

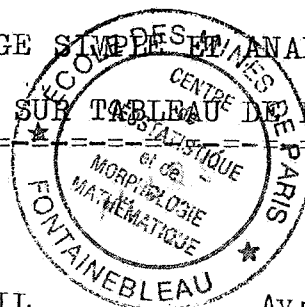
Fontainebleau

N-325

KRIGEAGE SIMPLE ET ANALYSE DE
DONNEES SUR TABLEAU DE DISTANCE

J.P. ORFEUIL

Avril 1973



Ayant constaté une analogie formelle entre la structure du système du Krigeage simple et une formule d'Analyse Factorielle dite "Formule de transition", nous avons essayé de préciser cette analogie. L'idée est la suivante : soit $\{Z_i\}_{i=1,n}$ une réalisation d'une fonction aléatoire en n points expérimentaux. On suppose que cette fonction est stationnaire, et qu'on connaît sa covariance et son variogramme. On définit alors la distance entre deux points i et i' par : $d^2(i,i') = E(Z_i - Z_{i'})^2$. On cherche un espace vectoriel tel qu'on puisse y placer les points i en respectant leurs distances : il faut donc le munir d'une forme quadratique convenable. On effectue l'Analyse Factorielle de ce nuage de points. La formule de transition permet de placer un point supplémentaire, si on connaît ses distances aux n points expérimentaux. On montre alors que cette formule est la formule du krigeage.

RAPPELS SUR L'ANALYSE DES TABLEAUX DE DISTANCE

Ce rappel est basé sur la note [2] de J.P. BENZECRI.

I - DEFINITION D'UN TRIPLE ET D'UNE DE SES REALISATIONS

Soit J un ensemble fini, à n éléments, notés $j, j' \dots$

Soit d une fonction positive ou nulle, sur $J \times J$,
symétrique et nulle sur la diagonale de $J \times J$

$$\forall j, \forall j' \quad d^{jj'} = d^{j'j} \geq 0$$

$$\forall j \quad d^{jj} = 0$$

Cette fonction joue le rôle d'un carré de distance.

Soit enfin un système μ_j de masses sur J , μ la masse totale, $f_j = \mu_j/\mu$ le système des masses relatives

On appelle triple la donnée (J, d^{JJ}, μ_J)

Soit maintenant un ensemble de n points M^j d'un espace affine construit sur un espace vectoriel muni d'une forme quadratique m , munis des masses μ_j .

On dit que cet ensemble est une réalisation du triple si :

$$\forall j, \forall j' \quad m(M^j M^{j'}, M^j M^{j'}) = d^{jj'}$$

2 - EXISTENCE DE REALISATIONS

Il est facile de construire une réalisation du triple.

Soit en effet R^n , muni d'une base $\{e_i / i = 1, n\}$

et définissons M^i par : $OM^i = e_i$

Définissons la forme quadratique par :

$$\forall j \neq i : M(e_i, e_j) = - \frac{d^{ij}}{2}$$

$$\forall i \quad M(e_i, e_i) = 0$$

On a :

$$M(M^i M^{i'}, M^i M^{i'}) = M(OM^{i'}, OM^{i'}) + M(OM^i, OM^i) - 2 M(OM^i, OM^{i'})$$

$$M(M^i M^{i'}, M^i M^{i'}) = d^{ii'}$$

3 - EQUIVALENCE DE TOUTES LES REALISATIONS

Soit E un espace réalisant le triple, et E* son dual.

Définition 1 : Centre de gravité

On appelle centre de gravité du nuage le seul point G satisfaisant à :

$$\vec{OG} = \sum_{j \in J} f_j OM^j$$

Définition 2 : Forme quadratique d'inertie τ

Soient u et v deux éléments de E*

On définit τ par :

$$\tau(u, v) = \sum_{j \in J} f_j u(GM^j) v(GM^j)$$

Ainsi considérée, τ est une application de $E^* \times E^*$ dans $\mathbb{R}$.
On peut aussi considérer τ comme une application de E^* dans E :
on définit $\tau(u)$ comme l'unique vecteur de E tel que, pour tout v
 $v(\tau(u)) = \tau(u, v)$. Alors :

$$\tau(u) = \sum_{j \in J} f_j u(GM^j) GM^j$$

Définition 3

Soit R l'application de E^* dans $\mathbb{R}^n$ qui, à $u \in E^*$, fait correspondre $\varphi^j \in \mathbb{R}^n$ tel que : $j : \varphi^j = u(GM^j)$.

