r_1' + G_1$. Par la méthode des multiplicateurs de Lagrange on trouve que $a_1$et b_1 sont respectivement, le vecteur propre associé à la plus grande valeur propre de $E_0' G_0 G_0' E_0$ et de $G_0' E_0 E_0' G_0$. Et dans le cas particulier de la régression PLS1, où G_0 est réduit à une seule variable réponse g_1, on a $b_1 = 1$ et

$$a_1' = [cor\,(e_1, g_1)\,, ...cor\,(e_p, g_1)]\,/(\sum_{j=1.p} cor^2\,(e_j, g_1))^{\frac{1}{2}}. \tag{1}$$

Etape 2 : On itère l'étape 1 en remplaçant les tableaux de départ E_0 et G_0 par les tableaux des résidus E_1 et G_1 . On obtient alors deux nouvelles composantes t_2 et u_2 . On régresse ensuite les deux tableaux E_1 et G_1 sur l'axe t_2 d'où les équations : $E_1 = t_2 p_2' + E_2$ et $G_1 = t_2 r_2' + G_2$. Ou encore :

$$E_0 = t_1 p_1' + t_2 p_2' + E_2 \text{ et } G_0 = t_1 r_1' + t_2 r_2' + G_2.$$

Etape 3 : On réitère la procédure A fois, jusqu'à ce que les composantes $t_1, ..., t_A$ expliquent suffisamment G_0. Les $t_1, ..., t_A$ sont des combinaisons linéaires des p colonnes de E_0 de variance 1 et non corrélées mutuellement, et A est un entier compris entre 1 et le rang de E_0. Enfin de l'expression de G_0 on obtient les équations de regression PLS : $g_k = \beta_{k0} + \beta_{k1} e_1 + ... + \beta_{kp} e_p + G_{Ak}$ pour $k = 1, q$, où G_{Ak} est la k^{ieme}colonne de G_k résidu à la A^{ieme}étape. Tenenhaus (1998) présente les propriétés mathématiques de la régression PLS, et plus de développements sur la méthode.

3. Rappels sur l'estimation linéaire en statistique spatiale

3.1. Position du problème

Soit un domaine D de R^p muni d'un espace de probabilité et sur lequel est étudiée une fonction aléatoire $\Phi : D \times \Omega \to R$. Pour $M \in D$, notons par $F(M)$ la variable aléatoire définie par $F\,(M) : \omega \longmapsto \Phi\,(M, \omega)\,, \omega \in \Omega$; et notons par F le processus spatial $F = \{F(M), M \in D\}$. Par ailleurs, pour une réalisation ω_0 donnée, supposons qu'on ait observé la fonction aléatoire Φ en n points $M_1, ..., M_n$ de D, et posons $f\,(M_i) = \Phi\,(M_i, \omega_0)$ la valeur observée au point $M_i, i = 1, n$. Compte

tenu du phénomène physique étudié, qui ne présente presque jamais de répétitions, notamment en géostatistique, Matheron (1970) considère $\{f(M_i), i = 1, n\}$ comme la réalisation unique de la fonction aléatoire Φ.

Toutefois, à cause de cette situation de réalisation unique, l'inférence statistique ne peut être envisagée qu'en introduisant des hypothèses supplémentaires au modèle probabiliste. Rappelons que l'hypothèse de stationnarité d'ordre 2 suppose que $E(F) = m$ est une constante réelle sur D et que $\forall M_1, M_2 \in D$, $\operatorname{cov}(F(M_1), F(M_2)) = C(h)$ ne dépend que de la distance euclidienne h entre M_1 et M_2, et non des points eux mêmes. Le modèle est dit isotrope si toutes les directions de l'espace ont la même fonction de covariance, sinon il est dit anisotrope. Par ailleurs il est d'usage de noter par $M + h$ le point de D translaté de M par un vecteur de norme h. Alors suivant cette notation, la stationarité intrinsèque suppose que ce sont les accroissements $F(M + h) - F(M)$ qui sont stationnaires; le processus F lui même peut ne pas avoir dans ce cas de moment d'ordre 2. On definit alors le variogramme de la fonction aléatoire intrinsèque F par : $\Gamma(h) = \frac{1}{2} \operatorname{var}(F(M + h) - F(M))$. Enfin on émet aussi l'hypothèse d'ergodicité de F qui assure que si la mesure de Lebesgue de D, $L(D)$ augmente indéfiniment, alors la variance d'une estimation portant sur la globalité du champs D tend vers zéro; cette hypothèse n'est pas vérifiée pour tout processus stationnaire, comme par exemple le processus $F(M) = Y, \forall M \in D$ où Y est une quelconque variable aléatoire non nulle dans $L^2(\Omega)$. On peut se référer ici à Cressie (1991) ou Lantuéjoul (1991) pour plus de développements sur cette notion. Dans toute la suite on suppose que F est un processus stationnaire d'ordre 2 et ergodique.

