Econométrie et données spatiales

Une introduction à la pratique

Hubert JAYET

Cahiers d'économie et sociologie rurales, n° 58-59, 2001

Hubert JAYET*

Econometrics on spatial data: a beginner's guide

Summary *– This contribution is an introduction to the main topics of spatial econometrics. We start analyzing the main problems raised by spatial data. The first one is heterogeneity: statisticians must take account of the fact that spatial units may not be directly comparable. They must correct for differences in size, form, structure and so on. The second one is interaction among units located in space, the intensity of which decreases with distance. These interactions lead to spatial autoregression and spatial autocorrelation. Then, the paper introduces to the main instruments used to represent and analyze spatial autocorrelation and autoregression: spatial graphs, weight matrices, contiguity matrices. It presents the main tests used to detect spatial autocorrelation, color tests on qualitative data, Moran and Geary tests for quantitative data. It shows how these tests can be interpreted. An illustrative example is also provided. Last, the paper shows how to deal with spatial autocorrelation and autoregression on the example of linear models. The main types of spatial linear models are presented: spatially autoregressive, spatially autocorrelated and their combination. Then, we explain why least squares methods are not well suited to estimate this type of models. Most often, econometric analysis will rest upon maximum likelihood methods. The paper shows how to use these methods in the specific context of spatial models, in order to find parameters estimates and to make tests on them.*

Key-words:
econometrics, spatial data, autocorrelation

Econométrie et données spatiales: une introduction à la pratique

Résumé – Cet article d'initiation à l'économétrie sur données spatiales met l'accent sur les principaux problèmes rencontrés dans l'utilisation de ces données: hétérogénéité des observations, interactions liées à la proximité. La présence de ces dernières conduit à s'intéresser à l'autocorrélation spatiale. L'article montre comment la représenter en pratique. Il montre ensuite comment en tester la présence dans les données. Enfin, il présente les principaux modèles linéaires qui en tiennent compte et leurs procédures d'estimation.

Mots-clés:
économétrie, données spatiales, autocorrélation

* *Université des Sciences et Technologies de Lille, Laboratoire des mécanismes économiques et dynamiques des espaces européens, 59655 Villeneuve d'Ascq cedex.*
e-mail: jayet@pop.univ-lille1.fr

Cet article reprend, pour l'essentiel, une présentation à l'Ecole-Chercheurs du Croisic. Je remercie Virginie Piguet et Mohamed Hillal, de l'unité INRA ESR de Dijon, pour l'aide technique apportée sur certains points de l'article.

MANIPULER des données spatiales pose des problèmes auxquels le praticien doit prendre garde. Ces problèmes ne sont pas nécessairement spécifiques à la nature spatiale des données utilisées. On peut les rencontrer dans d'autres domaines. Mais, avec des données spatialisées, ils se posent avec une acuité particulière. De ce fait, il est important de les avoir toujours à l'esprit. C'est l'objectif de cet article d'en préciser la nature, avec le regard du praticien plus que du théoricien, et de donner quelques indications sur la manière d'aborder les données quand elles ont une dimension spatiale. Le lecteur ayant besoin d'aller plus loin pourra consulter Jayet (1993), Anselin (1988), Cliff et Ord (1981).

L'HÉTÉROGÉNÉITÉ DES DONNÉES SPATIALES

Précisons tout d'abord qu'en général la spatialisation d'une donnée statistique revêt deux formes principales. D'une part, les informations recueillies peuvent porter sur des points particuliers répartis dans l'espace. C'est par exemple le cas quand, à l'échelle nationale, on travaille sur les grandes villes françaises. On parlera alors de données ponctuelles. D'autre part, ces mêmes informations peuvent être des agrégats, des moyennes ou des taux relatifs à un ensemble de zones : les 22 régions françaises, les 96 départements métropolitains, les 341 zones d'emploi, etc. On parlera alors de données de zones.

Dans les deux cas, l'analyse d'une information spatialisée pose d'abord un problème d'hétérogénéité. En effet, toute analyse statistique d'une population suppose que les éléments de cette population ont des points communs, sur lesquels on peut fonder des comparaisons et asseoir des régularités. Or, qu'il s'agisse d'entités ponctuelles ou de zones, les unités spatiales sont généralement fortement hétérogènes, au moins par leur taille, leur forme et leur structure[(1)].