Soit E_λ^* le sous-espace propre relatif à la valeur propre λ non nulle de $m_0 \tau$ (qui est l'opérateur à diagonaliser pour trouver les axes principaux d'inertie).

Il lui correspond le sous-espace $R_\lambda^J = R(E_\lambda^*)$ dans $\mathbb{R}^n$.

On dit que deux représentations du triple sont équivalentes, si elles engendrent les mêmes spectres et les mêmes sous-espaces spectraux dans R^n (correspondant aux valeurs propres différentes de 0).

Théorème : Toutes les réalisations d'un même triple sont équivalentes

En effet, u sera direction propre de $m_0 \tau$ si

$$m_0 \tau(u) = \lambda(u)$$

Ce qui s'écrit aussi :

$$x \in E \quad m(\tau(u), x) = \lambda u(x)$$

Comme on ne s'intéresse qu'au support du nuage (c'est-à-dire aux valeurs propres non nulles), il suffit d'écrire cette équation pour un système générateur du support du nuage, donc pour $x_i = GM^i$. Compte tenu de la définition de $\tau(u)$, l'équation devient :

$$\textcircled{1} \quad \forall i : \sum f_j u(GM^j) m(GM^i, GM^j) = \lambda u(GM^i)$$

Soit, pour l'espace R_λ^J et les fonctions φ^J :

$$\textcircled{2} \quad R_\lambda^J = \{ \varphi^J \mid \forall i : \sum f_j \varphi^j m(GM^i, GM^j) = \lambda \varphi^i \}$$

A ce stade, il est clair que spectre et sous-espaces spectraux dans R^n sont déterminés par le système $\{f_j, j \in J\}$ et le système $\{m(GM^j, GM^{j'}) \mid j \in J, j' \in J\}$.

Il suffit donc, pour prouver l'équivalence de toutes les réalisations d'un même triple, de montrer que $m(GM^j, GM^{j'})$ ne dépend que du triple, et non de sa réalisation.

$$\text{Or } m(GM^j, GM^{j'}) = m(\sum_k f_k M^k M^j, \sum_l f_l M^l M^{j'})$$

$$m(GM^j, GM^{j'}) = \sum_k \sum_l f_k f_l m(M^k M^j, M^l M^{j'})$$

On connaît la formule classique des triangles :

$$2 m (AB, AC) = m(AB, AB) + m(AC, AC) - m(BC, BC)$$

Soit, pour 4 points A, B, C, D :

$$2 m (AB, CD) = 2 m(AB, AD) - 2 m(AB, AC)$$

$$2 m (AB, CD) = m (AD, AD) - m(BD, BD) - m (AC, AC) + m (BC, B$$

Dans le cas présent :

$$2 m(M^k M^j, M^l M^{j'}) = d^{kj'} - d^{jj'} - d^{kl} + d^{jl}$$

L'équation 2 ne dépend que du triple, et non de sa réalisation. Le théorème est démontré.

4 - CALCUL PRATIQUE DES COEFFICIENTS DE 2

Appliquons l'algorithme du triangle :

$$\textcircled{3} \quad 2 m(GM^j, GM^{j'}) = m(GM^j, GM^j) + m(GM^{j'}, GM^{j'}) - d^{jj'}$$

Il suffit de calculer $m(GM^j, GM^j)$.

Appliquons pour cela le théorème d'Huyghens. En désignant par I_a l'inertie par rapport à un point a quelconque.

$$\textcircled{4} \quad I_{Mi} = I_G + m(GM^i, GM^i)$$

Or :

$$I_{Mi} = \sum f_j m(M^i M^j, M^i M^j) = \sum_j f_j d^{ij}$$

$$\text{Posons : } d^{\circ i} = \sum_j f_j d^{ij}$$

En sommant 4 pour tous i :

$$\sum_i f_i I_{Mi} = I_G + \sum f_i m(GM^i, GM^i) = 2 I_G$$

$$\text{Donc : } 2 I_G = \sum_i f_i d^{\circ i} = \sum_i \sum_j f_i f_j d^{ij}$$

$$\text{Posons : } d^{\circ \circ} = \sum_j f_j d^{\circ j} = \sum_i \sum_j f_i f_j d^{ij}$$

