REVUE DE STATISTIQUE APPLIQUÉE

MICHEL ARNAUD
XAVIER EMERY
CHANTAL DE FOUQUET
MARINUS BROUWERS
MICHEL FORTIER

L'analyse krigeante pour le classement d'observations spatiales et multivariées

Revue de statistique appliquée, tome 49, n° 2 (2001), p. 45-67
<http://www.numdam.org/item?id=RSA_2001__49_2_45_0>

Rev. Statistique Appliquée, 2001, **XLIX** (2), 45-67

L'ANALYSE KRIGEANTE POUR LE CLASSEMENT D'OBSERVATIONS SPATIALES ET MULTIVARIÉES

Michel Arnaud[1], Xavier Emery[2], Chantal de Fouquet[2], Marinus Brouwers[1], Michel Fortier[1]

1. CIRAD Avenue du Val de Montferrand, BP5035, 34032 Montpellier Cedex

2. Ecole des Mines de Paris, Centre de Géostatistique, 35, rue Saint-Honoré, 77305 Fontainebleau Cedex

RÉSUMÉ

Lorsque l'ingénieur, le chercheur... désire classer des observations multivariées, un certain nombre de méthodes statistiques sont mises à sa disposition. Lorsque, en outre, les observations sont disposées dans un espace géographique et que l'information spatiale est riche de sens et doit donc être prise en compte, les méthodes adéquates sont beaucoup moins nombreuses. L'article propose une méthodologie d'analyse pour obtenir des groupes de sites ayant les mêmes caractéristiques. Basée sur les techniques d'analyses de données et de géostatistique elle permet également à l'utilisateur de choisir le niveau de perception (échelle) pour le classement.

Mots-clés : *Classification, Géostatistique, Analyse en Composantes Principales, Analyse krigeante.*

ABSTRACT

There are various statistical methods that consultants, researchers, etc can use to classify multivariate observations. However, if their observations concern a specific geographic area and the corresponding spatial information is highly relevant and thus needs to be taken into account, the choice of appropriate methods is much more limited. This article proposes an analysis method aimed at identifying groups of sites with the same characteristics. It is based on data analysis and geostatistical techniques, and also enables users to select a perception level (scale) for classification.

Keywords : *Classification, Geostatistics, Principal Component Analysis, Multivariate Factorial Kriging.*

1. Introduction

Cartographier une zone géographique $\mathcal{D}$ à partir d'observations multivariées et géoréférencées est un objectif courant dans de nombreuses disciplines traitant des phénomènes naturels qui s'étendent dans l'espace : météorologie, agronomie, environnement, géographie... On retrouve cet objectif en sciences du sol pour des études détaillées destinées par exemple à :

• l'analyse des causes attribuables au sol et pouvant expliquer la variabilité spatiale constatée dans les paramètres du rendement au sein d'une parcelle;

• la mise en place d'essais agronomiques afin que la variabilité de l'essai, attribuable aux variations spatiales existant au sein de la parcelle dans les caractéristiques du sol, soit la plus faible possible;

• mieux connaître l'extension spatiale du ou des types de sols au sein de la parcelle sur lesquels on fera en un endroit déterminé des multiples mesures dans le temps destinées à mieux comprendre la croissance de la culture, son bilan hydrique...

Lorsqu'il désire cartographier une zone déterminée, le pédologue prélève des échantillons de sol et relève avec précision leur position sur le terrain. Ces échantillons seront, par la suite, analysés au laboratoire et caractérisés par un ensemble de valeurs quantitatives (les variables physico-chimiques : granulométrie, CEC, bases échangeables, pH, N, C...) ainsi que par leurs coordonnées dans la zone. Pour chaque variable, une première «cartographie» sommaire de la zone, peut s'obtenir, sans avoir recours à des outils compliqués. Un moyen, parmi d'autres, consiste à découper l'histogramme de la variable en plusieurs classes de valeurs (faible, moyenne et forte) et à identifier, par des symboles (ou des couleurs) différents les positions des échantillons selon leur classe d'appartenance. Il est possible alors soit de repérer des secteurs «riches», des secteurs «pauvres», des gradients ou, au contraire, d'obtenir une répartition homogène des symboles sur toute l'étendue de la zone sans lien avec la position des échantillons. Le problème se complique lorsqu'on veut utiliser simultanément l'ensemble des variables pour fournir une cartographie globale et synthétique.

La cartographie multivariée consiste, à partir d'observations ponctuelles pour lesquelles on a mesuré plusieurs variables et relevé leurs coordonnées, à synthétiser par un nombre réduit de nouvelles variables, appelées ici « facteurs », à la fois les valeurs observées sur les emplacements échantillonnés et les dépendances spatiales dues à la proximité, au sens géographique, entre les sites. En tenant compte de la position des sites de prélèvement dans l'espace, on admet implicitement que les sites géographiquement proches ont des valeurs voisines et que les différences, lorsqu'elles existent, ne sont dues qu'aux fluctuations d'échantillonnage. Cette hypothèse aura pour effet de simplifier la visualisation des cartes et de supprimer une grande partie de l'effet de « mitage ».

État de l'art

Quels outils et quelles méthodes statistiques ou géostatistiques sont disponibles pour résoudre cette question?

En première approche, on peut utiliser les outils classiques bien connus tels ceux de l'analyse factorielle : Analyse en Composantes Principales (ACP) ou Analyse Factorielle des Correspondances (AFC) (Cailliez et Pagès, 1976), en liaison avec les techniques de classification comme les Nuées dynamiques (Diday, 1972) ou la Classification Ascendante Hiérarchique (Jambu, 1978). Malheureusement, on obtient souvent une cartographie de type mosaïque où les classes obtenues apparaissent très fragmentées et très dispersées dans l'espace. Ce « mitage » provient de l'absence de prise en compte de l'information spatiale par ces techniques.

Pour cette raison, la dimension spatiale a été introduite de diverses façons. D'abord en modélisant chaque variable en fonction des coordonnées géographiques. Dans un cadre monovarié, il s'agit de découper l'espace géographique par un arbre dichotomique. L'arbre est construit par détection de seuils de coupure, selon les deux axes de coordonnées, de telle manière que l'homogénéité des observations dans chaque unité spatiale soit maximum (Breiman *et al*, 1984). D'autres auteurs (Ambroise *et al*, 1997) proposent d'utiliser l'algorithme « EM », où le problème de classification est formulé en un problème d'optimisation. Ils introduisent l'information spatiale en ajoutant au critère à maximiser une quantité ayant une signification géographique.

Dans une autre optique (analyse des données à la française), certains auteurs ont introduit la contrainte de contiguïté entre les observations. Mesurée par le coefficient de Geary (Lebart, 1969), elle est utilisée dans la Classification Hiérarchique avec des contraintes de voisinages (Lebart, 1978). Ces travaux ont été poursuivis en introduisant des opérateurs de voisinage dans l'analyse des données spatio-temporelles (Méot *et al*, 1993) et avec Chessel et Sabatier (1993) qui proposent l'étude de deux ensembles de variables tenant compte de la proximité des observations mais dans une optique explicative. Enfin Thioulouse *et al* (1995) réalisent des ACP et des AFC qui maximisent les structures globales, locales et totales.

La géostatistique (Matheron, 1962) offre une autre perspective. En général, ces techniques permettent de répondre à des problèmes d'estimation locale ou globale de paramètres mesurés à l'intérieur du domaine géographique $\mathcal{D}$. Le caractère multivarié est pris en compte dans les techniques de cokrigeage. La géostatistique permet ainsi de construire des blocs diagrammes à trois dimensions ou de dessiner des cartes d'isovaleurs. Cependant, ces produits finaux sont monovariés : ils ne cartographient qu'une variable à la fois.

A partir de l'analyse géostatistique et des méthodes de classification, Oliver et Webster (1989) proposent d'appliquer un algorithme de classification non hiérarchique sur une matrice de dissimilarité particulière. Cette procédure considère, non seulement les différences entre les valeurs mesurées sur les variables mais également la structure spatiale par l'intermédiaire des variogrammes modélisés.

Le caractère multivarié et l'information spatiale peuvent aussi être analysés dans une méthodologie utilisant de façon combinée : géostatistique, analyse multivariée et classification (Arnaud et Pichot, 1996). Après découpage de l'espace géographique en cellules et cokrigeage par bloc de ces cellules, les valeurs estimées sont analysées par ACP et classification pour obtenir une cartographie de l'espace géographique. Cette méthodologie n'est pas sans critique en raison, notamment, de la non indépendance spatiale des facteurs de l'ACP (Wackernagel, 1998) utilisés dans la classification.

Plus récemment Allard *et al* (1999) proposent, dans le cas monovarié, une méthodologie de classification où la composante spatiale est prise en compte par l'estimation de la variance intra-groupe lorsque le nombre de groupes est fixé *a priori*.

Dans cet article, nous proposons d'utiliser l'analyse krigeante (Matheron, 1982; Wackernagel, 1998) qui considère simultanément les caractères multivariés et géographiques des observations. Nous verrons comment cette méthode permet de répondre à un problème de classification et comment elle l'enrichit en donnant des

indications sur le caractère plus ou moins régulier du phénomène régionalisé, les différents niveaux d'échelle, la détection d'éventuelles anisotropies...