L'hétérogénéité de taille est la première qui apparaît. Peut-on comparer, sans précaution, l'agglomération parisienne et sa dizaine de millions d'habitants, et une petite ville de quelques milliers d'habitants ? Le département du Nord avec ses 2,5 millions d'habitants et celui de la Corse du Sud, qui dépasse à peine les 100 000 habitants ? Cette hétérogénéité se manifeste au moins dans l'ordre de grandeur des agrégats, beaucoup d'entre eux étant d'autant plus élevés que la taille de l'entité est grande.

[(1)] Ces problèmes d'hétérogénéité sont analysés de manière très détaillée par G. Arbia (1989).

Ce problème d'hétérogénéité de taille n'est pas spécifique aux données spatiales. Ces dernières se distinguent par le fait que le problème se pose de manière quasi systématique[2], en particulier avec les données utilisées par les sciences humaines. Les spécialistes des données temporelles, qui travaillent en général sur des périodes de durée approximativement égale (le jour, la semaine, le mois, l'année), le rencontrent peu. Mais il se pose parfois, les conduisant à des pratiques comme la correction des jours ouvrables. De même, les statisticiens utilisant des données individuelles travaillent sur des entités qui, issues d'une même population, sont généralement considérées *a priori* comme directement comparables. Il n'empêche que, quand les individus sont des ménages, des problèmes de différences de taille se posent, justifiant l'usage des unités de consommation. De même, quand les entités sont des firmes, il faut rendre compte de leurs différences de taille.

Dans tous les cas, le statisticien tiendra compte des différences de taille en choisissant un indicateur de dimension (le nombre de jours ouvrables, le nombre d'unités de consommation, la population ou la superficie de la zone), et en utilisant les taux plutôt que les agrégats comme variables soumises à l'analyse statistique : le taux de chômage plutôt que le nombre de chômeurs, la valeur ajoutée par tête plutôt que la valeur ajoutée globale. Il faut être bien conscient du fait qu'il peut y avoir plusieurs indicateurs de taille : superficie, population totale, population active, emploi, nombre de logements, etc. Dans certains cas, l'un d'entre eux s'imposera *a priori*. Mais, le plus souvent, il faudra faire des tests pour trancher.

A l'hétérogénéité de taille s'ajoutent souvent des hétérogénéités de forme et de position. En effet, peut-on comparer sans précaution le département du Nord, de forme longue et étroite avec près de la moitié de sa frontière le séparant de la Belgique, avec la Sarthe, de forme presque carrée et située à l'intérieur du territoire national ? Typiquement géographique, ce type d'hétérogénéité est très difficile à appréhender, tant dans ses conséquences que pour la construction d'indicateurs permettant d'en mesurer les effets et de les corriger. Il faut cependant toujours l'avoir à l'esprit pour identifier les situations où il pourrait poser problème. On verra un exemple plus loin avec l'analyse des interactions spatiales.

Enfin, on ne saurait négliger les hétérogénéités de structure. Un exemple simple permet d'en illustrer l'importance. Peut-on, sans précaution, comparer le revenu moyen par tête à Paris et dans un département rural ? La réponse est non, car la structure de la population n'est pas la même dans les deux cas. Or, les revenus des individus sont fortement liés à la nature de l'activité qu'ils exercent et à leur niveau de qualification.

[2] Les biomètres, grands consommateurs de données spatialisées, y échappent plus facilement car ils sont souvent capables d'utiliser des données réparties de manière régulière dans l'espace, par exemple sur la base de quadrillages.

Pour déterminer l'influence effective qu'exerce une localisation particulière par rapport à d'autres, il faut comparer ce qu'un même individu obtiendrait dans les différentes localisations. Ce qui conduit à un modèle intégrant des variables de structure : dans l'exemple des revenus, on prendra en compte les structures de qualification, d'activité économique, voire de taille d'établissement.

C'est ce qui justifie la pratique, fréquente en analyse spatiale, de l'analyse *« shift-share »* des Anglo-saxons, terme qu'on traduira ici par analyse structurelle-géographique. Celle-ci revient à choisir des indicateurs de référence (dans notre exemple, les revenus moyens nationaux par tête pour chacun des niveaux de qualification) et à calculer l'effet structurel, qui est égal aux performances qu'aurait chaque zone étudiée, avec sa structure actuelle, si elle se comportait comme la zone de référence. La comparaison entre zones portera alors sur les effets géographiques, égaux à la différence entre performances actuelles et effets structurels (pour une présentation détaillée, voir Jayet, 1993).