On déduit :

$$m(GM^j, GM^j) = d^{°j} - \frac{1}{2} d^{°°}$$

Et :

$$m(GM^j, GM^{j'}) = \frac{1}{2} [d^{°j} + d^{°j'} - d^{°°} - d^{jj'}]$$

5 - REPRESENTATION D'UN POINT SUPPLEMENTAIRE

Supposons qu'on connaisse les distances d'un point jsu à tous les autres points de J, mais qu'on accorde à ce point jsu une masse nulle : il ne contribue pas à la définition des axes, mais peut être placé après coup, par :

$$\lambda \varphi^{jsu} = \sum_j f_j \varphi^j m(GM^j, GM^{jsu})$$

C'est sur cette équation que nous reviendrons pour montrer qu'elle équivaut à l'équation du krigeage simple dans le cas particulier que nous traiterons.

REPRESENTATION EUCLIDIENNE ET FONCTION ALEATOIRE

I - FONCTION ALEATOIRE ET DISTANCE

Soit $\{Z_x\}_{x \in R^n}$ une fonction aléatoire, et $\{z_x\}_{x \in R^n}$ une de ses réalisations. Pour fixer les idées, supposons que R^n soit l'espace ambiant de dimension 3, et que z_x soit la teneur en un certain minéral, au point x .

On dit que la fonction aléatoire est stationnaire d'ordre 2 si :

- . $E(Z_x)$ est indépendant de x
- . $E(Z_x - E(Z_x))(Z_y - E(Z_y))$ ne dépend que du vecteur $x-y$.

On note :

$$C(h) = E(Z_x - E(Z_x))(Z_{x+h} - E(Z_{x+h}))$$

la covariance centrée de la fonction aléatoire.

On appelle variogramme de la fonction aléatoire la fonction γ définie par :

$$\gamma(h) = \frac{1}{2} E(Z_{x+h} - Z_x)^2$$

On remarque que :

$$\gamma(h) = C(0) - C(h)$$

Supposons connue la réalisation de la fonction aléatoire en n points $\{x_j\}$. Soit J l'ensemble des points connus.

On peut définir une distance entre deux points connus x_j et $x_{j'}$, par :

$$d^2(x_j, x_{j'}) = 2 \gamma(x_j - x_{j'})$$

Pour rejoindre le problème du triple, il suffit d'accorder des poids à nos échantillons connus. Il faut cependant avoir présent à l'esprit le fait suivant : le système de poids qu'on accorde aux échantillons permet de définir les sous-espaces de dimension donnée les plus "proches" du nuage de points, au sens de l'inertie. Mais, en ce qui concerne la disposition des points dans l'espace R^n tout entier, le système de poids n'intervient pas, et les figures correspondant à deux systèmes de poids différents y sont superposables.

II - L'ESTIMATION DE LA MOYENNE

Posons, à titre de curiosité, en termes d'analyse de distance, le problème de l'estimation linéaire optimale de l'espérance $E(Z_x)$, à partir des $\{z_{x_j}/j = 1, n\}$.

Parmi tous les centres de gravité possibles de l'ensemble des points J, quel est le point par rapport auquel l'inertie est ~~maximale~~
minimisée.

Supposons que les $\{z_{x_j}, j=1, n\}$ sont représentés par des points M^j d'un espace vectoriel quelconque, muni d'une métrique telle que $m(M^j M^{j'}, M^j M^{j'}) = 2\gamma(x_j, x_{j'})$.