2. Présentation des méthodes géostatistiques et de l'analyse krigeante

La géostatistique a été développée pour résoudre le problème de l'estimation des variables régionalisées (Matheron, 1962). Une variable régionalisée $z(\boldsymbol{s})$, définie pour $\boldsymbol{s}$ variant dans le domaine $\mathcal{D}$, est interprétée comme une réalisation, un tirage particulier, d'une fonction aléatoire $Z(\boldsymbol{s})$. L'aspect erratique de $z(\boldsymbol{s})$ est pris en compte par un modèle probabiliste, l'aspect régionalisé par la corrélation spatiale ou plus généralement par la loi spatiale de la fonction aléatoire $Z(\boldsymbol{s})$.

On peut généraliser cette approche au cas multivarié : chaque variable régionalisée $z_i(\boldsymbol{s})$, $i = 1, \ldots, N$, est considérée comme une réalisation d'un processus $\{Z_i(\boldsymbol{s}), \boldsymbol{s} \in D\}$. Deux outils, les fonctions de covariance et les variogrammes, reliés entre eux dans certaines conditions, vont rendre possible l'analyse de la structure spatiale *conjointe* des variables $Z_i(\boldsymbol{s})$. L'introduction d'hypothèses de stationnarité va être nécessaire : la stationnarité des deux premiers moments des variables considérées (stationnarité conjointe du second ordre) ou la stationnarité de deux premiers moments des *accroissements* des variables (hypothèse intrinsèque conjointe). Ces hypothèses vont rendre possible l'inférence statistique à partir des valeurs expérimentales pour estimer les fonctions de covariance ou les variogrammes. Leur modélisation se situe, en général, dans le cadre du modèle linéaire de corégionalisation. Celui-ci va permettre de décomposer les fonctions aléatoires en facteurs dont les propriétés de non corrélation spatiale vont donner les moyens de cartographier la zone d'étude.

2.1. Fonctions de covariance et variogrammes : propriétés et inférence

On dit que les fonctions aléatoires Z_i sont **conjointement stationnaires du second ordre** si leurs deux premiers moments (espérance et covariance) existent et sont invariants par translation :

$$\begin{cases} E[Z_i(\boldsymbol{s})] = m_i & \forall i = 1, \ldots, N \text{ et } \forall \boldsymbol{s} \in \mathcal{D} \\ \operatorname{cov}\big[Z_i(\boldsymbol{s}+\boldsymbol{h}),\ Z_j(\boldsymbol{s})\big] = C_{ij}(\boldsymbol{h}) & \forall i, j = 1, \ldots, N \text{ et } \forall \boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h} \in \mathcal{D} \end{cases}$$

Les $C_{ij}(\boldsymbol{h})$ sont les fonctions de covariances simples ($i = j$) et croisées ($i \neq j$). Ces fonctions ne sont pas quelconques et doivent vérifier certaines contraintes mathématiques.

Les fonctions aléatoires Z_i sont **conjointement intrinsèques** si leurs accroissements sont conjointement stationnaires du second ordre. On peut résumer cette condition par les relations suivantes :

$$\begin{cases} E\big[Z_i(\boldsymbol{s}+\boldsymbol{h}) - Z_i(\boldsymbol{s})\big] = 0 & \forall i = 1, \ldots, N \text{ et } \forall \boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h} \in \mathcal{D} \\ \operatorname{cov}\big[Z_i(\boldsymbol{s}+\boldsymbol{h}) - Z_i(\boldsymbol{s}), Z_j(\boldsymbol{s}+\boldsymbol{h}) - Z_j(\boldsymbol{s})\big] & \\ \qquad = 2\gamma_{ij}(\boldsymbol{h}) & \forall i, j = 1, \ldots, N \text{ et } \boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h} \in \mathcal{D} \end{cases}$$

Les $\gamma_{ij}(\boldsymbol{h})$ sont les variogrammes simples ($i = j$) et croisés ($i \neq j$). On montre que, pour tout vecteur $\boldsymbol{h}$, la matrice $\Gamma(\boldsymbol{h}) = [\gamma_{ij}(\boldsymbol{h})]_{i,j=1\ldots N}$ est symétrique et de type positif.

Les fonctions de covariance et les variogrammes simples et croisés fournissent une description élémentaire de la dépendance entre les valeurs prises entre deux points séparés de $\boldsymbol{h}$ par les différentes variables. Ils modélisent la « structure » spatiale du phénomène multivarié étudié.

L'hypothèse stationnaire conjointe implique l'hypothèse intrinsèque conjointe, mais la réciproque n'est pas vraie. On qualifie de strictement intrinsèques des fonctions aléatoires intrinsèques mais non stationnaires. Dans le cas où l'hypothèse stationnaire conjointe est vérifiée, on a la relation :

$$\forall i, j = 1, \ldots, N,\ \forall \boldsymbol{h},\ \gamma_{ij}(\boldsymbol{h}) = C_{ij}(\boldsymbol{0}) - \frac{1}{2}\left[(C_{ij}(-\boldsymbol{h}) + C_{ij}(+\boldsymbol{h})\right]$$

qui montre que le variogramme croisé ne contient que la partie paire de la fonction de covariance croisée, et renferme donc moins d'information structurale. En dépit de cette restriction, on préfère, le plus souvent, utiliser les variogrammes simples et croisés, car ils sont définis dans le cadre intrinsèque, plus général que le cadre stationnaire du second ordre.

A partir des données expérimentales, on peut estimer les fonctions de covariance et les variogrammes. Notons $\boldsymbol{s}_\alpha^i$ ($\alpha = 1 \ldots n_i$) les sites de mesure de la variable z_i. Ces sites peuvent être différents d'une variable à l'autre (cas d'*hétérotopie*, partielle ou totale). La fonction de covariance simple ou croisée $C_{ij}(\boldsymbol{h})$ est habituellement estimée par :

$$\widehat{C}_{ij}(\boldsymbol{h}) = \frac{1}{|N_{ij}(\boldsymbol{h})|} \sum_{N_{ij}(\boldsymbol{h})} \left[Z_i(\boldsymbol{s}_\alpha^i) - \overline{Z}_i\right]\left[Z_j(\boldsymbol{s}_\beta^j) - \overline{Z}_j\right]$$

où $\overline{Z}_i = \dfrac{1}{n_i}\sum_{\alpha=1}^{n_i} Z_i(\boldsymbol{s}_\alpha^i)$ et $\overline{Z}_j = \dfrac{1}{n_j}\sum_{\beta=1}^{n_j} Z_j(\boldsymbol{s}_\beta^j)$

$N_{ij}(\boldsymbol{h}) = \left\{(\alpha, \beta) \text{ tel que } \boldsymbol{s}_\alpha^i - \boldsymbol{s}_\beta^j = \boldsymbol{h}\right\}$

$|N_{ij}(\boldsymbol{h})|$ est le nombre de paires distinctes de l'ensemble $N_{ij}(\boldsymbol{h})$.

Quant au variogramme $\gamma_{ij}(\boldsymbol{h})$, il est estimé par :

$$\widehat{\gamma}_{ij}(\boldsymbol{h}) = \frac{1}{2|N_{ij}(\boldsymbol{h})|} \sum_{N_{ij}(\boldsymbol{h})} \left[Z_i(\boldsymbol{s}_\alpha^i) - Z_i(\boldsymbol{s}_\beta^i)\right]\left[Z_j(\boldsymbol{s}_\alpha^j) - Z_j(\boldsymbol{s}_\beta^j)\right]$$

où $N_{ij}(\boldsymbol{h}) = \{(\alpha, \beta))$ tel que $\boldsymbol{s}_\alpha^i = \boldsymbol{s}_\alpha^j$, $\boldsymbol{s}_\beta^i = \boldsymbol{s}_\beta^j$ et $\boldsymbol{s}_\alpha^i - \boldsymbol{s}_\beta^i = s_\alpha^j - \boldsymbol{s}_\beta^j = \boldsymbol{h}\}$

Le calcul de $\widehat{\gamma}_{ij}(\boldsymbol{h})$ (mais pas celui de $\widehat{C}_{ij}(\boldsymbol{h})$) est impossible en cas d'hétérotopie totale, auquel cas l'ensemble $N_{ij}(\boldsymbol{h})$ est vide.

2.2. Le modèle linéaire de corégionalisation

Les variogrammes (ou covariances) simples et croisés d'un ensemble de variables $Z_1 \dots Z_N$ ne peuvent être modélisés indépendamment, car il existe entre eux des contraintes mathématiques. En pratique, pour satisfaire ces contraintes, on a recours au *modèle linéaire de corégionalisation*, qui permet de faire face à la plupart des situations. On peut en donner une définition dans les deux cas : stationnaire et intrinsèque. L'interprétation de ce modèle sera détaillée dans le paragraphe 2.3 concernant l'analyse krigeante.

N fonctions aléatoires $Z_i(\boldsymbol{s})$ conjointement stationnaires d'ordre deux obéissent au modèle linéaire de corégionalisation lorsque leurs covariances sont des combinaisons linéaires des mêmes covariances de base $\rho_u(\boldsymbol{h})$:

$$\forall i, j = 1, \dots, N,\ \forall \boldsymbol{h},\ C_{ij}(\boldsymbol{h}) = \sum_{u=1}^{S} b_{ij}^u \rho_u(\boldsymbol{h})$$

en notation matricielle : $\boldsymbol{C}(\boldsymbol{h}) = \sum_{u=1}^{S} \boldsymbol{B}_u \rho_u(\boldsymbol{h})$

où $\boldsymbol{C}(\boldsymbol{h})$ est la matrice des covariances : $\boldsymbol{C}(\boldsymbol{h}) = \left[C_{ij}(\boldsymbol{h})\right]_{i,j=1,\dots,N}$

$\rho_u(\boldsymbol{h})$ est une covariance telle que $\rho_u(\boldsymbol{0}) = 1$

$\boldsymbol{B}_u = (b_{ij}^u)$ est appelée matrice de corrégionalisation.

