UN SURVOL DES METHODES ÉLÉMENTAIRES EN STATISTIQUE SPATIALE

M. HANNOUN[(*)]

[(*)]INSEE, Département de l'action régionale

Résumé

Le développement conjoint des bases de données géoréférencées et des systèmes d'information géographique place la carte au premier plan des outils d'analyse spatiale en démographie et en économie. Le statut des méthodes statistiques s'en trouve transformé et renforcé : nécessité d'adapter les outils « classiques » à la spécificité spatiale, appropriation d'outils mis au point dans des disciplines éloignées comme la géologie ou la météorologie. Ce papier survole trois domaines : la statistique descriptive, le géostatistique et l'économétrie.

Introduction

Les données spatialisées ou géoréférencées constituent aujourd'hui une matière première (sur) abondante produite par de nombreuses disciplines :
Prospection minière et pétrolière, géologie, géographie, météorologie, télédétection, océanographie, astronomie, écologie, épidémiologie, chirurgie, démographie, économie et des disciplines plus exotiques (« crime incident location data », modélisation de la détection et de la délimitation des champs de mines ou des bancs de poissons, etc...).

Cette croissance du volume des données disponibles est concomitante à une offre de plus en plus riche de logiciels de cartographie et de systèmes d'information géographique (SIG). Cette vulgarisation des outils graphiques induit une production croissante de cartes thématiques. Cette évolution quantitative se double d'une modification du statut même de la carte thématique qui voit son rôle illustratif se transformer souvent en générateur d'hypothèses ou même de théories.

Le syndrome de la « belle carte »

L'impact visuel définit de manière ambiguë le pouvoir explicatif de la carte, c'est-à-dire sa capacité à mettre en valeur des configurations et des rapports spatiaux qui ne soient pas que le pur produit du hasard ou de nos a priori :

● La « belle carte » est alors celle qui conforte notre intuition ou nos préjugés.

● La « belle carte » peut aussi être « fausse » en ce sens que les corrélations spatiales qu'elle exhibe n'ont aucune consistance statistique, les hypothèses qui sont alors formulées reposent alors soit sur une configuration due au hasard soit sur un artéfact statistique, biais de collecte par exemple.

Les outils de statistique spatiale qui permettent de mesurer le degré de signification statistique des configurations et des relations spatiales des données géoréférencées sont des instruments complémentaires indissociables de la démarche purement cartographique.

Spécificité du domaine spatial

L'arsenal des procédures statistiques classiques (descriptives, analyse des données, économétrie) est relativement inopérant en l'état dans les applications mettant en œuvre des données spatiales. De même que l'analyse des séries chronologiques fait appel à des techniques ad hoc, les spécialistes de l'analyse et la modélisation des données géoréférencées proposent des outils adaptés aux deux spécificités majeures des mesures spatiales :

❶ L'hétérogénéité, chaque localisation est unique, influence de la géographie physique par exemple.

❷ La dépendance, l'autocorrélation spatiale, est la règle, condition ontologique de l'existence de la carte :
« Everything is related to everything else, but near things are more related than distant things » : Tobler, première loi de la géographie (1979).

L'autocorrélation spatiale est donc la corrélation d'une variable géoréférencée avec elle-même. Il existe alors une relation fonctionnelle entre les points de l'espace plus ou moins proches. L'autocorrélation sera positive si les variables proches ont des valeurs semblables et si les variables éloignées ont des valeurs différentes. L'autocorrélation sera négative si les variables proches ont des valeurs différentes et si les variables éloignées ont des valeurs semblables. Enfin il n'y a pas d'autocorrélation spatiale si les tests ne décèlent aucune relation entre la valeur et la localisations des variables.

Le phénomène d'autocorrélation spatiale trouve ses origines dans deux catégories de processus :

➡ Des processus de diffusion spatiale, les zones proches du foyer d'une épidémie seront, en général, plus touchées que les zones éloignées, les dégâts seront fonction de la distance à l'épicentre d'un tremblement de terre.

➡ Des processus d'interactions spatiale, l'interaction spatiale prend en compte les mouvements dans l'espace qui permettent d'expliquer les flux entre entités : zones de hautes et de basses pressions en météorologie, phénomène de symbiose en écologie, complémentarité technique entre donneurs et preneurs d'ordre dans l'industrie, zones de résidence et zones d'activité, migrations pendulaires ou définitives, en démographie.

Deux différences principales entre approche spatiale et temporelle

❶ L'axe du temps est unidirectionnel, l'espace est omnidirectionnel. Dans le temps les interactions sont unidirectionnelles, passé ➡ présent ➡ futur, dans l'espace les interactions sont omnidirectionnelles. Les séries temporelles sont extrapolées du passé vers le futur, les données spatiales sont interpolées dans le plan ou dans un volume (géologie).
❷ Les observations temporelles sont généralement régulièrement espacées, les observations spatiales le sont rarement.

Trois catégories de données spatiales :

❶ **Données continues**, existent et mesurables théoriquement en tout point mais en pratique, que pour les stations de mesures. Par exemple, l'altitude ou la température, la teneur en minerais d'un gisement, sont des variables continues interpolables entre les points de mesure proches.
❷ **Données ponctuelles**, existent et ne sont mesurées qu'en un nombre fini de points, par exemple : l'épicentre d'un séisme auquel on peut attacher des variables

comme la magnitude, la durée, la profondeur ; les arbres d'une forêt, localisation, espèces, âge, taille ; les pixels d'une image satellite.

❸ **Données de superficie**, ce sont des zones contiguës, le plus souvent administratives, de collecte statistique etc... Des données ponctuelles continues peuvent aussi être transformées en données de superficie par carroyages ou triangulations de Delaunay.

Trois « boites à outils » pour le « bricolage » des données spatiales :

❶Analyse exploratoire des données spatiales, concerne les données continues, ponctuelles et de superficie.

❷ Analyse géostatistique, ne concerne, à l'origine, que les variables d'un processus spatial continu.

❸ Econométrie spatiale, concerne des ensembles finis de points ou de zones.

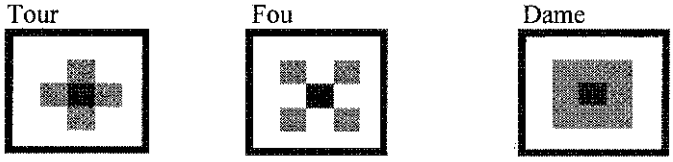
Un outil clef : les matrices de contiguïté et de pondération spatiales

Deux grands types de matrices traduisent les interactions spatiales ou autres entre les variables géoréférencées : les matrices spatiales de contiguïté, les matrices spatiales de poids.

❶ Matrices de contiguïté

Deux zones sont contiguës si elles ont au moins une frontière commune. La matrice de contiguïté d'ordre un[1], le cas le plus général, est une matrice $W(n,n)$ avec n nombre de zones. La matrice est donc carrée et symétrique. $w_{ij} = 1$ si les zones ont au moins une frontière commune, $w_{ij} = 0$ sinon. Par convention $w_{ii} = 0$.

Dans la littérature (Cliff et Ord 1981), on distingue trois types de contiguïté conventionnels :



[1] La notion de contiguïté peut être généralisée à l'ordre k si k est le nombre minimal de frontières à franchir pour aller de la zone i à la zone j.

- Tour : La zone centrale a quatre voisins dans les directions Nord-Sud et Est-Ouest.

- Fou : La zone centrale a quatre voisins dans les directions Nord-Est Sud-Ouest et Sud-Est Nord-Ouest.

- Dame : La zone centrale a huit voisins dans les directions Nord-Sud, Est-Ouest, Nord-Est Sud-Ouest et Sud-Est Nord-Ouest.

Le nombre total de zones contiguës à une zone i est égale à la somme en ligne des éléments de la matrice carrée symétrique : $L_i = \sum_{j=1}^{n} w_{ij}$. Ces matrices peuvent être standardisées pour que la somme de chaque ligne soit égale à 1 : $\tilde{w} = \frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}}$

❷ Matrices spatiales de pondération.