3.2. L'estimation linéaire en statistique spatiale

En estimation spatiale, on peut voir dans Chauvet (1994) par exemple, la distinction entre les problèmes d'estimation globaux qui mettent en jeu la totalité du champ de la variable régionalisée étudiée, comme l'estimation de la moyenne et de la matrice de covariance, et les problèmes locaux qui ne se posent qu'au niveau d'une portion de ce champ, comme la prédiction de la valeur de la fonction aléatoire dans le voisinage d'un point de D. Nous présentons ci-dessous brièvement les premières pour traiter davantage de la prédiction de la variable en un point quelconque du champ.

3.2.1. Estimation de la matrice de covariance

L'estimation de la covariance, comme celle de la moyenne, repose sur des méthodes empiriques. L'estimateur empirique de la covariance, établi sous l'hypothèse de la stationarité d'ordre 2, pour quelques distances $h_\alpha, \alpha = 1, \nu$, est appelé covariance observée (ou variogramme observé quand on travaille sur les variogrammes). Ces estimations empiriques sont ensuite ajustés à un modèle parmi un ensemble de familles de modèles de référence, comme nous allons le voir ci-dessous. Dans cette sous-section on suppose que F est un processus stationnaire du second ordre, de fonction de covariance C qu'on se propose d'estimer empiriquement, pour une direction donnée dans le cas anisotrope ou bien pour toutes directions confondues dans le cas isotrope. Notons $M = \{M_i, i = 1, n\}$ l'en-

semble des points où on a observé F, et H l'ensembe des distances entre les éléments de M pris 2 à 2, et soit $\varepsilon \succ 0$ un paramètre dit de tolérance. Alors pour une suite d'intervalles $I_\alpha = [h_\alpha - \varepsilon, h_\alpha + \varepsilon], \alpha = 1, \nu$, de ν classes de distances dans l'espace qui forme une partition de H, on estime $C(h_\alpha)$ par $C^*(h_\alpha)$, moyenne empirique des covariances observées pour tout couple de points dans la classe de distance I_α, pris parmi toutes les observations disponibles $f_i = f(M_i), i = 1, n$. Le nuage des valeurs $(C^*(h_\alpha))_{\alpha=1,\nu}$ est ensuite ajusté à un des modèles de référence, que nous noterons en lettre minuscule $c(h)$. Il faut se référer à Cressie (1991), où on trouve un développement des méthodes d'ajustement utilisées. Dans R^2 on peut citer à titre d'exemples de modèles usuels : la covariance exponentielle définie par : $c(h) = c(0)\exp\left(-\frac{3h}{a}\right)$; la covariance sphérique définie par $c(h) = c(0)\left(1 - \frac{3}{2}\frac{h}{a} + \frac{1}{2}\frac{h^3}{a^3}\right)\chi_{\{h<a\}}$, où χ est la fonction indicatrice; la covariance gaussienne définie par $c(h) = c(0)\exp\left(-\frac{\sqrt{3}h}{a}\right)^2$; et la covariance sinus-cardinal définie par $c(h) = c(0)\frac{a}{h}\sin\left(\frac{h}{a}\right)$.

Tandis que la première, se caractérise par ses propriétés markoviennes, la seconde covariance présente une portée finie a, distance au delà de laquelle le processus n'est plus autocorrélé. Et pour les modèles exponentiels et gaussiens, le paramètre a est utilisé dans la pratique comme une approximation de la portée, car à cette distance les autocorrélations du processus deviennent négligeables. Enfin la dernière fonction de covariance présente des valeurs négatives pour certaines distances. Par ailleurs, la variance ponctuelle d'un point du processus, $C(0)$, est estimée par $c(0)$ appelé palier, et on a une relation entre la covariance usuelle $c(h)$ et le variogramme usuel $\gamma(h)$ donnée par : $\gamma(h) = c(0) - c(h)$. On peut citer ici, Christakos(1992) qui étudie les propriétés mathématiques, Isaaks et Srivastana(1989) et Ajerame (1997) qui présentent les caractéristiques et appliquent les différents modèles. Enfin Lantuéjoul (1991) a élargi la notion de portée, en définissant la portée intégrale A d'un processus stationnaire ergodique, telle que la variance d'une estimation portant sur la globalité du champ, soit approximativement proportionnelle à $A/L(D)$ où $L(D)$ est la mesure de Lebesgue de D. Par exemple si on estime $E(F)$ par m^*, alors l'expression de A est donnée par : $A = \lim_{D \longrightarrow \infty} L(D)\frac{\text{var}(m^*)}{c(0)}$, ce qui donne pour D suffisamment grand, var $(m^*) \approx c(0)(A/L(D))$, et si on pose $L(D)/A = N$, pour un entier N, alors var $(m^*) \approx c(0)/N$. Tout se passe comme si le champ D de mesure $L(D)$ était partagé en N parties indépendantes de mesure A.