INTERACTIONS ET PROXIMITÉ

Des observations réparties dans l'espace sont fréquemment interdépendantes : ce qui se passe dans une localisation particulière dépend de ce qui se passe dans d'autres localisations. Suivant un bon vieux principe de la géographie, ces interactions sont d'autant plus fortes que les localisations concernées sont plus proches. Le statisticien a donc besoin d'un instrument qui lui permette de représenter cette interaction entre observations et sa décroissance en fonction de la distance qui les sépare.

Cet instrument est la matrice d'interactions spatiales[(3)]. Avec N observations, on utilise une matrice carrée W à N lignes et N colonnes, dont les termes diagonaux sont nuls et dont le terme non diagonal w_{ij} est d'autant plus élevé que l'effet de l'observation j sur l'observation i est important.

Si W est une matrice d'interactions et $Y = (y_1, ..., y_N)'$ est un vecteur colonne de N observations d'une variable spatialisée, le produit matriciel WY a pour terme courant :

$$(WY)_i = \sum_j w_{ij} y_j \qquad (1)$$

Ce terme mesure l'intensité de l'effet global sur la i-ième observation des valeurs prises par la variable Y ailleurs dans l'espace. Cette variable sera utilisée ensuite dans les modèles statistiques pour représenter les effets que les localisations exercent les unes sur les autres.

(3) Dans la littérature anglo-saxonne, on trouve fréquemment le terme de *weight matrix*, matrice de poids.

Cependant, pour utiliser une matrice d'interactions dans un modèle statistique, il est nécessaire de lui donner une forme *a priori*. Il est en effet impossible, du moins dans le cadre d'un modèle purement spatial, d'estimer avec N observations chacun des $N(N-1)/2$ coefficients de la matrice d'interactions. En choisissant une définition particulière des distances entre observations et une forme fonctionnelle spécifique pour la relation entre la distance et l'intensité de l'interaction, l'économètre détermine une famille de matrices d'interactions dépendant d'un nombre faible de paramètres, dont la valeur sera déterminée par estimation.

L'exemple le plus classique est celui des matrices de contiguïté, fréquemment utilisées quand les données portent sur des zones géographiques. Deux zones sont contiguës quand elles ont une frontière commune. Plus généralement, on définit la distance de contiguïté entre deux zones comme le nombre minimal de frontières qu'il faut franchir pour aller de l'intérieur de l'une à l'intérieur de l'autre. Deux zones sont contiguës à l'ordre k quand leur distance de contiguïté est égale à k. Cette définition de la contiguïté est l'analogue spatial de la définition des retards en séries temporelles : la contiguïté d'ordre k correspond au retard d'ordre k.

Cette notion de contiguïté peut être commodément représentée par un graphe. A chaque observation spatiale (zone géographique), on associe un nœud du graphe. Les nœuds correspondant à deux zones contiguës sont reliés par un arc. La figure 1 est une bonne illustration de ce passage de la carte au graphe, pour quelques cantons de la région Rhône-Alpes.

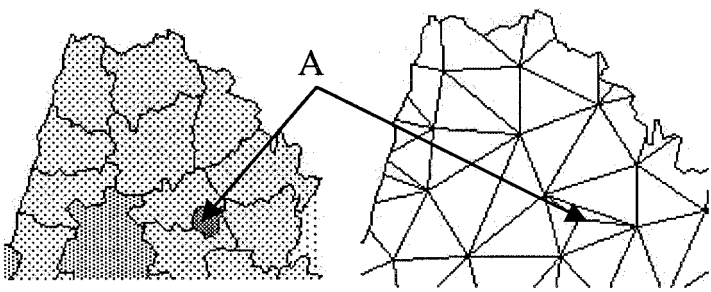


Figure 1.
Représentation de la contiguïté de quelques cantons de Rhône-Alpes

Pour un ensemble de N zones géographiques, la matrice $C^{(k)}$ de contiguïté à l'ordre k est l'analogue de l'opérateur retard. L'élément $c_{ij}^{(k)}$ de cette matrice est égal à l'unité quand les zones sont contiguës à l'ordre k, nul sinon. La matrice de contiguïté est éventuellement normalisée en divisant chacune de ses lignes par la somme de ses éléments. La matrice normalisée est donc une matrice stochastique[(4)]. On définit