Pour ces matrices, l'intensité des interactions entre zones ne reflète pas obligatoirement la proximité spatiale. Les valeurs de pondération ne sont pas nécessairement symétriques, par exemple la distance ou le temps de transport de i vers j est plus court que le temps de transport de j vers i. Un nombre important de distances sont en pratique mises à contribution : indicateurs non spatiaux, distance euclidienne, distance réelle par la route, le rail..., diverses relations fonctionnelles telles celles proposée par Cliff et Ord : d_{ij} = distance entre les zones (centroïdes) ou entre les points i et j, $\beta_{i(j)}$ = pourcentage de frontière commune entre les zones i et j, $\sum_{j=1}^{n} \beta_{i(j)} = 1$, alors : $w_{ij} = \frac{1}{d_{ij}^{a}} (\beta_{i(j)})^{b}$, avec a et b > 0 sont des paramètres estimés. $w_{ij} = 0$ pour toutes les zones non contiguës.

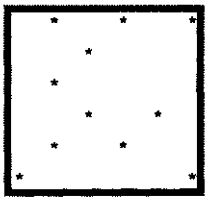
1. Analyse exploratoire des données spatiales.

1.1 Analyse de la configuration de points

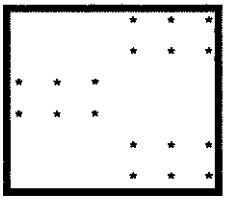
L'analyse de la configuration de points est un ensemble d'outils d'analyse de la forme de la distribution d'un ensemble fini de N points (variables d'intérêt) cartographié en coordonnées cartésiennes (x,y) à l'intérieur d'une zone d'étude Z : $Z \subset R^2$.

Exemples : Villes et leurs populations, Galaxies par formes et tailles, tremblements de terre par intensités et durées. L'analyse de la configuration de points met en évidence le ou les liens entre la forme géographique, l'intensité et la dépendance entre les variables géoréférencées. L'intensité est décrite par la densité, nombre moyen de points par unité de surface, la dépendance par les interactions entre les points. Les N points peuvent se répartir en trois grands types de configurations :

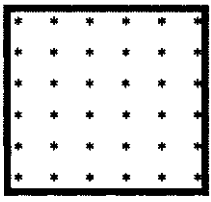
Aléatoire



Agrégée



Régulière



Aperçu sur les principaux outils.

1.1.1 Indicateur du voisin le plus proche *IVPP* :

C'est le plus ancien (Clark et Evans 1954), le plus simple et le moins performant des indicateurs. L'indicateur du voisin le plus proche compare les distances entre les points de la configuration à la distance attendue si la dite configuration était aléatoire.

Pour chacun des N points de la zone de surface A, on calcule :

La distance moyenne au voisin le plus proche (VPP) : ➡ $\bar{d} = \frac{\sum_{i=1}^{N} d_{ij}}{N}$

avec $d_{ij} = \sqrt{(x_i - x_j)^2 + (x_i - x_j)^2}$

La distance moyenne au VPP attendue si la configuration est aléatoire

➡ $E(\bar{d}) = \frac{1}{2}\sqrt{\frac{A}{N}}$

Variance de la distance au VPP si la configuration est aléatoire ➡ $V(\bar{d}) = \frac{(4-\pi)A}{4\pi N^2}$

Indicateur du voisin le plus proche ➡ IVPP = $\left(\frac{\sum_{i=1}^{N} d_{ij}}{N}\right) / \frac{1}{2}\sqrt{\frac{A}{N}}$

Si la configuration est aléatoire, IVPP est égal ou proche de 1, si la configuration est régulière, IVPP est supérieur à 1, si elle est agrégée, IVPP est compris entre 0, agrégation totale, et inférieur à 1.

La variable centrée réduite de $\bar{d}$ forme un test de signification de l'IVPP

➡ $z = \frac{\bar{d} - E(\bar{d})}{\sqrt{V(\bar{d})}}$

1.1.2 Fonction K de Ripley :

Cette technique (Ripley 1976) est basée sur le décompte du nombre de points situés en deçà d'une distance ou d'une classe de distances :

➡ $K(d) = \lambda^{-1} E(d)$ Avec :

$\lambda = n/A$, densité observée des n points dans la zone de surface A

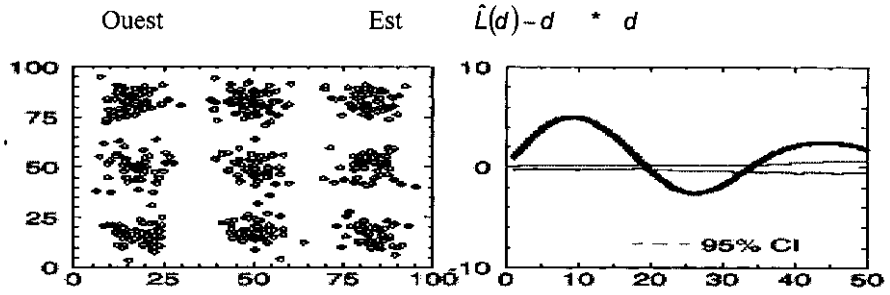
$E(d)$ nombre de points attendus en deçà de la distance d.

$E(d)$ est estimé de manière robuste par :

$\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_{ij}}{n}$ avec $\delta_{ij} = 1$ si $x_i - x_j \le d$ et $\delta_{ij} = 0$ si $x_i - x_j > d$. ➡ $\hat{K}(d) = \lambda^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_{ij}}{n}$.

Pour une configuration aléatoire la valeur attendue est : $K(d) = \pi d^2$, soit :

$\hat{L} = \sqrt{\frac{k(d)}{\pi}}$. On trace généralement le graphique $\hat{L}(d) - d \quad * \quad d$



1.1.3 Fonction K bivariée :

Cette extension de la fonction de Ripley prend en compte les distances entre deux types de variables géoréférencées, espèces végétales ou animales par exemple. En réarrangeant la formule univariée avec $\lambda = n / A$, on obtient :

➡ $\hat{K}(d) = \frac{A}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{ij}(d_{ij})$ qui, pour deux types de points 1 et 2, devient :

➡ $\hat{K}_{12}(d) = \frac{A}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \delta_{ij}(d_{ij})$

avec $\delta_{ij} = 1$ si les points *i* et *j* sont en deçà de la distance *d* et sont du type 1 et 2 et n_1 et n_2 taille de chacun des groupes.

1.2 Mesures de l'autocorrélation spatiale

Une bonne douzaine d'indicateurs ont été élaborés depuis le début de la décennie 50 en réponse aux problèmes de la détection et de la mesure des configurations spatiales originales. Dans cette brève présentation nous distinguerons :

- Les tests concernant les variables qualitatives, test de couleurs des cartes.
- Les test globaux d'autocorrélation spatiale, la carte présente-t-elle une autocorrélation spatiale ?
- Les tests locaux d'association spatiale (LISA), la valeur observée à la localisation *i* est-elle entourée d'un agrégat des valeurs plus hautes ou plus basses ? La valeur observée en *i* est-elle associée positivement aux valeurs des localisations voisines (ressemblance) ou négativement (dissemblance) ?

1.2.1 Tests d'autocorrélation spatiale des données qualitatives.

Dans le cas binaire, chaque zone ***Z*** est codée soit ***B*** (noir = 1) si elle possède le caractère, soit ***W*** (blanc=0) si elle ne le possède pas. Ces tests sont généralisables à n modalités.

On considère une matrice de contiguïté simple du type $w_{ij} = 1$ si les zones i et j ont au moins une frontière commune, $w_{ij} = 0$ sinon et $w_{ii} = 0$, $w_{ij} = w_{ji}$.