3.2.2. Prédiction spatiale

Abordons maintenant le problème d'estimation locale, objet de la présente étude, et qui est d'estimer, ou de prédire, pour reprendre l'expression de la géostatistique, la valeur de la variable aléatoire $F_0 = F(M)$ pour un point M non observé de D à partir d'observations $f_i = f(M_i), i = 1, n; M_i \in D$, faites dans un voisinage de M. Le processus F est, dans tout ce qui suit, considéré stationnaire, ergodique,

d'espérance $m \in R$ supposée connue et de fonction de covariance $c(h)$ prédéterminée et présentant un palier $c(0)$, estimation de la variance ponctuelle $\sigma^2 = \text{var}(F(M))$. L'expression de l'estimateur de F_0, étant une fonctionnelle des distances h_i entre M et les points d'observations M_i, elle permet d'obtenir une estimation pour tout point non observé dans le voisinage.

Nous présentons dans cette section les estimations par krigeage simple noté k_s, et par inverse de la distance noté i_d, puis dans la section suivante nous proposerons des estimations basées sur les méthodes explicatives multidimensionnelles. Notons f le vecteur des observations $f = (f_i)_{i=1,n}$, z le vecteur des observations centrées réduites $z = (z_i)_{i=1,n}$où $z_i = (f_i - m)/\sigma$, et Z_0 la variable $(F_0 - m)/\sigma$. Puis conservons les notations des méthodes multidimensionnelles, en posant $R_{12} = (c_1, ..., c_n)'$ le vecteur colonne où $c_i = \text{cov}(Z_o, z_i)$, R_{11} la matrice $n \times n$ dont l'élément ij est $k_{ij} = \text{cov}(z_i, z_j)$, et $R_{21} = R'_{12}$.

Enfin soit $z_0 = \lambda' z$ l'estimateur linéaire de la variable centrée réduite Z_0, où $\lambda = (\lambda_i)_{i=1,n} \in R^n$ est le vecteur des coefficients d'estimation, alors f_0 l'estimateur de F_0 donné par $f_0 = m + \sigma z_0$ est sans biais. Par ailleurs, en notant $\phi^2(z_0)$ et $\phi^2(f_0)$ les variances des estimateurs z_0 et f_0 respectivement, on a : $\phi^2(z_0) = E(z_0 - Z_0)^2 = \lambda' R_{11}\lambda + 1 - 2\lambda' R_{12}$, et $\phi^2(f_0) = \sigma^2\phi^2(z_0)$.

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 $\lambda' z$ minimise l'erreur quadratique $E = E\left(Z_0 - \lambda' z\right)^2$. La résolution de cette équation donne $\lambda = R_{11}^{-1} R_{12}$ que nous noterons par la suite λ_{ks}. L'estimateur de Z_0 est alors :

$$z_{ks} = R_{21} R_{11}^{-1} z. \tag{2}$$

Par ailleurs, le minimum atteint par l'erreur quadratique moyenne pour z_0 est : $\phi^2_{ks}(z_0) = E(z_{ks} - Z_0)^2 = \lambda'_{ks} R_{11} \lambda_{ks} + 1 - 2\lambda'_{ks} R_{12} = 1 - R_{21} R_{11}^{-1} R_{12} = 1 - \lambda'_{ks} R_{12}$.

Et pour la variable f_0 on a : $\phi^2_{ks}(f_0) = \sigma^2 \phi^2_{ks}(z_0)$.

L'estimation par inverse de la distance :

Par contre par la méthode de l'inverse de la distance, on donne simplement à chaque observation z_i un poids inversement proportionnel à la distance d_i entre Z_0 et z_i, puis on somme à 1 le vecteur des poids. L'estimateur par inverse de la distance s'écrit alors $z_{id} = \sum_{i=1,n} (\frac{1}{di} z_i) / \sum \frac{1}{di} = \lambda'_{id} z$, en notant $\lambda_{id} = \left((\frac{1}{di}) / \sum_{j=1,n} \frac{1}{d_j} \right)_{i=1,n}$ le vecteur des pondérations. Et l'erreur quadratique moyenne pour cette méthode est : $\phi^2_{id}(z_0) = E(z_{id} - Z_0)^2 = \lambda'_{id} R_{11} \lambda_{id} + 1 - 2\lambda'_{id}$, et $\phi^2_{id}(f_0) = \sigma^2 \phi^2_{id}(z_0)$.

Nous verrons dans la section « comparaisons », comment cette estimation présente un handicap car elle ne fait pas intérvenir la structure de covariance du processus.

4. Estimation spatiale basée sur les méthodes explicatives

4.1. L'analyse des redondances sur données spatiales

Nous allons procéder ici, à une analyse des redondances de Z_o sur le vecteur ligne des observations centrées réduites z. Il est clair que dans notre situation de processus spatial nous n'avons pas deux tableaux de données pour appliquer l'algorithme de l'analyse des redondances. Et en reprenant les notations de la section 2.1, on a $s = q = 1$. Par contre nous allons exploiter la structure de covariance existant entre les données pour calculer les quantités nécessaires. Nous allons montrer que l'estimateur obtenu en régressant Z_o sur l'axe des redondances est identique à l'estimateur obtenu par le krigeage simple.