[(4)] C'est-à-dire une matrice dont la somme des éléments de chaque ligne est égale à l'unité. On a $c_{ij}^{(n)} = 1/m_i^{(n)}$, où $m_i^{(n)}$ est le nombre d'observations contiguës à la zone i à l'ordre n.

alors une famille de matrices d'interactions spatiales sur la base des contiguïtés d'ordre au plus égal à K:

$$W(\rho_1, ..., \rho_K) = \sum_{k=1}^{K} \rho_k C^{(k)} \quad (2)$$

où $\rho_1, ..., \rho_K$ sont des paramètres à estimer. En pratique, on se restreint généralement au cas $K = 1$ et donc à $W(\rho) = \rho C^{(1)}$. L'interaction se restreint donc à l'influence des zones contiguës.

Les matrices de contiguïté usuelles ont cependant pour inconvénient que, lorsque $c_{ij}^{(k)}$ et $c_{il}^{(k)}$ diffèrent tous deux de zéro, $c_{ij}^{(k)} = c_{il}^{(k)}$. Ce qui signifie que toutes les observations contiguës à une zone donnée l'influencent de la même manière. On retrouve une homogénéité des interactions qui n'est pas plus vraisemblable que l'homogénéité des observations elles-mêmes. On en retrouve un exemple sur la figure 1 avec la petite zone désignée par la lettre A. Cette zone est contiguë à trois autres, ce qu'on retrouve bien sur le graphe. Cependant, avec la zone située à l'Est, la frontière commune se réduit à presque rien. Il s'en faudrait de peu pour qu'il n'y en ait pas, faisant chuter brutalement le coefficient de contiguïté à zéro.

On peut résoudre ce problème en faisant des $c_{ij}^{(k)}$ des fonctions des caractéristiques des zones j qui influencent i. Un exemple classique est l'utilisation de coefficients proportionnels à la longueur des frontières communes. On peut penser à d'autres déterminants de l'intensité des interactions: la distance entre les centres des deux zones, la taille de la zone émettrice de l'interaction, la capacité des réseaux de transport entre les zones, etc. C'est ici à l'économètre de choisir et de justifier les variables pertinentes.

Enfin, il y a de nombreuses situations où l'utilisation de la contiguïté s'avère impossible ou inadaptée. C'est en particulier le cas avec des données portant sur des points: contrairement au cas des zones, la détermination de la contiguïté entre deux points n'a guère de sens. Tout ce que l'on peut dire est que ces points sont plus ou moins éloignés les uns des autres. Même avec des données de zones, on peut supposer que toutes les observations (et pas seulement celles qui sont les plus proches les unes des autres) interagissent les unes avec les autres. Dans ce cas, on adoptera une formulation plus générale pour les coefficients w_{ij} de la matrice d'interaction, du type $w_{ij} = \rho f(d_{ij})$ où $f(d_{ij})$ est une fonction décroissante d'une distance entre observations, éventuellement combinée comme plus haut avec d'autres variables. Les fonctions les plus utilisées sont $f(d_{ij}) = \exp(-d_{ij})$, $f(d_{ij}) = d_{ij}^{-1}$ et, dans le prolongement des modèles gravitaires, $f(d_{ij}) = d_{ij}^{-2}$.

En conclusion, il est rare qu'un choix unique s'impose d'emblée pour représenter les interactions spatiales. Si le recours aux matrices de contiguïté a l'avantage de fournir un outil similaire à celui des opérateurs retard en économétrie des séries temporelles, il est loin de s'imposer de manière aussi générale. Bien souvent, il faudra tester plusieurs solutions avant d'en arrêter une.

TESTER LA PRÉSENCE D'INTERACTIONS SPATIALES : LES TESTS ÉLÉMENTAIRES D'AUTOCORRÉLATION SPATIALE

Les tests de couleur

Commençons par l'exemple très simple d'un ensemble de N zones avec, pour chacune d'entre elles, une réalisation d'une variable binaire X (Pour une analyse détaillée de cet exemple, voir Cliff et Ord, 1973). Pour suivre la tradition imagée de la statistique spatiale, on désignera par zones blanches celles pour lesquelles la variable binaire est égale à zéro ($X_i = 0$), par zones noires celles pour lesquelles elle est égale à l'unité ($X_i = 1$). Les zones blanches et noires sont-elles réparties aléatoirement dans l'espace ? Si ce n'est pas le cas, c'est parce que :

• Deux zones contiguës tendent à être de la même couleur (les valeurs de la variable dichotomique sont souvent les mêmes). On parlera d'autocorrélation spatiale positive.