Trois configurations de base sont possibles : ***BB***, ***BW*** et ***WW***.
Le nombre de configurations du type :

$$BB = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} w_{ij}Z_iZ_j\,,\quad BW = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} w_{ij}(Z_i - Z_j)^2\,,\quad WW = A - (BB + BW).$$

Avec : $A = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} w_{ij}$ nombre total de contiguïtés de la carte.

Sous l'hypothèse nulle H_0 d'absence d'autocorrélation spatiale, nous pouvons préciser pour ***BB*** et ***BW*** l'une des lois de probabilité suivantes :

- ***N*** normale, la couleur de chacune des zones est tirée aléatoirement, loi binomiale, tirage avec remise, avec une probabilité ***p*** pour les ***B*** (noir) et ***1-p*** pour les ***W*** (blanc). Sous H_0, absence d'autocorrélation spatiale, les zones sont indépendantes.
- ***R*** « randomisation », les valeurs n_1 de ***B*** et n_2 de ***W*** sont connues, $n = n_1 + n_2$. L'affectation d'une couleur à chaque zone est aléatoire, le tirage est sans remise, il y a $n\,!$ permutations possibles.

Pour chacune des hypothèses ***N*** et ***R***, on peut calculer l'espérance mathématique et la variance de ***BB*** et ***BW*** (cf Cliff et Ord 1981 chap 2). On montre que pour n grand :

$z_{BB} = \dfrac{BB - E(BB)}{\sqrt{V(BB)}}$ et $z_{Bw} = \dfrac{BW - E(BW)}{\sqrt{V(BW)}}$ suivent une loi normale centrée réduite.

On teste alors ***N*** et ***R*** sous H_0 d'absence d'autocorrélation spatiale, en comparant z_{BB} et z_{BW} à α de la loi normale à 1% (2,6) ou 5% (1,96).

Avec : n_1 nombre de zones ***B***,

$$S_0 = \sum_{i}^{n}\sum_{j,j\neq i}^{n} w_{ij}\,,\quad S_1 = \frac{1}{2}\sum_{i}^{n}\sum_{j,j\neq i}^{n}(w_{ij} + w_{ji})^2\,,\quad S_2 = \sum_{i}^{n}(w_{i.} + w_{.j})^2\,,\quad n^{(i)} = n(n-1)..(n-i-1)$$

- $E_N(BB) = \frac{1}{2}S_0p^2$,

$$Var_N(BB) = \frac{1}{4}p^2(1-p)[S_1(1-p) + S_2p]$$

- $E_R(BB) = \frac{1}{2} S_0 \frac{n_1^{(2)}}{n^{(2)}}$,

$$Var_R(BB) = \left[S_1 \left[\frac{n_1^{(2)}}{n^{(2)}} - 2\frac{n_1^{(3)}}{n^{(3)}} + \frac{n_1^{(4)}}{n^{(4)}} \right] + S_2 \left[\frac{n_1^{(3)}}{n^{(3)}} + \frac{n_1^{(4)}}{n^{(4)}} \right] + S_0^2 \left[\frac{n_1^{(4)}}{n^{(4)}} \right] - S_0 \left[\frac{n_1^{(2)}}{n^{(2)}} \right]^2 \right] * \frac{1}{4}$$

- $E_N(BW) = S_0 pq$, $Var_N(BW) = S_1 pq + \frac{1}{4} S_2 pq(1-4pq)$

- $E_R(BW) = S_0 \frac{n_1 n_2}{n^{(2)}}$,

$$Var_R(BW) = \frac{1}{2} S_1 \frac{n_1 n_2}{n^{(2)}} + \frac{1}{4}(S_2 - 2S_1)\frac{n_1 n_2}{n^{(2)}} + (S_0^2 + S_1 - S_2)\frac{n_1^{(2)} n_2^{(2)}}{n^{(4)}} - \left(S_0 \frac{n_1 n_2}{n^{(2)}} \right)^2$$

La significativité de *BB* et *BW* peut aussi être testée par des méthodes du type Monte-Carlo.

1.2.2 Tests globaux d'autocorrélation spatiale des données quantitatives.

Les tests les plus courants sont ceux de Moran (1950) et de Geary (1954).

1.2.2.1 Test I de Moran.

$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1, i \neq j}^{n} w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n} (z_i - \bar{z})^2}$, avec n nombre de zones et $S_0 = \sum_i^n \sum_{j, i \neq j}^n w_{ij}$ somme des éléments binaires de la matrice de contiguïté, z_i et z_j valeurs de la variable en i et j, $\bar{Z}$ valeur moyenne de la variable pour l'ensemble de la carte. Le numérateur est la covariance entre zones contiguës, le dénominateur la variance totale observée. L'espérance mathématique et la variance de I peuvent être calculées pour les hypothèses N et R :
Avec :

$$S_0 = \sum_i^n \sum_{j, i \neq j}^n w_{ij} \ , \ S_1 = \frac{1}{2} \sum_{i=1}^n \sum_{j=1, i \neq j}^n (w_{ij} + w_{ji})^2 \ , \ S_2 = \sum_{i=1}^n (w_{i.} + w_{.i})^2 \ , \ K = \frac{n \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)^2}$$

- $E_N(I) = -\frac{1}{(n-1)}$, $Var_N(I) = \frac{n^2 S_1 - n S_2 + 3S_0^2}{S_0^2 (n^2 - 1)} - E_N^2(I)$

- $E_R(I) = -\frac{1}{(n-1)}$,

$$Var_R(I) = \frac{n((n^2 - 3n + 3)S_1 - nS_2 + 3S_0^2) - K((n^2 - n)S_1 - 2nS_2 + 6S_0^2)}{S_0^2 (n-1)(n-2)(n-3)} - E_R^2(I)$$

Le test H_0 : pas d'autocorrélation spatiale est du type : $z = \frac{I - E(I)}{\sqrt{V(I)}}$. La valeur de I varie entre -1 et 1, $I = 0$: absence d'autocorrélation spatiale, $0 < I \leq 1$: autocorrélation positive, $-1 \leq I < 0$: autocorrélation négative.

1.2.2.2 Test C de Geary.

$$c = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} w_{ij}(z_i - z_j)^2}{2\left(\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} w_{ij}\right)\sum_{i=1}^{n}(z_i - z_j)^2}$$

- $E_N(c) = 1$, $Var_N(c) = \left(\frac{1}{2(n+1)S_0^2}\left((2S_1 + S_2)(n-1) - 4S_0^2\right)\right)$
- $E_R(c) = 1$,

$$Var_R(c) = \frac{\left((n-1)S_1\left(n^2 - 3n + 3 - (n-1)k\right)\right) - \left(\frac{1}{4}\left((n-1)S_2\left(n^2 + 3n - 6 - (n^2 - n + 2)k\right)\right)\right) + \left(S_0^2\left(n^2 - 3 - (n-1)^2 k\right)\right)}{n(n-2)(n-3)S_0^2}$$

Trois valeurs remarquables de c : $0 < c < 1$: forte autocorrélation positive ; $c \geq 1$: forte autocorrélation négative ; $c = 1$: distribution aléatoire, pas d'autocorrélation spatiale.

On peut aussi tester I et c, par un test de permutations de type « Monte Carlo ».

1.2.3 Indicateurs locaux d'association spatiale

L'identification des modèles locaux d'association spatiale pose deux problèmes : la zone d'intérêt i est-elle entourée de valeurs significativement basses ou hautes ? La valeur observée en i est-elle associée positivement (similarité) ou négativement (dissemblance) aux observations voisines ?