Démonstration. On cherche l'unique variable $t = \lambda' z$ de variance 1 ($\lambda' R_{11}\lambda = 1$), qui maximise $cor^2\,(Z_o, t) = \left(\lambda' R_{12}\right)^2$. Il s'agit donc de maximiser l'expression E donnée par :

$$E = \lambda' R_{12} R_{21} \lambda - \mu(\lambda' R_{11}\lambda - 1)$$

où μ est un multiplicateur de lagrange.

L'annulation de la dérivée partielle de E par rapport à λ nous donne :

$$R_{12} R_{21} \lambda = \mu R_{11} \lambda \tag{3}$$

et l'annulation de la dérivée partielle par rapport à μ restitue la contrainte de normalisation

$$\lambda' R_{11} \lambda = 1. \tag{4}$$

d'où l'on déduit

$$R_{11}^{-1} R_{12} R_{21} \lambda = \mu\lambda. \tag{5}$$

$R_{21}\lambda$ étant un scalaire, λ est proportionnel à $R_{11}^{-1} R_{12}$, d'où tenant compte de la contrainte (4) :

$$\lambda = R_{11}^{-1} R_{12} / \left(R_{21} R_{11}^{-1} R_{12}\right)^{\frac{1}{2}} \tag{6}$$

On déduit alors de (3),(4) et (6) que :

$$\mu = \lambda' R_{12} R_{21} \lambda = \left(\lambda' R_{12}\right)^2 = R_{21} R_{11}^{-1} R_{12} \tag{7}$$

$$\lambda' R_{12}\lambda = \sqrt{\mu}\lambda = R_{11}^{-1}R_{12} \tag{8}$$

Régressons maintenant Z_o sur la variable $t = \lambda' z$ suivant la méthode proposée par Muller (1981). On obtient alors l'estimateur des redondances $z_{Rd} = t\theta$ avec $\theta = \text{cov}\,(Z_o, t)$, la variance de t étant égale à 1. Or $\text{cov}\,(Z_o, t) = \lambda' R_{12}$ et l'estimateur devient d'après (8) :

$$z_{Rd} = \left(\lambda' R_{12}\right)\lambda' z = \left(R_{11}^{-1}R_{12}\right)' z.$$

Et d'après (2), on a bien que $z_{Rd} = z_{ks}$, l'estimateur obtenu par la méthode du krigeage simple.

4.2. Estimation spatiale basée sur la régression PLS

4.2.1. Présentation

Nous allons maintenant nous inspirer de la régression PLS pour obtenir un nouvel estimateur linéaire de Z_0. Il s'agit donc de chercher la combinaison linéaire $t = \lambda' z$, avec $\|\lambda\| = 1$ telle que $\text{cov}\,(Z_0, t)$ soit maximale. Ceci revient à rechercher la première composante de la régression PLS de Z_0 sur $z = \{z_i = z\,(M_i)\,, i = 1...n, M_i \in D\}$. Comme pour l'analyse des redondances, z étant un processus spatial nous n'avons pas deux tableaux de données pour appliquer la régression PLS; par contre nous allons exploiter la structure de covariance existant entre les données pour calculer les quantités ci-dessus. De même, nous n'avons qu'une seule composante PLS, t car le tableau des régresseurs z est de dimension $(1 \times n)$.

4.2.2. Recherche de la combinaison linéaire t

Soit μ un multiplicateur de Lagrange, on va maximiser l'expression :

$$E = R_{21}\lambda - \mu\left(\lambda'\lambda - 1\right).$$

L'annulation de la dérivée partielle de E par rapport à λ nous donne :

$R_{12} - 2\mu\lambda = 0$ et l'annulation de la dérivée partielle de E par rapport à μ nous restitue la contrainte de normalité. On en déduit alors que : $2\mu = \lambda' R_{12}$ et que : $\lambda = \dfrac{1}{2\mu}R_{12}$. Et en remplaçant λ par son expression dans l'équation de la contrainte, on tire que : $\mu = \dfrac{1}{2}\left(R_{21}R_{12}\right)^{\frac{1}{2}}$, et donc que $\lambda = \dfrac{1}{\left(R_{21}R_{12}\right)^{\frac{1}{2}}}R_{12} = R_{12}/\,\|R_{12}\|$, qui est bien l'expression (1) rencontrée dans la section 2.2.

Remarque :

Ce résultat peut aussi s'obtenir comme suit : Notons par $\langle u, v\rangle = E\,(u'.v)$ le produit scalaire dans $L^2\,(D)$, et écrivons : $\text{cov}\,(Z_0, t) = R_{21}\lambda =< R_{12}, \lambda >$.