On obtient un modèle cohérent en imposant que chaque matrice $\boldsymbol{B}_u$ soit symétrique et de type positif. En particulier, on aura la relation :

$$\forall i, j = 1 \dots N,\ \forall u = 1 \dots S,\ |b_{ij}^u| \leqslant \sqrt{b_{ii}^u b_{jj}^u} \tag{1}$$

Cette inégalité implique que si, pour un modèle ρ_u quelconque, la valeur b_{ij}^u associée à la covariance croisée est non nulle, les valeurs b_{ii}^u et b_{jj}^u relatives aux deux covariances simples correspondantes sont nécessairement strictement positives.

N fonctions aléatoires $Z_i(\boldsymbol{s})$ conjointement intrinsèques obéissent au modèle linéaire de corégionalisation lorsque les variogrammes sont des combinaisons linéaires des mêmes variogrammes de base $g_u(\boldsymbol{h})$:

$$\forall i, j = 1 \dots N,\ \forall \boldsymbol{h},\ \gamma_{ij}(\boldsymbol{h}) = \sum_{u=1}^{S} b_{ij}^u g_u(\boldsymbol{h})$$

en notation matricielle : $\Gamma(\boldsymbol{h}) = \sum_{u=1}^{S} \boldsymbol{B}_u g_u(\boldsymbol{h})$

où $\Gamma(\boldsymbol{h})$ est la matrice des variogrammes,

$g_u(\boldsymbol{h})$ est un variogramme de base,

$\boldsymbol{B}_u = (b_{ij}^u)$ est appelée matrice de corrégionalisation.

Comme $\Gamma(\boldsymbol{h})$ doit être une matrice symétrique de type positif pour tout $\boldsymbol{h}$, on impose (condition suffisante) que *chacune* des matrices de corégionalisation $\boldsymbol{B}_u$ soit symétrique de type positif.

La recherche des coefficients des matrices $\boldsymbol{B}_u$ se fait par ajustement à partir des covariances ou variogrammes expérimentaux. Dans le cas bivarié, il s'effectue d'abord sur les variogrammes simples puis se poursuit par celui du variogramme croisé en utilisant la relation (1). Dans le cas de plus de deux fonctions aléatoires, Goulard *et al* (1992) proposent une procédure itérative qui fournit des matrices $\boldsymbol{B}_u$ ayant les bonnes propriétés, mais la convergence de l'algorithme n'a pu être démontrée.

2.3. L'analyse krigeante

L'analyse krigeante s'effectue en deux étapes successives : la première étape réalise la décomposition des fonctions aléatoires en une somme de *facteurs* indépendants; la seconde procède à l'estimation, par cokrigeage, de ces facteurs en chaque point $\boldsymbol{s} \in \mathcal{D}$.

2.3.1. Cas d'une seule variable Z(s)

2.3.1.1. Décomposition de la variable Z(s)

a) cas d'une fonction aléatoire stationnaire

Nous allons considérer dans un premier temps une fonction aléatoire $Z(\boldsymbol{s})$ stationnaire d'ordre deux, de moyenne (espérance) m et de fonction de covariance « gigogne », c'est-à-dire composée de plusieurs modèles élémentaires :

$$\forall \mathbf{h} : C(\boldsymbol{h}) = \sum_{u=1}^{S} b_u \rho_u(\boldsymbol{h})$$

où $\forall u = 1 \ldots S,\ b_u > 0$

$\forall u = 1 \ldots S,\ \rho_u(\boldsymbol{h})$ est une covariance telle que $\rho_u(\mathbf{0}) = 1$

On peut décomposer $Z(\boldsymbol{s})$ en une somme de S *composantes spatiales* stationnaires X^u :

$$\forall \boldsymbol{s} \in \mathcal{D},\ Z(\boldsymbol{s}) = \sum_{u=1}^{S} X^u(\boldsymbol{s}) + m$$

• d'espérances nulles : $\forall u = 1 \ldots S, \forall \boldsymbol{s} \in \mathcal{D}, E\big[X^u(\boldsymbol{s})\big] = 0$

• spatialement non corrélées et de fonctions de covariance $b_u \rho_u(\boldsymbol{h})$:

$$\forall u, v = 1 \dots S,\ \forall \boldsymbol{s},\ \boldsymbol{s}+\boldsymbol{h} \in \mathcal{D},\ \operatorname{cov}\left[X^u(\boldsymbol{s}+\boldsymbol{h}), X^v(\boldsymbol{s})\right] = \begin{cases} 0 & \text{si } u \neq v \\ b_u \rho_u(\boldsymbol{h}) & \text{si } u = v \end{cases}$$

La variable $Z(\boldsymbol{s})$ apparaît comme la superposition de composantes « indépendantes » qui agissent à des échelles spatiales différentes (chaque covariance est associée à une dimension caractéristique, donnée par sa *portée*, *i.e.* la distance à partir de laquelle elle est identiquement nulle).

A titre d'exemple, supposons que la fonction aléatoire stationnaire d'ordre deux $Z(\boldsymbol{s})$ admette une covariance gigogne composée de deux modèles sphériques de portées 15 m et 70 m :

$$C(\boldsymbol{h}) = 3C_{\text{sph}}(\boldsymbol{h}, 15\ m) + 5C_{\text{sph}}(h, 70\ m).$$

On rappelle que le modèle de covariance sphérique de portée a s'écrit :

$$C_{\text{sph}}(r, a) = \begin{cases} 1 - \frac{3}{2}\frac{r}{a} + \frac{1}{2}\frac{r^3}{a^3} & \text{pour } 0 \leqslant r \leqslant a \\ 0 & \text{pour } r \geqslant a \end{cases}$$

Il s'agit d'un des modèles de covariance les plus couramment utilisés.

$Z(\boldsymbol{s})$ peut alors se décomposer de la façon suivante :

$$Z(\boldsymbol{s}) = X^1(\boldsymbol{s}) + X^2(\boldsymbol{s}) + m$$

où $X^1(\boldsymbol{s})$ et $X^2(\boldsymbol{s})$ sont deux fonctions aléatoires stationnaires d'ordre deux, sans corrélation spatiale mutuelle, de moyenne nulle et de covariances respectives $C_1(\boldsymbol{h}) = 3\,\text{sph}(\boldsymbol{h}, 15\ m)$ et $C_2(\boldsymbol{h}) = 5\,\text{sph}(\boldsymbol{h}, 70\ m)$.

b) cas d'une fonction aléatoire intrinsèque stricte

Considérons à présent une fonction aléatoire $Z(\boldsymbol{s})$ obéissant au modèle linéaire de corégionalisation avec un variogramme gigogne composé de $S-1$ variogrammes stationnaires (*i.e.* qui se stabilisent à l'infini autour d'une valeur limite appelée *palier*) et d'un variogramme intrinsèque strict (non borné) :

$$\forall \boldsymbol{h},\ \gamma(\boldsymbol{h}) = \sum_{u=1}^{S-1} b_u g_u(\boldsymbol{h}) + b_S g_S(\boldsymbol{h}).$$

$Z(\boldsymbol{s})$ peut être décomposée en une somme de S *composantes spatiales* X^u, où $X^1 \dots X^{S-1}$ sont stationnaires et X^S strictement intrinsèque, de variogrammes respectifs $b_u g_u(\boldsymbol{h})$ et $b_S g_S(\boldsymbol{h})$:

$$\forall \boldsymbol{s} \in \mathcal{D},\ Z(\boldsymbol{s}) = \sum_{u=1}^{S-1} X^u(\boldsymbol{s}) + X^S(\boldsymbol{s}).$$

Les composantes stationnaires X^u et les accroissement de X^S sont d'espérances nulles et sans corrélation spatiale.

Ainsi donc, toute fonction aléatoire $Z(\boldsymbol{s})$ qui admet une covariance ou un variogramme gigogne peut être décomposée en plusieurs composantes spatiales spatialement non corrélées ou à accroissements spatialement non corrélés, agissant à différentes échelles.

2.3.1.2. Estimation des composantes spatiales par krigeage

Lorsqu'on considère le modèle intrinsèque décrit ci-dessus [1], on estime les valeurs des composantes spatiales $X^u(\boldsymbol{s})$ par krigeage à partir des mesures expérimentales $Z(\boldsymbol{s}_\alpha)$. On pose, pour tout $u = 1 \ldots S$:

$$\widehat{X}_u(\boldsymbol{s}) = \sum_\alpha \omega_\alpha^u Z(\boldsymbol{s}_\alpha)$$

où les ω_α^u sont des pondérateurs réels. Ils sont déterminés par le système d'équations suivant :

$$\begin{cases} \displaystyle\sum_\beta \omega_\beta^u \gamma(\boldsymbol{s}_\alpha - \boldsymbol{s}_\beta) - \mu_u = b_u g_u(\boldsymbol{s}_\alpha - \boldsymbol{s}) \text{ pour } \alpha = 1 \ldots n \\ \displaystyle\sum_\beta \omega_\beta^u = \begin{cases} 0 & \text{pour } u = 1 \ldots S-1 \ (X^u \text{ stationnaire}) \\ 1 & \text{pour } u = S \qquad\quad (X^S \text{ intrinsèque stricte}) \end{cases} \end{cases}$$

où μ_u est un multiplicateur de Lagrange.