1.2.3.1 L'indicateur G_i(d) de Getis et Ord (1992).

$$G_i(d) = \frac{\sum_{j, j\neq i}^{n} w_{ij}x_j}{\sum_{j, j\neq i}^{n} x_j}, \quad E(G_i(d)) = \sum_{j, j\neq i}^{n} w_{ij}E(x_j) \Big/ \sum_{j, j\neq i}^{n} x_j \;, = \frac{W_i}{(n-1)},$$

$$Var(G_i(d)) = \frac{W_i(n-1-W_i)}{(n-1)^2(n-2)}\left(\frac{Y_{i2}}{Y_{i1}^2}\right)$$

Où : n est le nombre d'observations, x_i est la valeur observée au lieu i, $\{w_{ij}\}$ est une matrice de pondération spatiale symétrique binaire avec 1 pour w_{ij} si la zone j est à la distance d de i (ou contiguë au lieu i) et 0 sinon.

$$W_i = \sum_{j, j\neq i}^{n} w_{ij}, \quad Y_{i1} = \sum_{j\neq i}^{n} x_j/(n-1) \text{ et } \quad Y_{i2} = \sum_{j\neq i}^{n} x_j^2/(n-1) - Y_{i1}^2$$

Les tests de signification de G_i(d) ont une forme standardisée qui, dans Ord et Getis, (1994) est définie comme suit:

$$Z(G_i) = \frac{G_i - E(G_i)}{\sqrt{Var(G_i)}} = \frac{\sum_{j \neq i}^{n} w_{ij}(d)(x_j - \bar{x}_i)}{S_i \sqrt{w_i(n-1-w_i)/(n-2)}}$$

où x_i est la valeur observée au lieu i, $\bar{x}_i = \frac{1}{(n-1)} \sum_{j,j \neq i}^{n} x_j$, W_{ij} est la matrice de pondération spatiale symétrique binaire,

$w_i = \sum_{j,j \neq i}^{n} w_{ij}(d)$. et $S_i^2 = \frac{1}{(n-1)} \sum_{j,j \neq i}^{n} (x_j - \bar{x}_i)^2$.

Ord et Getis montrent que la distribution des permutations sous H_0: $G_i = E(G_i)$, absence d'autocorrélation spatiale autour de x_i, tend vers une loi normale. Une valeur significativement positive de $Z(G_i)$ révèle que la zone i est entourée par des zones où la variable prend des valeurs fortes, une valeur significativement négative de $Z(G_i)$ révèle que la zone i est entourée par des zones où la variable prend des valeurs faibles.

1.2.3.2 Les indicateurs Locaux d'association Spatiale - LISA

Anselin (1994) propose des indicateurs locaux d'association spatiale (LISA) comme une mesure alternative d'associations spatiales locales, qui incluent l'indice local de Moran et l'indice local de Geary. L'indice local de Moran permet l'identification des configurations spatiales agglomérées. L'indice local de Geary permet l'identification de similarités ou de dissemblances entre configurations spatiales. Un avantage de l'indice local de Moran et de l'indice local de Geary est qu'ils peuvent être associés aux indicateurs globaux (I de Moran et C de Geary) et peuvent être employés pour estimer la contribution des indicateurs individuels aux indicateurs globaux correspondants.

L'indice local de Moran pour chaque observation i est défini comme :

$$I_i(d) = Z_i \sum_{j \neq i}^{n} w_{ij} Z_j$$

où les observations Z_i et Z_j sont dans la forme standardisée (de moyenne nulle et de variance égale à un). La matrice de pondération spatiale w_{ij} est dans une forme standardisée par ligne. I_i est alors le produit de Z_i et de la moyenne des observations avoisinantes.

Un niveau de pseudo significativité de I_i peut être obtenu par une « randomisation conditionnelle » ou approche par permutation (Anselin, 1994). On construit un test sous l'hypothèse H_0 que toutes valeurs sont distribuées aléatoirement dans l'espace. La valeur observée de Z_i en i est fixée et les autres valeurs sont permutées aléatoirement sur toutes les zones avec une probabilité égale. Dans l'algorithme de calcul, chacune des données est tirée aléatoirement et sans remise de la population

des zones. Un test de signification p peut être obtenu en calculant la proportion de résultats des permutations qui donnent des valeurs de I_i supérieures, inférieures ou égales à la valeur observée de I_i.

L'interprétation de l'indicateur local de Moran est similaire à celle de la statistique G. Une valeur faible de p ($p < 0.10$ par exemple) indique que cette observation i est associée à des valeurs relativement fortes dans les zones voisines. Une valeur élevée de p (p > 0.9) indique que l'observation i est associée à des valeurs voisines relativement faibles.

Pour comparer l'indicateur de Moran avec la statistique G, on montre que le test local de Moran est exactement le même que le test G sous l'hypothèse conditionnelle que l'observation i est non aléatoire. L'indicateur local de Moran est alors une transformation linéaire de la statistique G_i. :

$$I_i = Z_i \sum_{j \neq i}^{n} w_{ij} Z_j = \frac{(x_i - \bar{x})(\sum_{j \neq i}^{n} x_{ij})}{\sum_{j \neq i}^{n} c_{ij} S_x^2} \cdot \frac{\sum_{j \neq i}^{n} c_{ij} x_j}{\sum_{j \neq i}^{n} x_{ij}} - \frac{(x_i - \bar{x})\bar{x}}{S_x^2} = a_i G_i + b_i \text{ , avec :}$$

$$S_x^2 = \frac{1}{n} \sum_{i}^{n} (x_i - \bar{x})^2 \text{ , } a_i = \frac{(x_i - \bar{x})(\sum_{j \neq i}^{n} x_{ij})}{\sum_{j \neq i}^{n} c_{ij} S_x^2}, \quad b_i = -\frac{(x_i - \bar{x})\bar{x}}{S_x^2}, \quad w_{ij} = \frac{c_{ij}}{\sum_{j \neq i}^{n} c_{ij}},$$

et C_{ij}=1 si j et i sont voisins et 0 autrement. L'espérance mathématique et la variance de I_i sont alors :

$E(I_i) = a_i E(G_i) + b_i$, $Var(I_i) = a_i^2 Var(G_i)$

$$\text{et : } Z(I_i) = \frac{I_i - E(I_i)}{\sqrt{Var(I_i)}} = \frac{a_i G_i + b_i - a_i E(G_i) - b_i}{\sqrt{a_i^2 Var(G_i)}} = \frac{G_i - E(G_i)}{\sqrt{Var(G_i)}} = Z(G_i)$$

L'indicateur global I de Moran peut être retrouvé en calculant la moyenne des I_i de l'indicateur local :

$$I = \frac{\sum_{i=1, j=1, i \neq j}^{n} w_{ij} Z_i Z_j}{S^2 \sum_{i=1, j=1, i \neq j}^{n} w_{ij}} = \frac{1}{n} \sum_{i=1}^{n} \left(Z_i \sum_{j=1, j \neq i}^{n} w_{ij} Z_j \right) = \frac{1}{n} \sum_{i=1}^{n} I_i(d), \text{ Les } Z_i \text{ sont standardisés, la}$$

matrice w_{ij} est standardisée en ligne, $S^2 = \frac{1}{n} \sum_{i=1}^{2} Z^2 = 1$ et $\sum_{i=1, j=1, i \neq j}^{n} w_{ij} = n$.

Une statistique locale de Geary pour chaque observation i peut aussi être définie (Anselin 1994) : $C_i(d) = \sum_{j \neq i}^{n} w_{ij} (Z_i - Z_j)^2$. Le calcul du niveau de pseudo - signification de la valeur du p–test est identique à celui de l'indicateur local de Moran.

Une grande valeur de p indique une petite valeur de C_i pour des valeurs extrêmes de i, qui suggère une association spatiale positive (similarité) de l'observation i avec

les observations voisines. Une petite valeur de p indique une valeur élevée de C_i pour des valeurs extrêmes, qui suggère une association spatiale négative (dissemblance) de l'observation *i* avec les observations voisines.

1.2.3.3 Formes générales de *G* et des tests locaux de Moran et de Geary (Shuming Bao 1996).

Pour l'indicateur G et les tests locaux de Moran et Geary, on fait l'hypothèse que les valeurs observées sont distribuées de façon aléatoire dans l'espace sous l'hypothèse H_o d'absence d'autocorrélation spatiale (Anselin (1994)). Chaque valeur (x_i) a une probabilité égale de réalisation pour chacune des n zones de l'espace : $(p(x_i)=1/n)$.