Alors maximiser $< R_{12}, \lambda >= \|R_{12}\| \, \|\lambda\| \cos(R_{12}, \lambda)$, sous la contrainte $\|\lambda\| = 1$, revient à prendre $\cos(R_{12}, \lambda) = 1$, ce qui est atteint pour $\lambda = \beta R_{12}$ avec $\beta \in R$. Et puisque λ est normé, on a $\beta = \dfrac{1}{\|R_{12}\|}$.

Par ailleurs la composante PLS est $t = \lambda' z$, et on a $\mathrm{var}(t) = \lambda' R_{11} \lambda = R_{21} R_{11} R_{12} / (R_{21} R_{12})$. Et la covariance maximale obtenue entre Z_o et t est égale à : $mx\mathrm{cov} = \lambda' R_{12} = (R_{21} R_{12})^{\frac{1}{2}}$. Enfin on régresse Z_o sur t pour obtenir l'estimateur PLS z_{pls} donné par :

$$\begin{aligned} z_{\mathrm{pls}} &= (\mathrm{cov}(z_o, t) / \mathrm{var}(t)) t \\ &= (R_{21} R_{12})^{\frac{1}{2}} \frac{R_{21} R_{12}}{R_{21} R_{11} R_{12}} \frac{R_{21}}{\|R_{12}\|} z = \frac{R_{21} R_{12}}{R_{21} R_{11} R_{12}} R_{21} z = \lambda'_{\mathrm{pls}} z \end{aligned}$$

en notant le vecteur des coefficients dans cette nouvelle estimation :

$$\lambda_{\mathrm{pls}} = \frac{R_{21} R_{12}}{R_{21} R_{11} R_{12}} R_{12}.$$

Et l'erreur quadratique moyenne pour cette méthode d'estimation est :

$$\begin{aligned} \phi^2(z_0) &= E(z_{\mathrm{pls}} - Z_0)^2 = \lambda'_{\mathrm{pls}} R_{11} \lambda_{\mathrm{pls}} + 1 - 2\lambda'_{\mathrm{pls}} R_{12} \\ &= 1 - (R_{21} R_{12})^2 / (R_{21} R_{11} R_{12}). \end{aligned}$$

Enfin pour f_0 on a : $\phi^2(f_0) = \sigma^2 \phi^2(z_0)$.

5. Comparaison des différents estimateurs

Nous allons relever dans cette section quelques éléments de comparaison entre les trois méthodes d'estimation étudiées :

1) La variance d'estimation

Tout d'abord le krigeage est l'unique méthode qui minimise l'erreur quadratique. Par conséquent f_{ks} donne plus de précision que les deux autres.

2) La structure de covariance

Comme il est déjà signalé ci-dessus, la méthode de l'inverse de la distance ne tient pas compte de la structure de covariance du champ étudié, et ses résultats deviennent médiocres dès que la fonction aléatoire étudiée présente une structure plus fine que la simple dépendance de la distance euclidienne. Par exemple l'estimation par inverse de la distance ne distingue pas entre les différents axes d'orientation dans le cas d'un champ anisotrope.

3) La simplification algorithmique

Comme nous l'avons vu plus haut, l'expression de λ_{ks} et donc de z_{ks} fait intervenir l'inverse de la matrice R_{11}, de dimension nxn qui peut être de taille trop

grande. On préconise alors de supprimer les observations éloignées uniquement pour réduire la taille de la matrice comme le souligne Journel et Huijbregts (1978). Mais d'autres auteurs ne sont pas satisfaits par cette procédure, et à ce sujet Davis et Grivet (1984) proposent des méthodes d'inversion de matrices de grande taille, pour éviter de supprimer des observations du modèle. Par contre les algorithmes d'estimation par inverse de la distance et par régression PLS sont plus simples, et ne font pas intervenir d'inversion de matrices.

4) Le signe des coefficients

C'est là une critique majeure faite à l'encontre de l'estimation par krigeage. En effet cette méthode présente des coefficients λ_{ks} pouvant être des deux signes. Et ceci peut même conduire parfois à des estimations négatives de variables devant être toujours positives, comme la teneur d'un élément faiblement dosé dans l'espace. Les utilisateurs du krigeage qui n'apprécient pas de telles situations improvisent devant chaque problème. Barnes et Johnson (1984) ont proposé d'imposer au krigeage une condition de positivité des poids, et des algorithmes pour résoudre ce problème sont développés également dans Szidarovsky *et al* (1987), puis Herzfeld (1989). Par contre les cefficients λ_{id} obtenus par inverse de la distance sont positifs, et il en est de même pour les coefficients λ_{pls} dans les cas où la fonction de covariance est positive; ce qui représente la grande majorité des cas, car les fonctions usuelles sont en général positives à l'exception du modèle sinus-cardinal. En fait il est clair à ce propos que les coefficients de la régression PLS sont mieux interprétables que ceux obtenus par les deux autres méthodes, car le vecteur λ_{pls} est directement proportionnel au vecteur R_{12} des covariances entre le point à estimer Z_0 et les observations z_i.