Ce système diffère du krigeage usuel par l'apparition dans le membre de droite des termes $b_u g_u(\boldsymbol{s}_\alpha - \boldsymbol{s})$ spécifiques de la composante $X^u(\boldsymbol{s})$ et, dans le cas où $X^u(\boldsymbol{s})$ est stationnaire, par la condition sur la somme des poids (Wackernagel,1998).

2.3.2. Décomposition de plusieurs fonctions aléatoires en corrélation intrinsèque

2.3.2.1. Décomposition des variables $Z_i(\boldsymbol{s})$

a) Cas stationnaire du second ordre

N fonctions aléatoires conjointement stationnaires d'ordre deux $Z_1(\boldsymbol{s}) \ldots Z_N(\boldsymbol{s})$, de moyennes respectives $m_1 \ldots m_N$ sont en *corrélation intrinsèque* si leurs fonctions de covariance sont proportionnelles entre elles :

$$\forall i, j = 1 \ldots N, \ \forall \boldsymbol{h}, \ C_{ij}(\boldsymbol{h}) = \sigma_{ij} \rho_0(\boldsymbol{h})$$

où $\rho_0(\boldsymbol{h})$ est une covariance telle que $\rho_0(\boldsymbol{0}) = 1$

$V = (\sigma_{ij})$ est la matrice de variance-covariance des $Z_i(\boldsymbol{s})$

[1] On se limite à une seule structure intrinsèque stricte car l'estimation par krigeage de X^u est impossible pour plus de deux composantes strictement intrinsèques.

On montre que les Z_i peuvent être décomposées en un ensemble de N *facteurs spatiaux* $Y_p(\boldsymbol{s})$:

$$\forall i = 1 \dots N,\ \forall \boldsymbol{s} \in \mathcal{D},\ Z_i(\boldsymbol{s}) = \sum_{p=1}^{N} a_p^i Y_p(\boldsymbol{s}) + m_i \tag{2}$$

- d'espérances nulles : $\forall p = 1 \dots N, \forall \boldsymbol{s} \in \mathcal{D} : E[Y_p(\boldsymbol{s})] = 0$
- spatialement non corrélés et de fonction de covariance $\rho_0(\boldsymbol{h})$:

$$\forall p, q = 1 \dots N,\ \forall \boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h} \in \mathcal{D},\ \operatorname{cov}\big[Y_p(\boldsymbol{s}+\boldsymbol{h}), Y_q(\boldsymbol{s})\big] = \begin{cases} 0 & \text{si } p \neq q \\ \rho_0(\boldsymbol{h}) & \text{si } p = q \end{cases}$$

Les facteurs spatiaux $Y_p(\boldsymbol{s})$ sont définis par la relation :

$$\big(\sqrt{\lambda_1} Y_1(\boldsymbol{s}), \dots, \sqrt{\lambda_N} Y_N(\boldsymbol{s})\big) = \big(Z_1(\boldsymbol{s}) - m_1, \dots, Z_n(\boldsymbol{s}) - m_N\big) \times \boldsymbol{Q}$$

obtenue lorsque l'on effectue l'analyse en composantes principales de la matrice de variance-covariance $\boldsymbol{V}$. On a alors en notation matricielle :

$$\boldsymbol{V} = \boldsymbol{Q} \Lambda \boldsymbol{Q}^t \text{ avec } \boldsymbol{Q}^t \boldsymbol{Q} = \boldsymbol{I} \text{ (matrice identité) et } \Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_N)$$

où $\boldsymbol{Q}$ est une matrice orthogonale de vecteurs propres de $\boldsymbol{V}$ et Λ la matrice diagonale des valeurs propres associées.

Chaque facteur spatial Y_p est associé à une valeur propre positive λ_p de $\boldsymbol{V}$ qui mesure la « quantité » d'information contenue dans ce facteur. La comparaison des valeurs propres permet de classer les facteurs par ordre d'importance.

Lorsque les $Z_i(\boldsymbol{s})$ sont exprimées dans des unités différentes, il est préférable de travailler sur les variables standardisées (centrées et réduites) :

$$\widetilde{Z}_i(\boldsymbol{s}) = \frac{Z_i(\boldsymbol{s}) - m_i}{\sqrt{\sigma_{ii}}}$$

qui deviennent alors comparables (car sans dimensions) et l'analyse en composantes principales est réalisée sur la matrice des corrélations au lieu de celle de variance-covariance.

A titre d'exemple, supposons que les deux fonctions aléatoires stationnaires d'ordre deux, $Z_1(\boldsymbol{s})$ et $Z_2(\boldsymbol{s})$ soient en corrélation intrinsèque avec les covariances simples et croisés suivantes :

$$\begin{aligned} C_{11}(\boldsymbol{h}) &= 4C_{\text{sph}}(\boldsymbol{h}, 70\ m) \\ C_{12}(\boldsymbol{h}) &= \ C_{\text{sph}}(\boldsymbol{h}, 70\ m) \\ C_{22}(\boldsymbol{h}) &= 9C_{\text{sph}}(\boldsymbol{h}, 70\ m) \end{aligned}$$

soit en notation matricielle :

$$\begin{pmatrix} C_{11}(\boldsymbol{h}) & C_{12}(\boldsymbol{h}) \\ C_{21}(\boldsymbol{h}) & C_{22}(\boldsymbol{h}) \end{pmatrix} = \begin{pmatrix} 4 & 1 \\ 1 & 9 \end{pmatrix} C_{\text{sph}}(\boldsymbol{h}, 70\ m).$$

On obtient alors la décomposition des $Z_i(\boldsymbol{s})$ en une somme de facteurs spatiaux $Y_i(\boldsymbol{s})$:

$$\begin{pmatrix} Z_1(\boldsymbol{s}) \\ Z_2(\boldsymbol{s}) \end{pmatrix} = \begin{pmatrix} 1.92 & 0.57 \\ -0.37 & 2.98 \end{pmatrix} \begin{pmatrix} Y_1(\boldsymbol{s}) \\ Y_2(\boldsymbol{s}) \end{pmatrix} + \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}$$

Les valeurs propres associées sont $\lambda_1 = 9.19$ et $\lambda_2 = 3.81$. Le facteur Y_1 contient donc 70 % de l'information totale.

b) Cas intrinsèque

N fonctions aléatoires intrinsèques $Z_1(\boldsymbol{s}) \dots Z_N(\boldsymbol{s})$ sont en **corrélation intrinsèque** si leurs variogrammes sont proportionnels entre eux :

$$\forall i, j = 1 \dots N,\ \forall \boldsymbol{h},\ \gamma_{ij}(\boldsymbol{h}) = p_{ij} g_0(\boldsymbol{h})$$

où $g_0(\boldsymbol{h})$ est un variogramme de base,

$P = (p_{ij})$ est appelée matrice des paliers.

Les $Z_i(\boldsymbol{s})$ peuvent être décomposées en un ensemble de N *facteurs spatiaux* $Y_p(\boldsymbol{s})$:

$$\forall i = 1, \dots, N,\ \forall \boldsymbol{s} \in \mathcal{D},\ Z_i(\boldsymbol{s}) = \sum_{p=1}^{N} a_p^i Y_p(\boldsymbol{s}).$$

Les Y_p ont pour variogramme $g_0(\boldsymbol{h})$; leurs accroissements sont d'espérances nulles et sans corrélation spatiale mutuelle. Ils sont définis par la relation :

$$\forall \boldsymbol{s} \in \mathcal{D},\ \left(\sqrt{\lambda_1} Y_1(\boldsymbol{s}), \dots, \sqrt{\lambda_N} Y_N(\boldsymbol{s})\right) = \left(Z_1(\boldsymbol{s}), \dots, Z_n(\boldsymbol{s})\right) \times \boldsymbol{Q} \qquad (3)$$

où $\boldsymbol{Q}$ est une matrice orthogonale de vecteurs propres de $\boldsymbol{P}$ et $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_N)$ la matrice diagonale de valeurs propres associées :

$$\boldsymbol{P} = \boldsymbol{Q} \Lambda \boldsymbol{Q}^t \text{ avec } \boldsymbol{Q}^t \boldsymbol{Q} = \boldsymbol{I}.$$

Comme dans le cas stationnaire, la comparaison des valeurs propres $\lambda_1, \dots, \lambda_N$ permet de hiérarchiser les facteurs par ordre d'importance.

2.3.2.2. Estimation des facteurs spatiaux $Y_p(\boldsymbol{s})$ en un point $\boldsymbol{s} \in \mathcal{D}$

a) si le point $\boldsymbol{s}$ coïncide avec l'un des points de données $\boldsymbol{s}_\alpha$ où toutes les variables Z_j sont connues, alors $Y_p(\boldsymbol{s})$ sera évalué par l'une des relations (2) ou (3) (celle correspondant à l'hypothèse adoptée : stationnaire d'ordre deux ou intrinsèque).

b) si $\boldsymbol{s}$ n'est pas un point de données, on estimera $Y_p(\boldsymbol{s})$ par cokrigeage à partir des $Z_j(\boldsymbol{s}_\alpha)$, ou, ce qui revient au même, à partir des $Y_j(\boldsymbol{s}_\alpha)$ précédemment calculés. Or les facteurs $Y_j(\boldsymbol{s})$ étant en corrélation intrinsèque, leur cokrigeage, en cas d'homotopie, se réduit à leur krigeage séparé, ce qui simplifie considérablement les calculs.