• Deux zones contiguës tendent à être de couleurs opposées (les valeurs de la variable dichotomique diffèrent en général). On parlera d'autocorrélation spatiale négative.

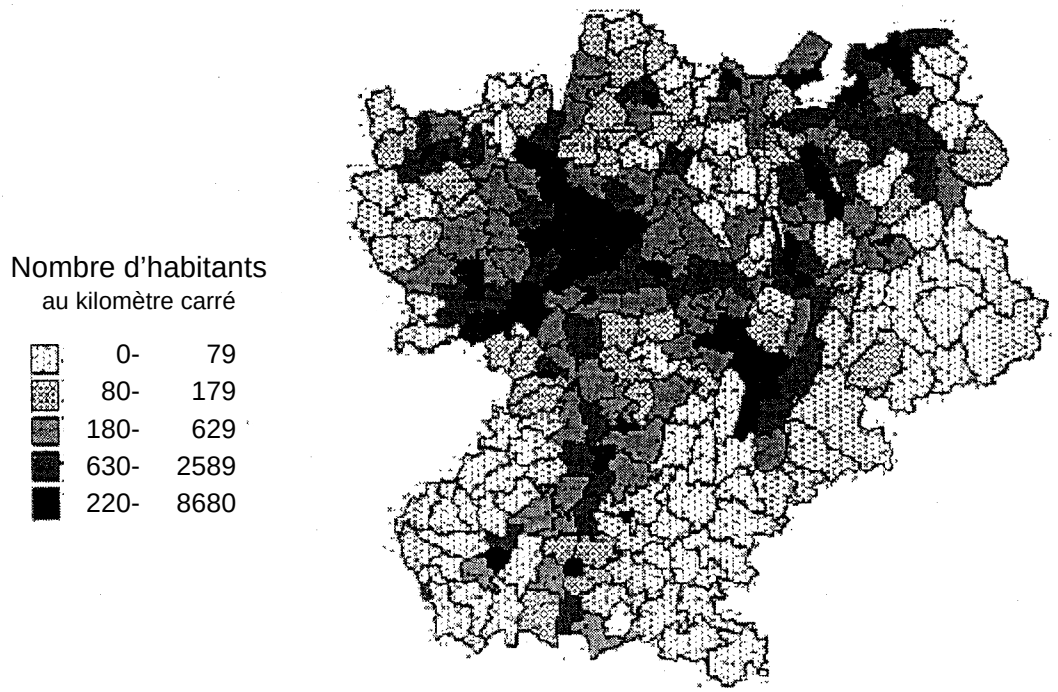
La carte de la figure 2 illustre bien cette notion. Les cantons à haute densité sont bien regroupés, en particulier autour de Lyon. Les cantons à basse densité sont également regroupés, principalement dans les zones de montagne. Deux cantons voisins ont donc tendance à avoir des densités voisines, ce qui correspond bien à la définition de l'autocorrélation spatiale positive.

Avant de parler de tests, il faut bien insister sur le fait que toute autocorrélation spatiale est relative à une définition particulière de la contiguïté, et nous venons de voir qu'aucune d'entre elles ne s'imposait sans discussion. Un exemple classique proposé initialement par Cliff et Ord (1973) illustre bien ce point. Prenons un échiquier. Sur ce dernier, on peut proposer trois définitions de la contiguïté, chacune associée au mouvement d'une pièce particulière. Pour la Tour, sont contiguës deux cases ayant un côté commun. Ces deux cases sont de couleurs systématiquement opposées : il y a autocorrélation spatiale négative. Pour le Fou, deux cases contiguës sont sur la même diagonale et ont un sommet commun. Elles ont toujours la même couleur. Pour le Roi et la Reine, sont contiguës deux cases ayant un côté ou un sommet commun. Chaque case intérieure est contiguë à quatre cases de même couleur et à quatre cases de couleur opposée : il n'y a pas d'autocorrélation spatiale.

Cet exemple montre bien la relativité de l'autocorrélation spatiale. Il ne doit cependant pas décourager. Il nous confirme ce qui était signalé plus haut : pour analyser l'interaction spatiale, il ne faut pas se restreindre à une seule définition de cette dernière. Il faut en étudier plu-

sieurs. La confrontation des résultats de plusieurs définitions est riche en enseignements sur la structure spatiale des données.