Un centre de gravité est un point G défini par :

$$OG = \sum_j \lambda_j OM_j, \text{ avec } \sum_j \lambda_j = 1$$

L'inertie par rapport à G est :

$$I_G = \sum_j \lambda_j m(GM^j, GM^j)$$

D'après le paragraphe 1 :

$$m(GM^j, GM^j) = 2 \sum_{j'} \lambda_{j'} \gamma(x_j, -x_{j'}) - \sum_i \sum_j \lambda_i \lambda_j \gamma(x_i - x_j)$$

D'où :

$$I_G = \sum_i \sum_j \lambda_i \lambda_j \gamma(x_i - x_j)$$

I_G est minimum sous la contrainte $\sum \lambda_i = 1$ si :

$$\forall i : \frac{\partial I_G}{\partial \lambda_i} = \mu \quad (\text{Formalisme de Lagrange})$$

Soit :

$$\begin{aligned} \forall i \quad \sum_j \lambda_j \gamma(x_i - x_j) &= \mu \\ \sum \lambda_i &= 1 \end{aligned}$$

On retrouve l'équation du Krigage (cf. G. MATHERON [1]) de l'espérance de la Fonction aléatoire, à ceci près que le variogramme remplace la Covariance. Comme $\gamma(h) = C(0) - C(h)$ on peut aussi écrire :

$$\begin{aligned} \forall i : \quad \sum \lambda_j C(x_i - x_j) &= C(0) - \mu \\ \sum \lambda_i &= 1 \end{aligned}$$

III - KRIGEAGE PONCTUEL SIMPLE ET ELEMENTS SUPPLEMENTAIRES

a) KRIGEAGE SIMPLE A MOYENNE CONNUE

Ce rappel est basé sur la référence [1] de G. MATHERON. Soit $Z(x)$ une fonction aléatoire de moyenne connue m .

Alors $Y(x) = Z(x) - m$ est une fonction aléatoire de moyenne nulle. C'est sur Y qu'on va travailler, étant entendu qu'il est aisé de remonter à Z . Soit $C(h)$ la covariance de Y .

Supposons connue la réalisation de Y aux point x_j , $j = 1, n$. On notera $C_{jj'} = C(x_j, -x_{j'})$.

La réalisation est inconnue au point x_i . Quel est le meilleur estimateur linéaire $Y_{x_i}^*$ de Y_{x_i} en fonction des $\{Y_{x_j} / j = 1, n\}$.

$$\text{Soit } Y_{x_i}^* = \sum_j \lambda_j Y_{x_j}$$

On détermine le meilleur $Y_{x_i}^*$ en minimisant la variance de l'erreur $e_i = Y_{x_i} - Y_{x_i}^*$. La fonction aléatoire étant d'espérance nulle, la variance de e_i est :

$$E(e_i^2) = E(Y_{x_i}^* - Y_{x_i})^2 = C(o) - 2 \sum_j \lambda^j C_{ij} + \sum_j \sum_k \lambda^j \lambda^k C_{jk}.$$

Cette forme quadratique en λ^j est minimale si
 $\forall j : \frac{\partial}{\partial \lambda_j} E(e_i^2) = 0.$

Soit :

$$\forall j : \sum \lambda^k C_{jk} = C_{ij}$$

En termes matriciels, si on appelle C la matrice de terme général C_{jk} , C_i le vecteur colonne de terme général C_{ij} , Λ le vecteur colonne de terme général λ^k , on obtient :

$$C \Lambda = C_i$$

La variance de l'estimateur optimal est :

$$E(e_i^2) = C(o) - \sum_j \lambda^j C_{ij}$$

Il est facile de vérifier que le krigeage est un interpolateur exact (l'estimateur optimal d'un point connu Y_{x_j} est $Y_{x_j}^* = Y_{x_j}$ (en faisant $\lambda^j = 1$ et $j \neq j' \Rightarrow \lambda^{j'} = 0$, on obtient une variance d'erreur nulle).

Le formalisme du krigeage s'étend à une gamme de problèmes beaucoup plus vaste : d'une part on peut s'affranchir de la stationnarité, d'autre part on peut estimer des combinaisons linéaires (finies ou non) de points de la fonction aléatoire. Enfin le cokrigeage permet de résoudre les problèmes d'estimation optimale lorsque plusieurs fonctions aléatoires sont en jeu. Pour tous ces problèmes, se reporter à G. MATHERON ([1]).

b) PROJECTION D'ELEMENT SUPPLEMENTAIRE ET KRIGEAGE SIMPLE

Soit Y_x une fonction aléatoire, stationnaire d'ordre deux, de moyenne inconnue, et soit $\{Y_x\}$ une de ses réalisations, connue aux points $\{x_j / j = 1, n\}$. Le triple associé est :

J : ensemble des points connus

d^{JJ} : distance sur $J \times J$ $d^{jj'} = 2\gamma(x_j - x_{j'})$

μ_J : système de poids.