2.3.3. Cas de plusieurs fonctions aléatoires obéissant au modèle linéaire de corégionalisation

2.3.3.1. Décomposition des N variables $Z_i(\boldsymbol{s})$

a) Cas stationnaire du second ordre : dans un premier temps, en appliquant les résultats du paragraphe 2.3.1 et en supposant que les $Z_i(\boldsymbol{s})$ obéissent au modèle linéaire de corégionalisation, on peut décomposer chaque fonction aléatoire $Z_i(\boldsymbol{s})$ en un ensemble de composantes spatiales $\{X_i^u(\boldsymbol{s})\}_{u=1\ldots S}$:

$$\forall i = 1 \ldots N,\ \forall \boldsymbol{s} \in \mathcal{D}, Z_i(\boldsymbol{s}) = \sum_{u=1}^{S} X_i^u(\boldsymbol{s}) + m_i$$

Les $X_i^u(\boldsymbol{s})$ sont spatialement non corrélées si $u \neq v$; si $u = v$, leur covariance est donnée par la matrice $\boldsymbol{B}_u \rho_u(\boldsymbol{h})$.

Dans un deuxième temps, comme les composantes spatiales $\{X_i^u(\boldsymbol{s})\}_{i=1\ldots N}$ sont en corrélation intrinsèque, elles peuvent être à leur tour décomposées en **facteurs spatiaux** $Y_p^u(\boldsymbol{s})$ non corrélés spatialement (paragraphe 2.3.2). On obtient alors le résultat général suivant :

N fonctions aléatoires conjointement stationnaires d'ordre deux $Z_1(\boldsymbol{s}) \ldots Z_N(\boldsymbol{s})$, de moyennes respectives $m_1 \ldots m_N$ et de covariances de la forme :

$$\forall i, j = 1 \ldots N,\ \forall \boldsymbol{h},\ C_{ij}(\boldsymbol{h}) = \sum_{u=1}^{S} b_{ij}^u \rho_u(\boldsymbol{h})$$

peuvent être décomposées en un ensemble de facteurs spatiaux $Y_p^u(\boldsymbol{s})$:

$$\forall i = 1 \ldots N,\ \forall \boldsymbol{s} \in \mathcal{D},\ Z_i(\boldsymbol{s}) = \sum_{u=1}^{S} \sum_{p=1}^{N} a_{up}^i Y_p^u(\boldsymbol{s}) + m_i$$

• d'espérances nulles :

$$\forall u = 1 \ldots S,\ \forall p = 1 \ldots N,\ \forall \boldsymbol{s} \in \mathcal{D},\ E\big[Y_p^u(\boldsymbol{s})\big] = 0$$

• spatialement non corrélés :

$$\forall p, q = 1 \dots N,\ \forall u, v = 1 \dots S,\ p \neq q \text{ ou } u \neq v,\ \forall \boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h} \in \mathcal{D},$$
$$\operatorname{cov}\left[Y_p^u(\boldsymbol{s}+\boldsymbol{h}),\ Y_q^v(\boldsymbol{s})\right] = 0$$

• de fonction de covariance $\rho_u(\boldsymbol{h})$:

$$\forall p = 1 \dots N,\ \forall u = 1 \dots S,\ \forall \boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h} \in \mathcal{D},\ \operatorname{cov}\left[Y_p^u(\boldsymbol{s}+\boldsymbol{h}), Y_p^u(\boldsymbol{s})\right] = \rho_u(\boldsymbol{h})$$

Les facteurs spatiaux $Y_p^u(\boldsymbol{s})$ sont définis par la relation :

$$\forall \boldsymbol{s} \in \mathcal{D} : \left(\sqrt{\lambda_1^u} Y_1^u(\boldsymbol{s}), \dots, \sqrt{\lambda_N^u} Y_N^u(\boldsymbol{s})\right) = \left(X_1^u(\boldsymbol{s}) - m_1, \dots, X_N^u(\boldsymbol{s}) - m_N\right) \times \boldsymbol{Q}_u \quad (4)$$

où $\boldsymbol{Q}_u$ est une matrice orthogonale de vecteurs propres de $\boldsymbol{B}_u$ et $\Lambda_u = \operatorname{diag}(\lambda_1^u, \dots, \lambda_N^u)$ la matrice diagonale de valeurs propres associées :

$$\boldsymbol{B}_u = \boldsymbol{Q}_u \Lambda_u \boldsymbol{Q}_u^t \text{ avec } \boldsymbol{Q}_u^t \boldsymbol{Q}_u = \boldsymbol{I}.$$

A titre d'exemple, supposons que les deux fonctions aléatoires stationnaires d'ordre deux, $Z_1(\boldsymbol{s})$ et $Z_2(\boldsymbol{s})$ admettent les covariances gigognes simples et croisées :

$$\begin{aligned} C_{11}(\boldsymbol{h}) &= 4\,\operatorname{sph}(\boldsymbol{h}, 5\ m) + 2\,\operatorname{sph}(\boldsymbol{h}, 50\ m) + 15\,\operatorname{sph}(\boldsymbol{h}, 500\ m) \\ C_{12}(\boldsymbol{h}) &= \operatorname{sph}(\boldsymbol{h}, 5\ m) - 3\,\operatorname{sph}(\boldsymbol{h}, 50\ m) + 16\,\operatorname{sph}(\boldsymbol{h}, 500\ m) \\ C_{22}(\boldsymbol{h}) &= 9\,\operatorname{sph}(\boldsymbol{h}, 5\ m) + 10\,\operatorname{sph}(\boldsymbol{h}, 50\ m) + 25\,\operatorname{sph}(\boldsymbol{h}, 500\ m) \end{aligned}$$

soit, en notation matricielle :

$$\begin{pmatrix} C_{11}(\boldsymbol{h}) & C_{12}(\boldsymbol{h}) \\ C_{21}(\boldsymbol{h}) & C_{22}(\boldsymbol{h}) \end{pmatrix} = \begin{pmatrix} 4 & 1 \\ 1 & 9 \end{pmatrix} \operatorname{sph}(\boldsymbol{h}, 5\,m) + \begin{pmatrix} 2 & -3 \\ -3 & 10 \end{pmatrix} \operatorname{sph}(\boldsymbol{h}, 50\,m) + \begin{pmatrix} 15 & 16 \\ 16 & 25 \end{pmatrix} \operatorname{sph}(\boldsymbol{h}, 500\,m).$$

On obtient alors la décomposition des $Z_i(\boldsymbol{s})$ en une somme de facteurs spatiaux $Y_i(\boldsymbol{s})$ que l'on peut écrire matriciellement :

$$\begin{pmatrix} Z_1(\boldsymbol{s}) \\ Z_2(\boldsymbol{s}) \end{pmatrix} = \begin{pmatrix} 1.92 & 0.57 \\ -0.37 & 2.98 \end{pmatrix} \begin{pmatrix} Y_1^1(\boldsymbol{s}) \\ Y_1^2(\boldsymbol{s}) \end{pmatrix} + \begin{pmatrix} 0.95 & -1.05 \\ 0.32 & 3.15 \end{pmatrix} \begin{pmatrix} Y_1^2(\boldsymbol{s}) \\ Y_2^2(\boldsymbol{s}) \end{pmatrix} + \begin{pmatrix} 1.45 & 3.59 \\ -1.07 & 4.89 \end{pmatrix} \begin{pmatrix} Y_1^3(\boldsymbol{s}) \\ Y_2^3(\boldsymbol{s}) \end{pmatrix} + \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}$$

b) Cas intrinsèque : le même raisonnement peut être effectué pour N fonctions aléatoires intrinsèques $Z_i(\boldsymbol{s})$, obéissant au modèle linéaire de corégionalisation :

$$\forall i,j = 1 \dots N,\ \forall \boldsymbol{h},\ \gamma_{ij}(\boldsymbol{h}) = \sum_{u=1}^{S-1} b_{ij}^u g_u(\boldsymbol{h}) + b_{ij}^S g_S(\boldsymbol{h})$$

où

• $g_u(\boldsymbol{h})$, $u = 1 \dots S-1$, sont des variogrammes stationnaires (bornés) de palier unité, *i.e.* tels que $g_u(\boldsymbol{h}) \xrightarrow[|\boldsymbol{h}|\longrightarrow\infty]{} 1$;

• $g_S(\boldsymbol{h})$ est un variogramme strictement intrinsèque (non borné).

Les $Z_i(\boldsymbol{s})$ peuvent être décomposées en un ensemble de facteurs spatiaux $Y_p^u(\boldsymbol{s})$:

$$\forall i = 1 \dots N,\ \forall \boldsymbol{s} \in \mathcal{D},\ Z_i(\boldsymbol{s}) = \sum_{u=1}^{S} \sum_{p=1}^{N} a_{up}^i Y_p^u(\boldsymbol{s})$$

où les facteurs stationnaires ($u = 1 \dots S-1$) et les accroissements des facteurs intrinsèques ($u = S$) sont d'espérances nulles et spatialement non corrélés.

Les facteurs ainsi définis ne contiennent pas d'information redondante. A chaque facteur $Y_p^u(\boldsymbol{s})$ est associée une valeur propre λ_p^u de la matrice de corégionalisation $\boldsymbol{B}_u$ correspondante, qui mesure son importance. Il faut remarquer que la matrice $\boldsymbol{B}_S$, correspond à une structure intrinsèque stricte et n'a pas le même sens que les autres; les facteurs non stationnaires doivent être privilégiés par rapport aux facteurs stationnaires.