Dans de nombreux cas concrets, l'unité spatiale (zone de collecte) n'est pas définie spatialement de façon homogène. Ainsi, par définition, les zones élémentaires de diffusion du Census américain sont peuplées de 4000 habitants. La surface de collecte varie ainsi fortement entre les zones urbaines et les zones rurales. Dans le modèle spatial de population distribué sur l'espace, il est raisonnable de supposer que la probabilité de la densité de population observée au lieu *i* est proportionnelle à la surface de la zone : $\left(p(x_i) = \frac{a_i}{\sum_j a_j} \right)$. Dans ce cas, les propriétés statistiques de *G* et des indicateurs locaux de Moran et de Geary définis ci-dessus ne sont plus adaptées. En particulier, les tests de permutation du type méthode de Monte-Carlo sont inadaptés si chaque observation est tirée avec une probabilité égale.

Shuming Bao a proposé en 1996 une généralisation de $G_i(d)$ et des tests de Moran et Geary du type :

$$G_i^*(d) = \frac{\sum_{j,j\neq i}^{n} w_{ij} p_j x_j}{\sum_{j,j\neq i}^{n} p_j x_j}, \quad E(Gi^*) = \sum_{j,j\neq i}^{n} w_{ij} p_j ,$$

$$Var(G_i^*(d)) = \frac{1}{Y_{i1}^2}[W_{i2}Y_{i2} + (W_{i1}^2 - W_{i2})Y_{i3}] - W_{i1}^2$$

où w_{ij} est la matrice spatiale de pondération binaire, p_j est la probabilité conditionnelle $(P(X = x_j \mid X \neq x_i)$

$$W_{i1} = \sum_{j,j\neq i}^{n} w_{ij} p_j, \quad W_{i2} = \sum_{j,j\neq i}^{n} w_{ij}^2 p_j^2, \quad b_j = p_j/(1-p_j), \quad Y_{i1} = \sum_{j\neq i}^{n} p_j x_j, \quad Y_{i2} = \sum_{j\neq i}^{n} p_j x_j^2 ,$$

et : $Y_{i3} = \sum_{j\neq i}^{n} p_j x_j \sum_{j\neq i}^{n} b_j x_j - \sum_{j\neq i}^{n} p_j b_j x_j^2$.

La forme générale pour les indicateurs locaux de Moran et Geary est :

$I_i^*(d) = Z_i \sum_{j,j\neq i}^{n} d_{ij} Z_j$.

Où Z_i est une série d'observations standardisées, $\{d_{ij}\}$ est la matrice spatiale de poids standardisée en ligne :

$(d_{ij} = w_{ij}p_j / \sum_{j,j\neq i}^{n} w_{ij}p_j),\ E(I_i^*(d)) = \sum_{j,j\neq i}^{n} d_{ij}E(Z_iZ_j) = Z_iD_{i1}U_{i1}$,

$Var(I_i^*) = Z_i^2\left[D_{i2}U_{i2} + (D_{i1}^2 - D_{i2})U_{i3}\right] - (D_{i1}Z_iU_{i1})^2$. avec : $p_j = P(X = x_j \mid X \neq x_i)$,

$D_{i1} = \sum_{j\neq i}^{n} d_{ij} = 1$, $D_{i2} = \sum_{j\neq i}^{n} d_{ij}^2$, $b_j = \frac{p_j}{1-p_j}$, $U_{i1} = \sum_{j\neq i}^{n} p_jZ_j$, $U_{i2} = \sum_{j\neq i}^{n} p_jZ_j^2$,

$U_{i3} = \sum_{j\neq i}^{n} p_jZ_j \sum_{j\neq i}^{n} b_jZ_j - \sum_{j\neq i}^{n} p_jb_jZ_j^2$.

L'indicateur local généralisé de Geary est alors :

$C_i^*(d) = \sum_{j,j\neq i}^{n} d_{ij}(Z_i - Z_j)^2$, $E(C_i^*) = \sum_{j\neq i}^{n} d_{ij}E(Z_i - Z_j)^2 = D_{i1}V_{i2}$,

$Var(C_i^*) = D_{i2}V_{i4} + (D_{i1}^2 - D_{i2})V_{i3} - D_{i1}^2V_{i2}^2$.

où, w_{ij} et p_j sont définis par : $D_{i1} = \sum_{j\neq i}^{n} d_{ij}$, $D_{i2} = \sum_{j\neq i}^{n} d_{ij}^2$, $b_j = \frac{p_j}{1-p_j}$,

$V_{i2} = \sum_{j\neq i}^{n} p_j(Z_i - Z_j)^2$,

$V_{i3} = \sum_{j\neq i}^{n} p_j(Z_i - Z_j)^2 \sum_{j\neq i}^{n} b_j(Z_i - Z_j)^2 - \sum_{j\neq i}^{n} p_jb_j(Z_i - Z_j)^4$, $V_{i4} = \sum_{j\neq i}^{n} p_j(Z_i - Z_j)^4$.

L'indicateur généralisé G et les indicateurs locaux de I de Moran et c de Geary sont identiques à leur formes initiales quand chaque probabilité conditionnelle est identique : ($P(X = x_j \mid X \neq x_i)$, $j = 1....n$) (par exemple : $p_1 = p_2 = ... = p_n$).

Comparées aux indicateurs standards, les formes généralisées rendent compte de la distribution spatiale réelle, sans déformation, en y incorporant la probabilité conditionnelle : $p_j = P(X = x_j \mid X \neq x_i) = a_j / \sum_{j\neq i}^{n} a_k$.

2. Analyse géostatistique.

Le champ de la géostatistique recouvre un ensemble de méthodes de traitement statistique des données spatiales issues des Géo-sciences.

Trois étapes jalonnent l'évolution de cette discipline :
● Dans les années 1950 à 1960 : méthodes d'estimation pour l'industrie minière, géologues de l'école sud-africaine, Krige, De Wijs.
● Dans les années 1960 à 1980, formalisation théorique par G. Matheron de L'Ecole de Mines de Fontainebleau.
● Depuis, sous l'impulsion de chercheurs comme Noël Cressie (1993), le champ d'utilisation des méthodes de géostatistique s'est étendu à de nouvelles disciplines : écologie, épidémiologie, météorologie, astronomie, démographie, économie.

2.1 Concepts de géostatistique.

A partir de Hans Wackernagel (1998), on peut fixer à grands traits le cadre de travail de la géostatistique. Le concept de ***variable régionalisée*** constitue la base de la démarche. Le phénomène étudié prend des valeurs dans l'espace. On considère ce phénomène réel comme une fonction $z(x)$ qui dépend de la position de x dans l'espace. Cette fonction s'appelle une variable régionalisée qui est une fonction numérique. On considère la variable régionalisée $z(x)$ comme une réalisation d'une fonction aléatoire $Z(x)$. Les données sont considérées comme étant générées par un processus aléatoire. La moyenne de $Z(x)$ au point x est notée $m(x)$. Aux points où aucune mesure n'a été réalisée, les valeurs $z(x)$ sont inconnues mais définies. C'est la réalisation des variables aléatoires correspondantes $Z(x)$. Faire des statistiques à partir d'une réalisation unique d'une fonction aléatoire nécessite des hypothèses de stationnarité que l'on peut définir de façon intuitive : la variable $z(x)$ a une moyenne et une variance qui ne dépendent pas de la position dans l'espace. Si $Z(x)$ est stationnaire alors :

❶ l'espérance mathématique $E[Z(x)]$ est invariante par translation, elle est constante dans l'espace.

❷ la covariance $Cov[Z(x),Z(y)]$ est invariante par translation. Elle ne dépend que de la position relative de x et de y et non de leur position absolue.
$Cov[Z(x),Z(y)]$ est une fonction de $h = x - y$.