6. Exemple

6.1. Présentation

Afin d'apprécier concrétement l'apport de l'estimation par régression PLS introduite ci-dessus par rapport aux deux méthodes classiques d'estimation spatiale linéaire, le krigeage simple et l'inverse de la distance, appliquons ces trois méthodes aux données relatives aux mesures d'élévation du sol en différents points du lac Walker dans l'État du Nevada aux U.S.A. Les données des 195 observations relatives au premier programme de mesures sont reproduites et très largement exploitées dans Isaaks et Srivastana (1989), de même qu'ils sont disponibles dans le logiciel Variowin de Pannatier (1996). Ces 195 observations sont régulièrement réparties sur une grille carrée, couvrant tout le lac.

6.2. Premier modèle

Pour faire de l'estimation spatiale, nous reprenons le modèle stationnaire établi dans Isaaks et Srivastana (1989) pour la variable V, fonctionnelle de l'élévation du sol. La moyenne de V est $m = 275$ mètres, l'écart-type ponctuel est $\sigma = 250$, et la fonction de covariance retenue est une combinaison anisotrope de covariances sphériques. En notant par χ la fonction indicatrice d'un ensemble, l'expression de la

fonction de covariance est donnée par :

$$c(h) = 22500 + \sigma^2 - 40000\,\mathrm{sph}(h_1) - 45000\,\mathrm{sph}(h_2)$$

avec : $\mathrm{sph}(h_i) = \left(\frac{3}{2}h_i - \frac{1}{2}(h_i)^3\right)\chi\{0 < h_i < 1\}, i = 1, 2,$ et $h_i = \left(h_{ix}^2 + h_{iy}^2\right)^{\frac{1}{2}}$; où les composantes $h_{ij}, i = 1, 2; j = x, y$, obtenues par changement de coordonnées polaires suivant les axes d'anisotropie sont données par les produits matriciels :

$$\begin{pmatrix} h_{1x} \\ h_{1y} \end{pmatrix} = \begin{bmatrix} \frac{1}{25} & 0 \\ 0 & \frac{1}{30} \end{bmatrix} \begin{bmatrix} \cos 14 & \sin 14 \\ -\sin 14 & \cos 14 \end{bmatrix} \begin{pmatrix} h_x \\ h_y \end{pmatrix},$$

et

$$\begin{pmatrix} h_{2x} \\ h_{2y} \end{pmatrix} = \begin{bmatrix} \frac{1}{50} & 0 \\ 0 & \frac{1}{150} \end{bmatrix} \begin{bmatrix} \cos 14 & \sin 14 \\ -\sin 14 & \cos 14 \end{bmatrix} \begin{pmatrix} h_x \\ h_y \end{pmatrix}$$

avec h_x et h_y les coordonnées cartésiennes du vecteur joignant deux points du plan; les axes du repère cartésien étant l'axe est-ouest en abscisse, et nord-sud en ordonnée.

On se propose ici d'estimer V_0, valeur de V en un point $M_0(x_0, y_0)$ du lac. Prenons pour mieux apprécier les estimations, un point déjà observé, soit $x_0 = 211$, et $y_0 = 250$; que l'on soustrait aux observations. La valeur observée en ce point étant $v_0 = 166.9$, peut alors être comparée aux valeurs estimées présentées dans le tableau 1 ci-dessous ($v_{id} = 238.85, v_{ks} = 180.52, v_{\mathrm{pls}} = 171.21$). Toutefois, compte tenu de la grande étendue de la variable V (le minimum observé pour V est 0, et le maximum est 975.3 mètres), pour être significatives, les comparaisons valeurs estimées – valeurs observées devraient être faites pour toutes les observation conduisant à des écarts moyens selon la méthode Jacknife. Dans notre situation, connaissant les variances théoriques des estimateurs, nous pouvons comparer directement les précisions des estimations.

TABLEAU 1. *(194 obs.)*

	inverse distance	krigeage simple	régression PLS
estimation de Z_0	$z_{id} = -0.1446$	$z_{ks} = -0.3779$	$z_{\mathrm{pls}} = -0.4154$
variance de z_0	$\phi_{id}^2(z) = 0.8529$	$\phi_{ks}^2(z) = 0.4823$	$\phi_{\mathrm{pls}}^2(z) = 0.5771$
estimation de V_0	$v_{id} = 238.8506$	$v_{ks} = 180.5229$	$v_{\mathrm{pls}} = 171.2147$
variance de v_0	$\phi_{id}^2(v) = 53309$	$\phi_{ks}^2(v) = 30145$	$\phi_{\mathrm{pls}}^2(v) = 36067$

L'examen des résultats du tableau 1 montre que l'estimation de V_0 par krigeage est bien sûr celle qui présente la plus petite variance$\left(\phi_{ks}^2(v) = 30145\right)$; par contre l'estimation par régression PLS est plus précise $(\phi_{\mathrm{pls}}^2(v) = 36067)$, que celle obtenue par inverse de la distance $\left(\phi_{id}^2(v) = 53309\right)$.