On détermine μ_J , de manière à maximiser l'inertie sous la contrainte $\sum \mu_j = 1$ (cf. II : solution équivalente à la détermination optimale de la moyenne par krigeage).

Pour former l'équation des facteurs, on doit d'abord former les produits scalaires $m(GM^j, GM^{j'})$. On a vu page (6) que :

$$m(GM^j, GM^{j'}) = \frac{1}{2} [d^{0j} + d^{0j'} - d^{00} - d^{jj'}]$$

$$m(GM^j, GM^{j'}) = \frac{1}{2} [\sum_j \mu_j d^{jj'} + \sum_{j'} \mu_{j'} d^{jj'} - \sum_j \mu_j \mu_{j'} d^{jj'} - d^{00}]$$

$$m(GM^j, GM^{j'}) = E(Y^j - m^*)(Y^{j'} - m^*) = C_{jj'}$$

en appelant $C_{jj'}$ la covariance centrée entre Y_{x_j} et $Y_{x_{j'}}$.

L'équation des facteurs s'écrit :

$$\forall i \quad \forall \alpha \quad \sum_j \mu_j \varphi_\alpha^j C_{ji} = \lambda \varphi_\alpha^i$$

L'équation de projection d'un point supplémentaire i_{su} s'en démarque immédiatement :

$$\sum_j \mu_j \varphi_\alpha^j C_{j i_{su}} = \lambda \varphi_\alpha^{i_{su}}$$

Elle exprime les coordonnées du vecteur $GM^{i_{su}}$ (représentant le point $Y_{x_{i_{su}}} - m^*$, sur les axes factoriels orthogonaux e_α).

Exprimons, parallèlement, l'équation du krigeage simple de $Y_{x_{isu}} - m^*$, en fonction des $Y_{x_j} - m^*$.

Vectoriellement, elle s'écrit :

$$G^* M^{isu} = \sum_j v_j G M^j$$

avec

$$C v = C_{isu}$$

C : matrice des covariances des points connus

v : matrice colonne des pondérateurs optimaux

C_{isu} : matrice colonne, avec $C_{j\ isu} = E(Y_{x_j} - m^*)(Y_{x_{isu}} - m^*)$.

Soit D_μ la matrice dont tous les termes extradiagonaux sont nuls, et dont les termes diagonaux sont les coefficients du krigeage de m^*

$$D_\mu = \begin{pmatrix} \mu_1 & & 0 \\ & \mu_2 & \\ 0 & & \ddots & \mu_n \end{pmatrix}$$

L'équation $C v = C_{isu}$ est équivalente, si aucun des μ_i n'est nul, à :

$$D_\mu C v = D_\mu C_{isu}$$

Posons :

$$K = D_\mu C$$

$$K_{isu} = D_\mu C_{isu}$$

Alors $\forall i \ \forall j$

$$k_{ij} = \mu_i C_{ij}$$

$\forall i$

$$k_{i\ isu} = \mu_i C_{i\ isu}$$

L'équation devient : $K v = K_{isu}$.

Dans la base des vecteurs propres de K, K devient une matrice diagonale K^0 , le ième terme de la diagonale étant λ_i , le ième terme de v devient la projection du point isu sur le ième axe ; K_{isu} (matrice colonne des coordonnées d'un vecteur) se transforme de manière contravariante :

Comme on a :

$$GM^j = \sum_i \sqrt{\lambda_i} \varphi_i^j e_i$$

$$K_{i, isu}^0 = \sum_j \sqrt{\lambda_i} \varphi_i^j k_{j, isu}$$

L'équation du krigeage, dans ces axes, s'écrit donc :

$$\lambda_i \sqrt{\lambda_i} \varphi_i^{isu} = \sum_j \sqrt{\lambda_i} \varphi_i^j \mu_j c_{j, isu}$$

On retrouve donc l'équation de projection d'un point supplémentaire en Analyse de tableau de distance.