Figure 2. **Densité de population cantonale en Rhône-Alpes, 1990**



Comment tester l'existence d'autocorrélation spatiale, positive ou négative ? Il faut d'abord définir soigneusement l'absence d'autocorrélation spatiale, c'est-à-dire comment a été obtenue une répartition purement aléatoire des zones blanches et noires. On trouve habituellement deux définitions dans la littérature :

– Dans la première, appelée hypothèse N, on suppose que les valeurs blanches et noires ont été obtenues par des tirages aléatoires indépendants et de même loi dans chacune des zones, la probabilité d'une valeur noire étant égale à p.

– Dans la seconde, appelée hypothèse R, on suppose qu'il y avait au départ I réalisations de la valeur aléatoire, une par zone, dont une fraction p était noire. Ces valeurs ont été affectées à chacune des zones par tirage aléatoire sans remise.

Un test d'autocorrélation spatiale est un test de l'une des deux hypothèses nulles de répartition aléatoire, N ou R, l'hypothèse alternative étant la similarité (autocorrélation spatiale positive) ou la dissimilarité (autocorrélation spatiale négative) des zones contiguës. Le principe de ces tests, appelés tests de couleur, est très simple. Rappelons que l'ensemble

des zones peut être représenté par un graphe de contiguïté, chaque zone correspondant à un nœud relié par des arcs aux nœuds des zones contiguës. Attribuons à chaque nœud la couleur de la zone qu'il représente. En présence d'autocorrélation spatiale, chaque arc tendra à relier des nœuds de même couleur : deux nœuds noirs (on parlera d'arc noir-noir) ou deux nœuds blancs (on parlera d'arc blanc-blanc). Les arcs reliant deux nœuds de couleurs opposées (on parlera d'arc noir-blanc) seront rares. A l'opposé, en présence d'autocorrélation spatiale négative, chaque arc tendra à relier des nœuds de couleurs opposées, d'où une prédominance des arcs noir-blanc au détriment des arcs noir-noir et blanc-blanc.

L'espérance $E(NN)$ et la variance $V(NN)$ du nombre NN d'arcs noir-noir et l'espérance $E(NB)$ et la variance $V(NB)$ du nombre NB d'arcs noir-blanc sont connues. On en trouvera les valeurs en annexe 1. Or, on démontre qu'asymptotiquement [5], sous chacune des deux hypothèses nulles, la statistique NN (resp. la statistique NB) suit une loi normale de moyenne $E(NN)$ et de variance $V(NN)$ (resp. de moyenne $E(NB)$ et de variance $V(NB)$). La procédure de test est donc très simple.

– Pour un « test *NN* », on compte le nombre NN d'arcs noir-noir, puis on calcule la statistique centrée réduite $T_{NN} = \dfrac{NN - E(NN)}{\sqrt{V(NN)}}$ et on utilise un test de normalité de l'hypothèse $T_{NN} = 0$. Si T_{NN} est significativement positive (resp. négative), on conclut à l'existence d'autocorrélation spatiale positive (resp. négative).

– Pour un « test *NB* », on compte le nombre NB d'arcs noir-blanc, puis on calcule la statistique centrée réduite $T_{NB} = \dfrac{NB - E(NB)}{\sqrt{V(NB)}}$ et on utilise un test de normalité de l'hypothèse $T_{NB} = 0$. Si T_{NB} est significativement positive (resp. négative), on conclut à l'existence d'autocorrélation spatiale négative (resp. positive).

A titre d'exemple, on peut réaliser un test de couleur sur la carte de la figure 2 en la réduisant à deux groupes, les cantons à haute densité (au moins égale à 180 habitants par km^2) et les cantons à faible densité (au plus égale à 179 habitants par km^2). Sur les 309 cantons de la région Rhône-Alpes, 224 sont dans la première catégorie (qu'on assimilera à la couleur noire), 85 sont dans la seconde (qu'on assimilera à la couleur blanche). Le graphe de contiguïté entre ces 309 cantons comprend 853 arcs, dont 532 relient deux cantons à haute densité (arcs NN), 132 relient deux cantons à basse densité (arcs BB) et 189 un canton à haute densité et un canton à basse densité (arcs NB). L'utilisation des formules de l'annexe conduit aux résultats suivants :

[5] C'est-à-dire quand le nombre de nœuds et d'arcs du graphe devient suffisamment grand. On trouvera la justification de cette propriété dans Cliff et Ord, 1973.