6.3. Modèle réduit

L'examen des coefficients d'estimation par régression PLS λ_{pls} va nous permettre de réduire considérablement la dimension du modèle. En effet sur les 194 composantes de ce vecteur, seules 34 sont non nulles, et 21 seulement ont une valeur supérieure à 0.01, ces 21 valeurs sont en fait comprises entre 0.011 et 0.145. Par contre les coefficients du krigeage λ_{ks} ne permettent pas de procéder de façon aussi nette à une telle réduction du modèle. Les composantes de λ_{ks} sont comprises entre -0.018 et 0.301, dont 42 valeurs négatives, 98 nulles, et seulement 8 sont supérieures à 0.01. Les composantes nulles dans λ_{ks} et λ_{pls} proviennent des portées incluses dans les fonctions $\text{sph}\,(h_i)\,, i = 1, 2$ qui constituent la fonction de covariance du phénomène, alors que le vecteur λ_{id} a toutes ses composantes strictement positives.

TABLEAU 2. *(21 obs.)*

	inverse distance	krigeage simple	régression PLS
estimation de v_0	$v_{id} = 208.3107$	$v_{ks} = 182.3297$	$v_{\text{pls}} = 170.5499$
variance de v_0	$\phi^2_{id}(v) = 40683$	$\phi^2_{ks}(v) = 30359$	$\phi^2_{\text{pls}}(v) = 35657$

La figure 1 montre la position dans l'espace des 21 observations ayant eu un coefficient de régression PLS supérieur à 0.01, ainsi que celle du point $M_0\,(x_0, y_0)$ dont on veut estimer la valeur V_0. On voit que cette configuration n'est pas symétrique par rapport au point à estimer. Ceci est dû au caractère anisotrope de la fonction de covariance $c\,(h)$ qui régit le phénomène étudié. Et avec le même modèle que précédemment, appliquons les trois méthodes d'estimation aux 21 observations retenues. Les résultats de ces estimations figurent dans le tableau 2.

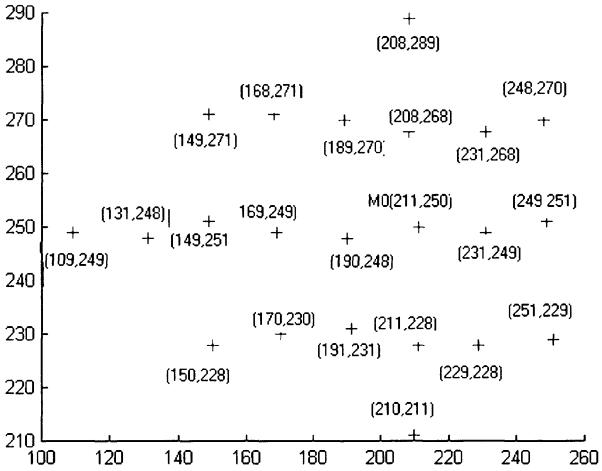

FIGURE 1

On remarque à la lecture du tableau 2 que les modifications occasionnées par la réduction du modèle à 21 observations, dans les précisions des estimateurs, sont tout à fait négligeables pour le krigeage simple et pour la régression PLS. Les résultats de l'estimation par inverse de la distance se sont par contre beaucoup améliorés, ce qui prouve que les observations supprimées, étaient très mal pondérées par cette méthode.

Enfin le tableau 3 présente, les coordonnées cartésiennes x et y des 21 observations retenues, les trois vecteurs des coefficients d'estimation, ainsi que le vecteur des covariances entre la variable à estimer Z_0 et les observations z, pour le modèle réduit.

TABLEAU 3

Coordonnées spatiales, coefficients d'estimations, et le vecteur cov (Z_0, z)

x	y	λks	λpls	λdist	cov	x	y	λks	λpls	λdist	cov
109	249	−0.008	0.016	0.016	0.069	208	268	0.178	0.092	0.091	0.407
131	248	0.005	0.036	0.021	0.159	208	289	−0.011	0.011	0.043	0.048
150	228	−0.022	0.018	0.026	0.080	229	228	−0.002	0.052	0.059	0.229
149	251	0.047	0.061	0.027	0.267	231	249	0.304	0.147	0.083	0.650
170	230	0.005	0.039	0.037	0.173	231	268	−0.001	0.065	0.062	0.287
169	249	0.053	0.090	0.040	0.398	251	229	0.001	0.037	0.037	0.164
168	271	0.001	0.035	0.035	0.153	249	251	0.061	0.097	0.044	0.427
191	231	−0.002	0.061	0.060	0.270	248	270	−0.001	0.043	0.040	0.189
189	270	−0.009	0.056	0.056	0.247	190	248	0.265	0.139	0.079	0.613
210	211	−0.012	0.011	0.043	0.050	149	271	−0.022	0.019	0.026	0.085
211	228	0.102	0.066	0.076	0.293						

On peut remarquer sur les colonnes λks et cov de ce tableau que certaines observations, comme $M_3\,(191, 231)$, $M_4\,(189, 270)$ ou $M_5\,(231, 268)$, ayant eu un poids négatif par le krigeage simple ont une covariance non négligeable avec le point à estimer; ce qui rend difficile l'interprétation des coefficients d'estimation du krigeage. En fait, on s'aperçoit que l'estimation f_{ks} est basée principalement sur les deux observations les plus proches du point à estimer à savoir $M_1\,(231, 249)$ et $M_2\,(190, 248)$, alors que l'estimation par régression PLS s'est faite sur la base d'une répartition des poids, élargie aux 21 observations.