La différence de la variable entre deux points ne dépend que du vecteur reliant ces deux points, alors : la moyenne $E[Z(x+h) - Z(x)] = 0$ et la variance $E[Z(x+\vec{h})\ Z(x)] = 2\gamma(\vec{h})$. $2\gamma(\vec{h})$ est le variogramme.

La géostatistique utilise deux outils principaux : le semivariogramme, description de la variabilité moyenne d'un phénomène dans l'espace, le Krigeage, méthode d'estimation, du type régression, des valeurs d'un phénomène en tout point.

2.2 Le semivariogramme.

Le semivariogramme peut être estimé par la variance statistique de la différence entre 2 points :

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} (z(x_i + h) - z(x_i))^2$$

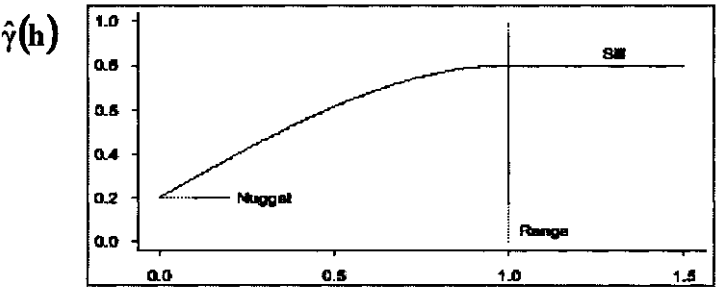
avec : $N(h)$ = nombre de points séparés par la distance h,
$z(x_i + h)$ et $z(x_i)$ = valeurs du couple d'échantillon de la même variable aléatoire séparés par la distance h.

Le modèle de semivariogramme $\gamma(h)$ est une fonction caractérisant la corrélation spatiale de la fonction aléatoire sous-jacente. Il permet d'interpoler entre les données en accord avec la corrélation spatiale et de calculer la variance de l'erreur d'estimation. Il doit enfin ajuster tous les variogrammes expérimentaux dans les différentes directions.

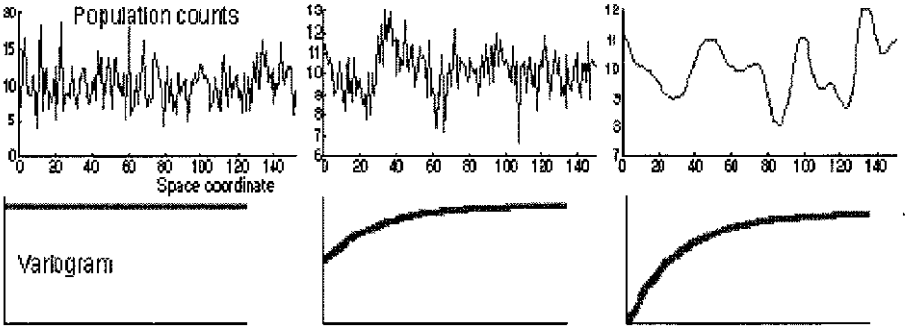
Cressie propose une estimation robuste peu sensible aux valeurs aberrantes qui est retenue dans la procédure VARIOGRAM de **SAS®** Version 8 :

$$\hat{\gamma}(h) = \frac{\left\{ \frac{1}{N(h)} \sum_{i=1}^{N(h)} (z(x_i + h) - z(x_i))^{1/2} \right\}^4}{.914 + [.988 / N(h)]}.$$

Le graphique semivariogramme * distance fait apparaître trois valeurs caractéristiques : Le palier ou filon (sill), la portée ou zone d'influence (range), la pépite (nugget).



Formes caractéristiques du semivariogramme :

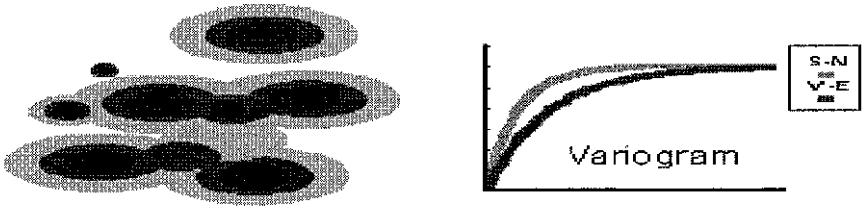


Alexei Sharov : Population Ecology1996

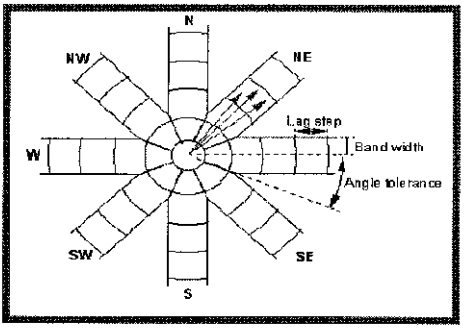
Effet pépite pur, pas d'autocorrélation spatiale	Combinaison autocorrélation spatiale et effet aléatoire	Forte autocorrélation spatiale sans effet aléatoire

Le variogramme est le plus souvent calculé pour tout couple de points dans des directions privilégiées Nord-Sud et Est-Ouest par exemple (variogramme directionnel). Il révèle alors souvent une anisotropie plus ou moins marquée :

(Alexei Sharov : Population Ecology1996)

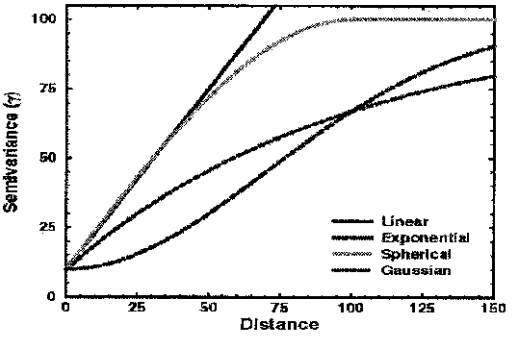


La procédure VARIOGRAM de SAS® Version 8 permet de contrôler les différents paramètres de retards et de direction :



(Alexei Sharov : Population Ecology1996)

Le variogramme peut revêtir différentes formes paramétriques, fonctions de la nature isotropique ou anisotropique du modèle :



D.L Urban Landscape Ecology feb 1999

Trois exemples de modèles stationnaires :

Modèle exponentiel : $\gamma(h) = C_0 + C\left[1 - e^{-h/A_0}\right]$,

Modèle sphérique :
$\gamma(h) = C_0 + C\left[1.5(h/A_0) - 0.5(h/A_0)^3\right]$ *si* $h \le A_0$ *et* $\gamma(h) = C_0 + C$ *si* $h > A_0$

Modèle gaussien : $\gamma(h) = C_0 + C\left[1 - e^{\left(-h^2/A_0\right)}\right]$.

Un exemple de modèle non stationnaire, le modèle linéaire : $\gamma(h) = C_0 + [h(C/A_0)]$
Avec : C_0 ordonnée à l'origine, pépite, $C_0 + C$ portée et A_0 palier.

2.3 Le krigeage.

Le terme krigeage trouve son origine dans le patronyme de D.G. Krige, géologue sud-africain qui a mis au point dans les années 50 des méthodes empiriques d'évaluation de la teneur en minerai globale à partir d'un nombre limité de sondages.

La méthode initiale de Krige est connue sous le nom de krigeage ordinaire ou ponctuel. Elle a connu de nombreuses extensions.

Le krigeage est un ensemble de méthodes d'estimation, interpolation et extrapolation. Le krigeage est une combinaison linéaire des données tenant compte du nombre et de la configuration des données, de la position des données par rapport au point à estimer, de la structure spatiale de la variable (variogramme).

Il existe de nombreux types de krigeages, les principaux sont :

- le krigeage ponctuel, les points inconnus sont estimés à partir des points connus
- le krigeage par blocs, on évalue non un point mais la moyenne d'un ensemble de points, etc...