2.3.3.2. Estimation des facteurs spatiaux $Y_p^u(\boldsymbol{s})$

Quel que soit le point $\boldsymbol{s} \in \mathcal{D}$, on estime la valeur des facteurs spatiaux $Y_p^u(\boldsymbol{s})$ par cokrigeage à partir des mesures expérimentales. Pour chaque fonction Z_i, on utilise les sites voisins échantillonnés les plus proches de $\boldsymbol{s}$ et l'estimateur du facteur $Y_p^u(\boldsymbol{s})$ en un point $\boldsymbol{s} \in \mathcal{D}$ est de la forme :

$$\widehat{Y}_p^u(\boldsymbol{s}) = \sum_{i=1}^{N} \sum_{\alpha=1}^{n_i} Z_i(\boldsymbol{s}_i^\alpha).$$

Dans le cas où toutes les variables sont stationnaires, de moyennes m_i inconnues, on utilise les fonctions de covariance et le système de cokrigeage s'écrit :

$$\begin{cases} \displaystyle\sum_{j=1}^{N} \sum_{\beta=1}^{n_j} \omega_\beta^j C_{ij}(\boldsymbol{s}_\alpha^i - \boldsymbol{s}_\beta^j) + \mu_i = a_{up}^i \rho_u(\boldsymbol{s}_\alpha^i - \boldsymbol{s}) & \text{pour } i = 1 \dots N,\ \alpha = 1 \dots n_i \\ \displaystyle\sum_{\beta=1}^{n_j} \omega_\beta^i = 0 & \text{pour } i = 1 \dots N \end{cases}$$

Dans le cas plus général où les Z_i vérifient l'hypothèse intrinsèque conjointe avec une seule structure strictement intrinsèque (on ne peut réaliser le cokrigeage des facteurs dans le cas de plus d'une structure non stationnaire), il est également possible d'estimer par cokrigeage les différents facteurs spatiaux. On obtient le système de cokrigeage :

$$\begin{cases} \displaystyle\sum_{j=1}^{N}\sum_{\beta=1}^{n_j} \omega_\beta^j \gamma_{ij}(\boldsymbol{s}_\alpha^i - \boldsymbol{s}_\beta^j) - \mu_i = a_{up}^i g_u(\boldsymbol{s}_\alpha^i - \boldsymbol{s}) \text{ pour } i = 1 \ldots N, \alpha = 1 \ldots n_i \\ \displaystyle\sum_{i=1}^{N} a_{uq}^i \sum_{\beta=1}^{n_j} \omega_\beta^i = \begin{cases} 0 & \text{pour } u = 1 \ldots S-1 \ (Y_p^u \text{ stationnaire}) \\ \delta_p^q & \text{pour } u = S \ (Y_p^S \text{ intrinsèque strict}) \end{cases} \text{ pour } q = 1 \ldots N \end{cases}$$

où δ_p^q est le symbole de Kronecker.

2.3.4. Conclusion sur l'analyse krigeante : atouts et limitations

Le lecteur aura pu noter des similitudes entre l'analyse krigeante et l'analyse en composantes principales (ACP). Nous allons les préciser. L'ACP a pour but la recherche des vecteurs propres de la matrice de variance-covariance des variables Z_i indépendamment de la position des sites $(\boldsymbol{s}_i)_{i=1\ldots N}$ d'observation dans l'espace géographique. Elle fournit des facteurs $Y_p(\boldsymbol{s})$ qui, s'ils ont la propriété d'être orthogonaux point à point :

$$\forall \boldsymbol{s} \in \mathcal{D},\ \mathrm{cov}\big[Y_p(\boldsymbol{s}), Y_q(\boldsymbol{s})\big] = 0 \text{ si } p \neq q$$

ne sont en général pas spatialement décorrélés (Wackernagel, 1998) :

$$\forall \boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h} \in \mathcal{D},\ \mathrm{cov}\big[Y_p(\boldsymbol{s}+\boldsymbol{h}), Y_q(\boldsymbol{s})\big] \neq 0 \text{ si } p \neq q.$$

En revanche, l'analyse krigeante, dans le cadre de l'hypothèse de stationnarité d'ordre deux, décompose la matrice de variance-covariance $\boldsymbol{V}$ en une somme de S matrices de corégionalisation ($\boldsymbol{V} = \displaystyle\sum_{u=1}^{S} \boldsymbol{B}_u$) et les analyse séparément pour extraire de chacune d'elles des facteurs qui, eux, sont spatialement non corrélés :

$$\forall\, u, v = 1 \ldots S,\ \forall \boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h} \in \mathcal{D},\ \mathrm{cov}\big[Y_p^u(\boldsymbol{s}+\boldsymbol{h}), Y_q^v(\boldsymbol{s})\big] = 0 \text{ si } p \neq q \text{ ou } u \neq v.$$

En outre, l'analyse krigeante peut se réaliser sous l'hypothèse intrinsèque stricte, alors que la matrice de variance-covariance n'existe pas (les variances sont infinies). L'analyse krigeante apparaît donc plus riche que l'ACP car elle appréhende simultanément les caractéristiques structurales de la corégionalisation et le caractère multivarié et spatialisé des observations.

L'estimation par cokrigeage des facteurs spatiaux en chaque point de l'espace $\mathcal{D}$ va fournir les éléments pour établir leur cartographie.

Les cartes obtenues pour une structure particulière correspondent à une échelle déterminée, donnée par exemple par la portée de la fonction de covariance. On peut, par exemple, ne s'intéresser qu'aux variations du phénomène à grande distance et vouloir mettre en évidence les éventuelles dérives (« tendances ») dans la zone étudiée. On a la possibilité également de choisir le rang du facteur dans la structure. Comme en ACP, il est caractérisé par une valeur propre et par conséquent son taux d'inertie donne une mesure de la quantité d'information qu'il contient.

La cartographie d'un facteur spatial synthétise l'information multivariée et s'appuie sur la structure spatiale du phénomène en intégrant le contenu de l'information des points de données voisins. A partir des facteurs spatiaux les plus significatifs (ceux associés aux plus grandes valeurs propres), on peut appliquer des algorithmes de classification qui proposeront des zones homogènes.

Néanmoins, une des limitations de l'analyse krigeante repose sur le relatif arbitraire concernant le choix des structures (variogrammes ou covariances) que l'on identifie au cours de l'analyse variographique. En effet, les facteurs spatiaux dérivent directement de ces structures. A partir d'un ensemble de données l'ajustement d'un modèle est une opération subjective. Pour réduire cet arbitraire, il convient d'intégrer les informations qualitatives que l'expérimentateur est à même de connaître (genèse du phénomène, différents processus en synergie ou en opposition, fiabilité et erreurs des mesures).

3. Exemple sur des données de sol

Nous allons illustrer les concepts de l'analyse krigeante sur des données de sol. Il s'agit de l'analyse de 110 prélèvements de terre à trois niveaux (0-20 cm, 40-60 cm et 80-100 cm) faits à la tarière, répartis sur une grille régulière carrée de six mètres de côté. Pour étudier la variabilité à courte distance, trois croix de sondage ont été implantées avec des sites distants de un mètre. La figure 1 localise l'implantation des prélèvements dans la zone étudiée. La zone étudiée fait partie d'une plaine peu large polygénique parcourue d'un cours d'eau cadré de reliefs calcaires d'âge secondaire. Elle présente une topographie plane, faiblement inclinée vers la droite où se trouve le cours d'eau. Diverses variables pédologiques ont été mesurées mais nous ne retiendrons pour l'exemple que le pourcentage de terre fine (fraction de taille < 2 mm), le pourcentage d'argile + limon et l'humidité au point de flétrissement permanent (1.6 MPa), par convention, humidité retenue à une tension telle qu'elle est indisponible pour les plantes. La structure conjointe des trois variables étudiées a été ajustée selon le modèle linéaire de corégionalisation, à l'aide de trois structures de base : un effet de pépite, un schéma exponentiel et un modèle puissance. Les deux dernières structures présentent une anisotropie géométrique, dont les directions principales sont orientées de 40° et $-50°$ par rapport à l'axe est-ouest. On trouve les différents paramètres dans le tableau 1, les variogrammes expérimentaux et leurs modèles dans la figure 2.

L'analyse krigeante est menée en trois étapes :

• La première étape consiste à estimer sur une grille régulière que l'on détermine *a priori* les composantes spatiales $X_i^u(\boldsymbol{s})$ associées à chaque variable $(i = 1, 2, 3)$ et à chaque structure $(u = 1, 2, 3)$.