Cette remarque permet de justifier à posteriori la formule de projection d'un élément supplémentaire en Analyse de données : c'est la projection optimale au sens de la distance choisie. Nous allons préciser au paragraphe suivant le sens de cette optimalité.

c) LE SENS DE L'OPTIMUM

Une façon de voir le krigeage simple est la suivante : les variables aléatoire Z_x appartiennent à un espace de Hilbert. Kriger le point Z_α , c'est le projeter orthogonalement sur le sous espace engendré par les variables aléatoires $\{Z_{x_i}, i = 1 \dots n\}$ dont la réalisation est connue. Pour cela, il faut et il suffit qu'il existe la relation suivante entre Z_α et son estimateur $\hat{Z}_\alpha$

$$\forall i : \quad \langle Z_\alpha, Z_{x_i} \rangle = \langle \hat{Z}_\alpha, Z_{x_i} \rangle$$

La même condition est vérifiée en Analyse des Données. Nous allons montrer que, si l'on appelle e_k les axes factoriels engendrant le support du nuage (c'est-à-dire correspondant à des valeurs propres non nulles), on a :

$$\langle Z_\alpha, e_k \rangle = m \langle GM^{jsu}, e_k \rangle$$

On a vu que, si les points M^j appartiennent à un espace E , la forme d'inertie τ peut être considérée comme une application de E^* dans E définie par :

$$\tau(u) = \sum_j f_j u(GM^j) GM^j$$

A un facteur $\{\varphi_K^j, j=1, n\}$, relatif à la valeur propre λ_K , est associé u tel que, pour tout j :

$u_K(GM^j) = \varphi_K^j$, et un vecteur e_K tel que :

$$e_K = \frac{\tau(u_K)}{\sqrt{\lambda_K}} = \frac{1}{\sqrt{\lambda_K}} \sum_j f_j \varphi_K^j GM^j$$

Il est facile de vérifier que e_K est l'axe factoriel, de norme 1, associé au projecteur $\{\lambda_K \varphi_K^j, j=1, n\}$. En effet :

$$m(GM^1, e_K) = \frac{1}{\sqrt{\lambda_K}} \sum_j f_j \varphi_K^j m(GM^j, GM^1) = \sqrt{\lambda_K} \varphi_K^1$$

$$m(e_K, e_K) = m(e_K)(e_K)$$

$$m(e_K, e_K) = \frac{1}{\lambda_K} m(\tau(u_K)(\tau(u_K)))$$

$$m(e_K, e_K) = \lambda_K / \lambda_K u_K(\tau(u_K)) = 1$$

Formons $m(GM^{jsu}, e_K)$

$$m(GM^{jsu}, e_K) = m(GM^{jsu}, \frac{1}{\sqrt{\lambda_K}} \sum_j f_j \varphi_K^j GM^j)$$

$$m(GM^{jsu}, e_K) = \frac{1}{\sqrt{\lambda_K}} \sum_j f_j \varphi_K^j m(GM^{jsu}, GM^j)$$

$$m(GM^{jsu}, e_K) = \sqrt{\lambda_K} \varphi_K^{jsu}$$

par définition de φ_K^{jsu}

$$m(GM^{jsu}, e_K) = m(G^* M^{jsu}, e_K)$$

Donc, pour tout e_K , c'est-à-dire pour tous vecteurs engendrant le support du nuage, l'égalité de projection orthogonale est vérifiée.

La projection d'un individu supplémentaire est m-orthogonale, et réalise donc le minimum de $m(GM^{jsu} - GM^{*jsu}, GM^{jsu} - GM^{*jsu})$.

Il faut remarquer que la projection de M^{jsu} étant orthogonale sur le support tout entier du nuage, elle l'est aussi sur chacun des axes principaux d'inertie, donc sur tous les sous-espaces engendrés par ceux-ci : lorsqu'on visualise un nuage de points sur un plan défini par deux axes quelconques la projection d'un point supplémentaire y est aussi optimale.

BIBLIOGRAPHIE

- [1] G. MATHERON Théorie des Variables Régionalisées et Applications.
Fascicule 5 des Cahiers du Centre de Morphologie
Mathématique. FONTAINEBLEAU

- [2] J.P. BENZECRI Représentation euclidienne d'un ensemble fini
muni de masses et de distances. Publication ISUP
PARIS VI°