Exemple : le krigeage ponctuel : la valeur de la variable $\hat{Z}(s_0)$ est estimée par la moyenne des valeurs d'un échantillon de n variables environnantes connues. Le krigeage est le meilleur estimateur linéaire sans biais (BLUE).

C'est une combinaison linéaire des données : $\hat{z}(s_0) = \sum_{i=1}^{n} w_i z(s_i)$, avec $\sum_{i=1}^{n} w_i = 1$; sans biais, $E\left[\hat{Z}(s_0)\right] = E\left[Z(s_0)\right]$, qui minimise au mieux la variance de l'estimateur local, $Var\left[\hat{Z}(s_0) - Z(s_0)\right]$. Les poids sont estimés en résolvant le système matriciel

$$\begin{bmatrix} \gamma(h_{11}) & \gamma(h_{12}) & \dots & \gamma(h_{1m}) & 1 \\ \gamma(h_{21}) & \gamma(h_{22}) & \dots & \gamma(h_{2m}) & 1 \\ \dots & \dots & \dots & \dots & \dots \\ \gamma(h_{m1}) & \gamma(h_{m2}) & \dots & \gamma(h_{mm}) & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix} * \begin{bmatrix} W_1 \\ W_2 \\ \dots \\ W_m \\ \lambda \end{bmatrix} = \begin{bmatrix} \gamma(h_{1p}) \\ \gamma(h_{2p}) \\ \dots \\ \gamma(h_{mp}) \\ 1 \end{bmatrix}$$

$$\Gamma * W = \hat{\Gamma}, \quad W = \Gamma^{-1} * \hat{\Gamma}.$$

Les $\gamma(h_{ik})$ sont les valeurs du semivariogramme qui correspondent aux distances h_{ik} entre les points x_i et x_k. Pour que la solution soit non-biaisée, la somme des poids, les W_i, doit être égale à 1. Cette dernière contrainte introduit un degré de liberté supplémentaire dans le problème. Ce degré supplémentaire est utilisé en ajoutant une variable libre, λ (un multiplicateur de Lagrange), dans le but de minimiser l'erreur d'estimation.

La variance de l'estimation de krigeage est égale à : $\delta_z^2 = \lambda + \sum_{i=1}^{n} w_i \gamma(h_{mp})$.

L'estimation des poids étant réalisée, nous pouvons évaluer les points $\hat{z}(s_0)$ a partir de $\hat{z}(s_0) = \sum_{i=1}^{n} w_i z(s_i)$. Quand ces valeurs sont estimées pour tous les points d'une grille régulière, nous obtenons une surface de densité de population que nous pouvons cartographier (lissage). La procédure KRIGE2D de SAS® Version 8 : permet une mise en œuvre relativement aisée de cette méthode d'estimation.

3. Econométrie spatiale.

Deux problèmes surgissent lorsque l'on désire modéliser des données géoréférencées :
1 / L'existence d'une dépendance entre les variables géoréférencées, autocorrélation spatiale.
2 / L'hétérogénéité spatiale qui différencie les comportements en fonction des localisations.

Ces caractéristiques de dépendance et d'hétérogénéité spatiale rendent en particulier inefficace la méthode des moindre carrés ordinaires (OLS) : les estimateurs ne sont pas convergents en présence de variables endogènes décalées, ils sont inefficients en présence d'autocorrélation des erreurs.

3.1 Principaux types de modèles spatiaux.

Soit une forme simple de modèle spatial autorégressif : $y = \rho Wy + \varepsilon$, avec y variable observée sur le domaine spatial $D : \{Y(s_i) : s_i \in D, i = 1...n\}$, W matrice (n,n) de contiguïté, ρ le paramètre spatial autorégressif et les résidus $\varepsilon \approx N(0, \delta^2)$. Les MCO estimés sont biaisés et inconsistants :

$$\hat{\rho} = \left[(Wy)'(Wy)\right]^{-1}(Wy)'y = \rho + \left[(Wy)'(Wy)\right]^{-1}(Wy)'\varepsilon \ , \ E(\hat{\rho}) \neq \rho \ .$$

La littérature récente (Anselin 1988 et 1993, Lesage 1999) distingue trois familles de spécifications dérivées de la forme simple :

3.1.1 Modèle spatial autorégressif simple.

C'est un modèle purement autorégressif où la variable explicative est le décalage spatial de la variable dépendante : $y = \alpha + \rho Wy + \varepsilon$, ou encore, $(y - \alpha) = \rho W(y - \alpha) + \varepsilon$. Avec, α, terme constant, ρ, coefficient spatial autorégressif, Wy, décalage spatial de y et ε terme d'erreurs classique, indépendance et distribution identiquement normale, avec une moyenne nulle et une variance constante. Si le terme d'erreurs ε est corrélé avec la variable explicative Wy, le cas le plus fréquent, l'estimateur MCO est biaisé et inefficace.

3.1.2 Modèle spatial autorégressif général.

Dans ce modèle, les variables explicatives incluent un décalage spatial de la variable dépendante en plus de l'ensemble des variables exogènes : $y = \rho Wy + X\beta + \varepsilon$, X vecteur des variables explicatives que l'on suppose non corrélées avec le terme d'erreurs, ρ, coefficient spatial autorégressif, Wy, décalage spatial de y, β vecteur $(K,1)$ des coefficients de régression, X matrice (N,K) des variables exogènes et ε terme d'erreurs classique, indépendance et distribution identiquement normale, avec une moyenne nulle et une variance constante. Ici encore, la méthode des MCO n'est pas valide du fait de la corrélation spatiale entre le décalage Wy et ε.

3.1.3 Modèle spatial autorégressif avec autocorrélation spatiale des erreurs.

Dans ce modèle, les variables explicatives se composent seulement des variables exogènes mais le terme d'erreurs suit un processus autorégressif : $y = X\beta + \varepsilon$ et $\varepsilon = \lambda W\varepsilon + \mu$ avec μ vecteur purement aléatoire (bruit blanc). Une forme voisine de ce modèle est : $(y - X\beta) = \lambda W(y - X\beta) + \varepsilon$. En présence de dépendance spatiale des erreurs, l'estimation MCO est non biaisée mais inefficace. Ce modèle peut se généraliser en modèle autorégressif- autorégressif SARAR (Anselin 1988). Il inclut une variable endogène décalée et une autocorrélation des erreurs : $y = \rho W_1 y + X\beta + \varepsilon$ et $\varepsilon = \lambda W_2 \varepsilon + u$, avec W_1 matrice de contiguïté de la variable endogène et W_2 matrice de contiguïté des erreurs. Si $W_1 = W_2$ le modèle se simplifie en : $y = (\rho + \lambda)Wy - \rho\lambda W^2 y + X\beta + \mu$.

3.2 Estimation des modèles spatiaux.

La méthode MCO est doublement inadaptée à la modélisation spatiale. Les éléments de la variable endogène sont corrélés à ceux des erreurs et les paramètres du modèle ne peuvent être estimés d'une façon convergente par les MCO. Remarquons que, dans les modèles temporels autorégressifs, les estimateurs MCO restent convergents en présence de variables retardées si les erreurs ne sont pas corrélées. Les estimateurs MCO sont sans biais mais inefficaces car les erreurs ne sont pas homoscédastiques. L'utilisation de la méthode des moindres carrés quasi-généralisés ou celle des doubles moindres carrés n'est pas non plus convergente, d'ou le sous-titre de l'article de Kelejian et Prucha (1997) : a « serious problem ».

Anselin et LeSage et de nombreux auteurs par la suite préconisent la méthode d'estimation du maximum de vraisemblance (MLE).