• On calcule ensuite les matrices $\boldsymbol{Q}_u$ et Λ_u permettant de passer des composantes spatiales $X_i^u(\boldsymbol{s})$ aux facteurs spatiaux $Y_u^p(\boldsymbol{s})$. En règle générale, les variables

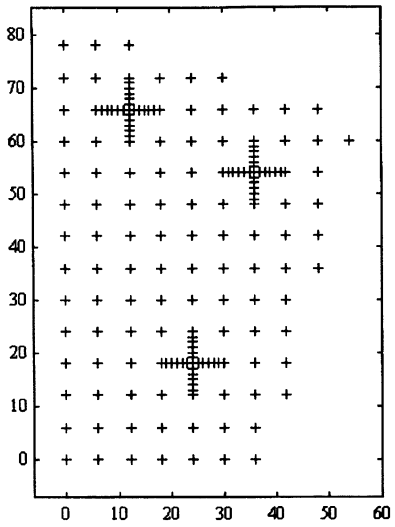

FIGURE 1
Carte d'implantation des sites échantillonnés (axes gradués en mètres)

TABLEAU 1
Matrices des corégionalisarions pour les trois schémas respectifs

Structure	Matrice de corégionalisation
Structure 1 : Effet de pépite $\gamma(\boldsymbol{h}) = \begin{cases} 0 & \text{si } \lvert\boldsymbol{h}\rvert = 0 \\ 1 & \text{sinon} \end{cases}$	$\boldsymbol{B}_1 = \begin{pmatrix} 20.00 & 15.00 & 1.56 \\ 15.00 & 13.00 & 1.00 \\ 1.56 & 1.00 & 0.33 \end{pmatrix}$
Structure 2 : Schéma exponentiel $\gamma(\boldsymbol{h}) = 1 - \exp(-3\lvert\boldsymbol{h}\rvert/a)$ portée pratique : $a = 40$ m (direction d'angle +40°) $a = 25$ m (direction d'angle $-50°$)	$\boldsymbol{B}_2 = \begin{pmatrix} 5.00 & 10.00 & 1.50 \\ 10.00 & 20.00 & 3.00 \\ 1.50 & 3.00 & 0.62 \end{pmatrix}$
Structure 3 : Schéma puissance $\gamma(\boldsymbol{h}) = (\lvert\boldsymbol{h}\rvert/a)^{1.8}$ facteurs d'échelle : $a = 5$ (direction d'angle +40°) $a = 1$ (direction d'angle $-50°$)	$\boldsymbol{B}_3 = \begin{pmatrix} 0.1100 & 0.0500 & -0.0160 \\ 0.0500 & 0.0240 & -0.0080 \\ -0.0160 & -0.0080 & 0.0032 \end{pmatrix}$

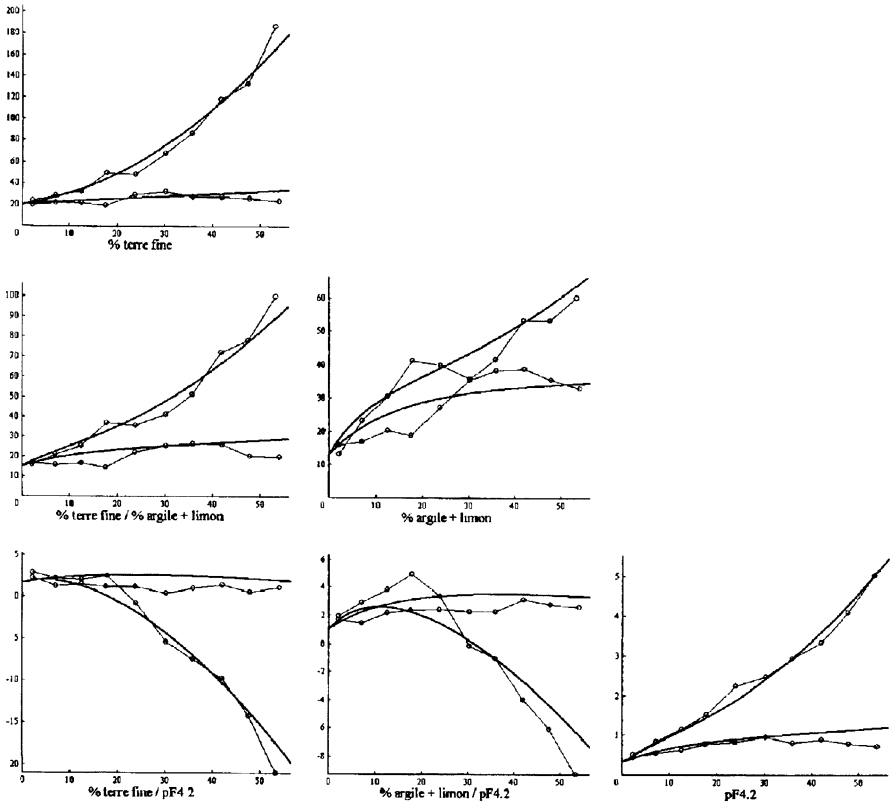

FIGURE 2
Variogrammes expérimentaux simples (diagonale) et croisés et modèles associés selon les deux directions d'anisotropie pour les trois variables analysées (en abscisse : la séparation en mètres; en ordonnée : les variogrammes simples et croisés)

sont standardisées lorsqu'elles sont d'unités différentes, c'est-à-dire qu'on les a divisées par leur écart-type. Si dans le cas stationnaire, les variances *a priori* sont finies, il n'en est pas de même en présence d'une structure strictement intrinsèque comme c'est le cas ici du schéma puissance. Les variances a priori sont alors infinies. Il convient donc de trouver d'autres coefficients de standardisation. Ils peuvent être fournis par les coefficients diagonaux de la matrice de corégionalisation $\boldsymbol{B}_3$ relative au schéma puissance, qui indiquent le degré d'« infinité » des variances *a priori*. Cela revient en fait à effectuer l'analyse krigeante avec les variables non standardisés, mais en recherchant des facteurs spatiaux orthogonaux par rapport à la métrique

$$\boldsymbol{M} = \begin{pmatrix} b_{11}^3 & 0 & 0 \\ 0 & b_{22}^3 & 0 \\ 0 & 0 & b_{33}^3 \end{pmatrix} = \begin{pmatrix} 0.1100 & 0 & 0 \\ 0 & 0.0240 & 0 \\ 0 & 0 & 0.0032 \end{pmatrix}$$

Les matrices $\boldsymbol{Q}_u$ et Λ_u cherchées sont alors solutions du problème de valeurs propres généralisé :

$$\boldsymbol{Q}_u^t \boldsymbol{B}_u \boldsymbol{Q}_u = \Lambda_u \text{ avec } \boldsymbol{Q}_u^t \boldsymbol{M} \boldsymbol{Q}_u = \boldsymbol{I}.$$

Dans le tableau 2, on trouve les matrices de vecteurs et valeurs propres relatives à la diagonalisation des matrices de corégionalisation.

TABLEAU 2

Matrices des vecteurs et valeurs propres pour les trois schémas respectifs

Structure	Diagonalisation	
	Matrice des vecteurs propres	Matrice des valeurs propres
Structure 1 : effet de pépite	$\boldsymbol{Q}_1 = \begin{pmatrix} -1.45 & -2.61 & 0.42 \\ -5.49 & 2.72 & -2.04 \\ -3.83 & 4.76 & 16.59 \end{pmatrix}$	$\Lambda_1 = \begin{pmatrix} 735.7 & 0 & 0 \\ 0 & 14.1 & 0 \\ 0 & 0 & 76.8 \end{pmatrix}$
Structure 2 : Schéma exponentiel	$\boldsymbol{Q}_2 = \begin{pmatrix} 2.94 & -0.63 & -0.27 \\ -1.47 & -5.79 & -2.44 \\ 0.00 & -6.87 & 16.29 \end{pmatrix}$	$\Lambda_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1027 & 0 \\ 0 & 0 & 45.5 \end{pmatrix}$
Structure 3 : Schéma puissance	$\boldsymbol{Q}_3 = \begin{pmatrix} 1.74 & -1.76 & -1.72 \\ 3.81 & 5.05 & -1.30 \\ -9.98 & 3.90 & -14.1 \end{pmatrix}$	$\Lambda_2 = \begin{pmatrix} 2.827 & 0 & 0 \\ 0 & 0.017 & 0 \\ 0 & 0 & 0.156 \end{pmatrix}$

• La troisième étape est l'estimation des facteurs spatiaux $\widehat{Y}_1^u(\boldsymbol{s})$ $(p = 1, 2, 3)$ aux nœuds de la grille à partir de la relation

$$\left(\sqrt{\lambda_1^u}\widehat{Y}_1^u(\boldsymbol{s}), \sqrt{\lambda_2^u}\widehat{Y}_2^u(\boldsymbol{s}), \sqrt{\lambda_3^u}\widehat{Y}_3^u(\boldsymbol{s})\right) = \left(\widehat{X}_1^u(\boldsymbol{s}), \widehat{X}_2^u(\boldsymbol{s}), \widehat{X}_3^u(\boldsymbol{s})\right) \times \boldsymbol{Q}_u$$

où les λ_i^u sont les éléments diagonaux de Λ_u.

Pour chaque structure, on obtient trois facteurs spatiaux, hiérarchisés selon les valeurs propres associées. Nous n'allons retenir que le premier facteur des structures relatives aux modèles exponentiel et puissance (les facteurs relatifs à l'effet de pépite n'ont aucun intérêt car ils sont nuls partout sauf aux points d'observations). La figure 3 visualise ces facteurs.

Le premier facteur relatif au schéma puissance, qui correspond à une structure intrinsèque stricte, affiche une nette dérive dans la direction nord-ouest/sud-est. En quelque sorte, il prend en charge la non-stationnarité du phénomène. Quant au premier facteur du schéma exponentiel, il peut être vu comme un résumé de la structure sous-jacente, lorsque le phénomène est débarrassé de sa dérive. Il correspond, par construction, à une structure stationnaire dont la portée s'étend entre 25 et 40 mètres selon les deux directions principales d'anisotropie.