7. Conclusions

Nous avons introduit ci-dessus la régression PLS sur données spatiales que nous avons confrontée à deux méthodes classiques d'estimation, le krigeage simple qui n'est d'ailleurs autre que la régression sur l'axe des redondances et l'inverse de la distance qui ne fait intervenir que les distances entre le point à estimer et les observations. Appliquée aux données du lac Walker, l'estimation par régression PLS a donné des résultats assez proches de l'estimation optimale obtenue par le krigeage simple. Par contre, la méthode de l'inverse de la distance, ne tenant pas compte de la structure de covariance, a été inadaptée au phénomène anisotrope étudié. Par ailleurs, la régression PLS, a fourni des coefficients mieux interprétables que ceux du krigeage simple, et qui ont permis de procéder naturellement à la réduction du modèle. Aussi, l'estimation spatiale par régression PLS, qui est obtenue par un algorithme numériquement plus simple, apporte des élements nouveaux et intéressants qui viennent en complément de l'estimation optimale, linéaire stationnaire à moyenne connue, qu'est le krigeage simple. Elle permet notamment d'apprécier l'apport de chaque observation dans la prédiction du point non observé, et de procéder aisément à une sélection des observations les plus explicatives de ce point.

Références bibliographiques

AJERAME (1997) : Géostatistique appliquée à la quantification du risque; thèse de doctorat à la faculté des sciences agronomiques de LOUVAIN.

ANTONIADIS *et al* (1992) : Régression non linéaire et applications; éditions ECONOMICA.

BARNES et JOHNSON (1984) : Positive kriging, 2nd NATO A.S.I. "Geostatistics for natural ressources characterisation" Part 2, D, Reidel Publ. Co. Dordrecht, Netherlands.

CHAUVET (1994) : Cahiers de géostatistique, fascicule 2, Ecole des mines de Paris.

CHRISTAKOS (1992) : Random Field models in earth science; Academic Press, Inc.

CLEROUX, LAZRAQ et LEPAGE (1994) : Indice de redondance basé sur les rangs et inférence non paramétrique; Revue de Statistique Appliquée 42 n° 2, pp. 79-98.

CRAMER et NICEWANDER (1979) : Some symetric invariant measures of multivariate association; Psychometrika 44, pp. 43-54.

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

DAVIS and GRIVET (1984) : Kriging in a global neighbourhood, Mathematical Geology 16 n° 3, pp. 249-265.

ELKETTANI (1984) : Etudes de modèles linéaires généralisés; Doctorat de 3eme cycle, Faculté des Sciences d'Orsay, Université Paris XI.

FAHMEIR et TUTZ (1991) : Multivariate statistical modelling based on generalized linear models; Springer.

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

JOURNEL and HUIJBREGHTS (1978) : Mining Geostatistics, Academic Press.

ISAAKS et SRIVASTANA (1989) : An introduction to applied geostatistics; Oxford University Press.

LANTUÉJOUL (1991) : Ergodicity and integral range; journal of microscopy 161, pp. 387-403.

MATHERON (1970) : The theory of regionalised variables; Centre de Géostatistique, Ecole des mines de Paris, Fontainebleau, France.

MULLER (1981) : Relationships between redundancy analysis, canonical correlation and multivarate regression; Psychometrika 46, pp. 139-142.

NELDER et MCCULLAGH (1983, 1989 2nde edition) : Generalized linear models; Chapman and Hall.

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

PREGIBON (1979) : Data analytic methods for generalized linear models; Ph-D thesis, Université de Toronto.

RIVOIRARD (1994) : Introduction to disjonctive kriging and non linear geostatistics; Clarendon Press, Oxford.

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

TENENHAUS, GAUCHY et MENARDO (1995) : Régression PLS et applications; Revue de Statistique Appliquée 43 n° 1, pp. 7-63

TENENHAUS, M. (1998) : La régression PLS, théorie et pratique; éditions TECHNIP

TOMASSONE, LESQUOY et MILLIER (1983) : La régression : nouveaux regards sur une ancienne méthode statistique, INRA et Masson – Paris.

VAN DEN WOLLENBERG (1977) : Redundancy analysis, an alternative for canonical correlation; Psychometrika 42, pp. 207-219.

WACKERNAGEL (1995) : Multivariate Geostatistics, Springer.

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.