En pratique, l'estimation peut s'effectuer à partir de la maximisation de la fonction de log-vraisemblance complète à l'aide de techniques d'optimisation non-linéaires. Le principe de la méthode est de résoudre une partie des équations associées aux conditions du premier ordre et d'introduire ensuite les solutions obtenues par la fonction de log-vraisemblance :

$L = -(n/2)ln(2\pi) - (1/2)ln|\Omega| + ln|A| + ln|B| - (1/2)v'v$ avec :

$v'v = (Ay - X\beta)B\Omega^{-1}B(Ay - X\beta)$ où, $A = I - \rho W_1$, $B = I - \rho W_2$, Ω est la matrice de covariance des erreurs et $v'v$ est la somme des carrés d'une forme particulière de transformation du terme d'erreurs.

La fonction du maximum de vraisemblance pour le modèle spatial autorégressif simple (3.1.1) est déduite de $(y - \alpha) = \rho W(y - \alpha) + \varepsilon$, si, $X = 1(n * 1)$, $A = I - \rho W$, $B = I$, $\beta = \alpha$ et $\Omega_{ii} = \delta^2(\Omega_{ij}) = 0$ *pour* $i \neq j$, la fonction de log-vraisemblance est :

$L = -(n-2)ln(2\pi) - (n/2)ln\delta^2 + ln|A| - (1/2\delta^2)(Ay - \alpha)'(Ay - \alpha)$

Pour le modèle spatial autorégressif général (3.1.2), la fonction du maximum de vraisemblance est déduite de $y = \rho Wy + X\beta + \varepsilon$ avec, $A = I - \rho W$, $B = I$, et $\Omega_{ii} = \delta^2(\Omega_{ij}) = 0$ *pour* $i \neq j$. La fonction de log-vraisemblance est :

$L = -(n-2)ln(2\pi) - (n/2)ln\delta^2 + ln|A| - (1/2\delta^2)(Ay - X\beta)'(Ay - X\beta)$.

Enfin pour un modèle spatial autorégressif avec autocorrélation spatiale des erreurs (3.1.3), la fonction du maximum de vraisemblance est déduite de

$(y - X\beta) = \lambda W(y - X\beta) + \varepsilon$

avec : $A = I$, $B = I - \lambda W$ et, $\Omega_{ii} = \delta^2(\Omega_{ij}) = 0$ *pour* $i \neq j$.

La fonction de log-vraisemblance est :

$L = -(n-2)ln(2\pi) - (n/2)ln\delta^2 + ln|B| - (1/2\delta^2)(y - X\beta)'B'(By - X\beta)$

La très riche, et très complexe, procédure MIXED de SAS® Version 8 permet aux esprits bricoleurs d'aborder pratiquement ces problèmes. Toutefois, certaines scènes explicites de joyeux bidouillages d'économétrie spatiale peuvent heurter gravement la sensibilité des théoriciens. On consultera avec profit le chapitre 9 de SAS® System for mixed models, SAS institute Inc 1996 ; la bibliothèque de programme en MATLAB® de LeSage www.econ.utoledo.edu/faculty/lesage/lesage.html et celle en SpaceStat® de Anselin luc@spacestat.com.

3.3 Tests des modèles spatiaux de régression.

La littérature (Anselin, LeSage, R Haining 1993, G Arbia 1996) propose deux familles de tests de diagnostic : les tests de dépendance spatiale (autocorrélation) et les test d'hétérogénéité, hétéroscédasticité en présence de dépendance spatiale. Dans les modèles spatiaux, les tests de dépendance et d'hétérogénéité dépendent de la nature de la matrice de pondération retenue.

3.3.1 Autocorrélation spatiale des erreurs.

3.3.1.1 Application du test de Moran au vecteur du terme des erreurs.

Ce test (Cliff et Ord 1981) est de la forme : $I = e'We / e'e$ avec e vecteur des résidus des MCO et W matrice spatiale de pondération. Pour un nombre de zones suffisamment élevé, entre 30 et 60 suivant le type de matrice, I tend asymptotiquement vers une loi normale : $N(\mu_I, \sigma_I^2)$. Ce test ne permet pas de discriminer l'autocorrélation des erreurs et l'autocorrélation des variables explicatives endogènes.

3.3.1.2. Test du multiplicateur de Lagrange.

Le test de dépendance spatial des erreurs du multiplicateur de Lagrange (cf Anselin, LeSage, Arbia) est du type :

$LM(err) = \left\{e'We / \delta^2\right\}^2 / tr\left[W'W + W^2\right] \approx \chi^2(1)$

avec : tr trace de la matrice, δ^2 estimation du maximum de vraisemblance pour la variance des erreurs $\left(\delta^2 = e'e / n\right)$. La log vraisemblance $LM(err)$ suit asymptotiquement un $\chi^2(1)$ sous l'hypothèse nulle d'absence d'autocorrélation spatiale : $(H_0 : \lambda = 0)$.

Anselin (1988) propose un autre test du type multiplicateur de Lagrange pour l'autocorrélation d'une variable endogène décalée :

$LM(lag) = \left\{e'Wy / \delta^2\right\}^2 / \left\{(WXb)'MWXb / \delta^2 + tr\left[W'W + W^2\right]\right\}$

avec Wy matrice de décalage spatial, b vecteur des MCO pour le paramètre β, $M = I - X\left(X'X\right)^{-1}X'$. La log vraisemblance $LM(lag)$ suit aussi asymptotiquement un $\chi^2(1)$ sous l'hypothèse nulle d'absence d'autocorrélation spatiale : $(H_0 : \rho = 0)$.

3.3.2 Test d'hétérogénéité spatiale, généralisation du test de Chow.

La présence d'autocorrélation dans les modèles spatiaux rend inopérants les tests classiques d'hétéroscédasticité. Le plus populaire des tests d'hétérogénéité ou de robustesse structurelle est le test de Chow. L'analogue en présence de dépendance spatiale peut se formuler : $H_0 : y = X\beta + \varepsilon; H_1 = \begin{bmatrix} X_i & 0 \\ 0 & X_j \end{bmatrix} \begin{bmatrix} \beta_i \\ \beta_j \end{bmatrix} + \varepsilon$

avec : $X_i(n^* k)$ et $X_j(n^* k)$ sous-ensemble des observations de la variable indépendante et β_i, β_j les coefficients de régression. Le terme d'erreur ε suit un processus spatial autorégressif, le test spatial de Chow est :

$C = \left\{ e_R'(I - \lambda W)'(I - \lambda W)e_R - e_U'(I - \lambda W)'(I - \lambda W)e_U \right\} / \delta^2$, avec : λ, estimation du maximum de vraisemblance pour le paramètre spatial, e_R, e_U résidus de la régression, et σ^2, l'estimation de l'erreur de la variance pour chacun des sous-modèles, le modèle complet ou les deux. Ce test suit une asymptotiquement un $\chi^2(k)$ sous l'hypothèse H_0 de non hétérogénéité spatiale : $\beta_i = \beta_j$.

Bibliographie sommaire :

L Anselin (1988). Spatial econometrics : methods and models. Kluwer academic publishers.

Giuseppe Arbia (1996). Analisi Econometrica di dati spaziali :
Universita D'Annunzio : www.dmqte.unich.it/users/arbia/econom.htm

Shuming Bao, Mark S Henry (1996). Heterogeneity issues in local measurement of spatial association Geographical Systems 1996 Vol 3 : 1-13.

Cliff , Ord (1981). Spatial processes : Model and applications. Pion, London.

Noël Cressie (1993). Statistics for spatial data, John Wiley & Sons, inc.

R Haining (1993). Spatial data analysis in the social and environmental sciences Cambridge university press.

H.H Kelejian et I.R Prucha (1997). Estimation of spatial régression models with autogressive errors by two-stage least squares procédures : a serious problem in International Regional Science rewiew.

J. Lesage. Spatial econometrics in Matlab® 1999 University of Toledo
www.econ.utoledo.edu/faculty/lesage/lesage.html

SAS® System for mixed models, SAS institute Inc 1996

Hans Wackernagel (1998). Multivariate Geostatistics, Springer