Il est intéressant d'interpréter les cartes obtenues en termes pédologiques pour s'assurer si elles confirment bien les connaissances qualitatives que l'on a sur la zone

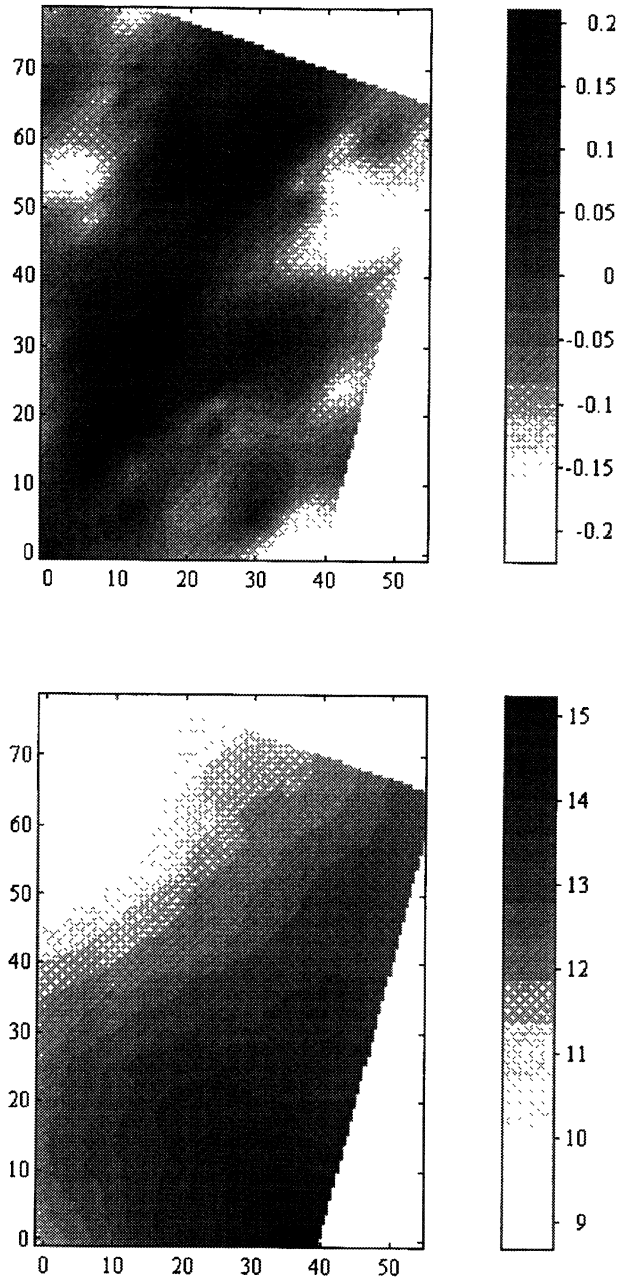

FIGURE 3
Visualisation du premier facteur relatif au modèle exponentiel (en haut), et au modèle puissance (en bas) (axes gradués en mètres)

étudiée ou si elles sont cohérentes avec d'autres informations que l'on a pu obtenir par d'autres méthodes. A cet effet, simultanément au travail d'échantillonnage et d'analyses au laboratoire, une carte pédologique sommaire a été établie sur cette même zone lors de la réalisation des sondages à partir d'observations qualitatives visuelles et tactiles comme la couleur, la texture de la terre fine de la terre extraite et de sa charge en éléments grossiers. Ce travail de terrain a permis d'identifier quatre unités pédologiques (figure 4), à savoir :

- colluvions ($\boldsymbol{C}$) matériaux **assez argileux à charge grossière** pouvant être importante venant du versant adjacent (à gauche sur la carte) sur toute la profondeur du sondage (110 cm environ);

- colluvions sur tuf calcaire ($\boldsymbol{C/T}$) dépôt lacustre à texture **limono-argileuse**, riche en **calcaire**, du pleistocène ancien;

- colluvions sur limon ($\boldsymbol{C/L}$);

- limon ($\boldsymbol{L}$) alluvions récentes essentiellement formées de **limons**.

On peut mettre en parallèle la carte pédologique avec celle du premier facteur de la structure exponentielle. Visuellement, on retrouve, sur cette dernière, les trois unités « colluvions sur tuf » ($\boldsymbol{C/T}$), « colluvions » ($\boldsymbol{C}$) et « limon » ($\boldsymbol{L}$). Les cartes des facteurs sont donc en cohérence avec des études plus qualitatives et confirment le bien fondé de la technique.

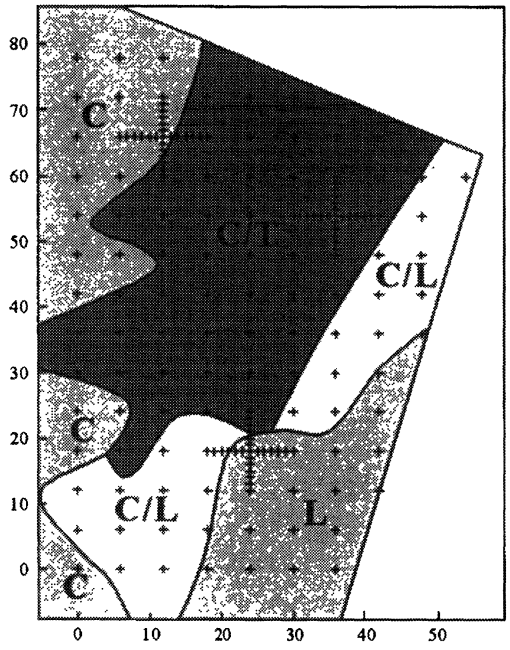

FIGURE 4
Carte pédologique
(axes gradués en mètres)

Conclusion

L'analyse krigeante a permis de prendre en compte le caractère spatial et multivarié des observations pour cartographier une zone géographique. Elle fournit des facteurs non corrélés spatialement et hiérarchisés en fonction de la quantité d'information qu'ils contiennent. Chaque facteur est associé à une échelle spatiale particulière. On peut ainsi soit choisir une échelle d'observation et cartographier les facteurs correspondants. On peut aussi combiner l'analyse des facteurs les plus significatifs, associés à des échelles différentes, et appliquer un algorithme de classification pour aboutir à une cartographie de la zone.

Bibliographie

ALLARD, D. & MONESTIEZ, P. *Classification par critère de variance en contexte spatial.* XXXI^{èmes} Journées de Statistiques, 17-21 mai 1999, Grenoble, France, pp. 945-948.

AMBROISE, C., DANG, M. & GOVAERT, G. *Clustering of spatial data by the EM algorithm.* In A. Soares *et al* (eds.), geoENV I - Geostatistics for Environmental Applications, Kluwer Academic Publishers, Dordrecht, 1997, pp. 493-504.

ARNAUD, M. & PICHOT, J. P. Poster : *A methodology to build a classification of spatialised and multivariate data.* Third International Conference/Workshop on Integrating GIS and Environmental Modeling. January 21-25, 1996. Santa FE, New Mexico, USA.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN R. A., STONE, C. H., *Classification and Regression Trees.* 1984 The Wadsworth Statistics / Probabilities Series. Monterey California. 358 p.

CAILLIEZ, F. & PAGES, J.-P. *Introduction à l'analyse des données*, 1976 (616p). SMASH.

CHESSEL, D. & SABATIER, R. *Couplage de triplets statistiques et graphes de voisinage.* Biométrie et analyse de données spatio-temporelles (B. Asselain et Coll., eds). Société Française de Biométrie, ENSA, Rennes, 1993, pp. 28-37.

DIDAY, E. *Une nouvelle méthode en classification automatique et reconnaissance des formes. La méthode des nuées dynamiques.* Analyse des données en architecture et urbanisme. centre de méthodologie mathématique et informatique. Institut de l'environnement. Nov. 1972, pp. 97-111.

GOULARD, M. & VOLTZ, M. *Linear Coregionalization Model : Tools for Estimation and Choice of Cross-Variogram Matrix.* Mathematical Geology, Vol 24, N° 3, 1992, pp. 269- 286.

JAMBU, M. *Classification automatique pour l'analyse des données 1- méthodes et algorithmes*, Dunod, 1978, 310 p.

LEBART, L. *Analyse statistique de la contiguïté.* Publication de l'ISUP, XVIII, 1969, pp. 81-112.

LEBART, L. *Programme d'agrégation avec contraintes*. Les cahiers de l'analyse des données. Dunod. Vol III, 1978, pp. 275-287.

MATHERON, G. *Traité de géostatistique appliquée*. 1962, Technip, Paris

MATHERON G. *Pour une Analyse Krigeante des données Régionalisées*. Publication N-732, Centre de Géostatistique, Fontainebleau, France, 1982, 22 p.

MÉOT, A., CHESSEL, D. & SABATIER, R. *Opérateurs de voisinage et analyse des données spatio-temporelles*. In : Biométrie et environnement (Lebreton, J.-D. & Asselin B. Eds), 1993, Masson, Paris. pp. 45-71.

OLIVER, M. A. WEBSTER, R. *A geostatistical Basis for Spatial Weighting in Multivariate Classification*. Mathematical Geology, 1989, Vol. 21, N° 1.

THIOULOUSE, J. CHESSEL, D. & CHAMPELY, S. *Multivariate analysis of spatial patterns : a unified approach to local and global structures*. Environmental and Ecological Statistics 2, pp. 1-14, 1995.

WACKERNAGEL, H. Cours de géostatistique multivariable. (4ème édit.), 1993, Ecole des Mines de Paris, Fontainebleau (France) [C-146], 80 p.

WACKERNAGEL, H., Multivariate Geostatistics : an introduction with applications, Springer-Verlag, Berlin, 1998, 291 p.

WACKERNAGEL, H., SANGUINETTI, H. Gold prospecting with factorial cokriging in Limousin, France. In J.C. Davis & U. Herzfeld, Computers in Geology : 25 years of progress, O.U.P., Oxford, 1993, 33-43 (Studies in Math. Geol. 5).