Tableau 1.
Tests de couleur sur les densités de population des cantons de la région Rhône-Alpes

	Hypothèse *N*	Hypothèse *R*
NN	532	532
E(NN)	448,3	447,7
V(NN)	1 113,7	126,7
$T_{NN} = \frac{NN - E(NN)}{\sqrt{V(NN)}}$	2,5	7,5
NB	189	189
E(NB)	340,2	341,3
V(NB)	551,5	168,5
$T_{NB} = \frac{NB - E(NB)}{\sqrt{V(NB)}}$	-6,4	-11,7

Rappelons que, avec une loi normale, pour un test unilatéral au risque de 5 %, le seuil est de 1,6 en valeur absolue ; il est de 2,3 pour un risque de 1 %. Les deux tests, sous les deux hypothèses, conduisent donc bien à la conclusion qu'il y a autocorrélation spatiale et que cette dernière est positive.

Mathématiquement, les statistiques *NN* et *NB* s'écrivent de la manière suivante :

$$NN = \frac{1}{2}\sum_{i \neq j} c_{ij} x_i x_j \qquad \text{et} \qquad NB = \frac{1}{2}\sum_{i \neq j} c_{ij} (x_i - x_j)^2 \tag{3}$$

où x_i et x_j sont les valeurs[(6)] de la variable *X* en chacun des nœuds *i* et *j*, et c_{ij} est le coefficient de contiguïté. Ces expressions servent de base à deux généralisations. La première consiste à remplacer les coefficients de contiguïté par les coefficients w_{ij} d'une matrice *W* d'interactions spatiales quelconque. Les propriétés statistiques des tests sont les mêmes. Evidemment, d'une matrice *W* à l'autre, c'est le type d'interaction spatiale testée qui change, chaque matrice correspondant à un type d'interaction particulier.

Les tests de Moran et Geary

La seconde généralisation consiste à construire des tests utilisables sur des variables quantitatives quelconques. De la statistique *NN*, on déduit la statistique de Moran (Moran, 1950) :

$$M = \frac{N}{P} \frac{\sum_{i \neq j} c_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{4}$$

[(6)] Rappelons que $x_j = 0$ pour un nœud blanc, $x_i = 1$ pour un nœud noir.

où N est le nombre d'observations, $\bar{x}$ est la moyenne des x_i et $P = \sum_{i \neq j} c_{ij}$ la somme des coefficients d'interaction. De la statistique *NB*, on déduit la statistique de Geary (Geary, 1954):

$$G = \frac{N-1}{2P} \frac{\sum_{i \neq j} c_{ij} (x_i - x_j)^2}{\sum_i (x_i - \bar{x})^2} \qquad (5)$$

Les statistiques de Moran et de Geary ont également une interprétation intuitive. La première est égale au ratio de la covariance entre observations contiguës à la variance totale de l'échantillon. A un facteur 1/2 près, la seconde est égale au ratio de la variance des écarts entre observations contiguës à la variance totale.

Comme pour les statistiques *NN* et *NB*, on connaît l'espérance $E(M)$ et la variance $V(M)$ de la statistique de Moran ainsi que l'espérance $E(G)$ et la variance $V(G)$ de la statistique de Geary, dont on trouvera les valeurs en annexe 1. De même, on démontre qu'asymptotiquement, sous chacune des deux hypothèses nulles, la statistique de Moran (resp. la statistique de Geary) suit une loi normale de moyenne $E(M)$ et de variance $V(M)$ (resp. de moyenne $E(G)$ et de variance $V(G)$)[7]. La procédure de test est donc très simple.

• Pour un test de Moran, on calcule la statistique centrée réduite $T_M = \dfrac{M - E(M)}{\sqrt{V(M)}}$ et on utilise un test de normalité de l'hypothèse $T_M = 0$. Si T_M est significativement positive (resp. négative), on conclut à l'existence d'autocorrélation spatiale positive (resp. négative).

• Pour un test de Geary, on calcule la statistique centrée réduite $T_G = \dfrac{G - E(G)}{\sqrt{V(G)}}$ et on utilise un test de normalité de l'hypothèse $T_G = 0$. Si T_G est significativement positive (resp. négative), on conclut à l'existence d'autocorrélation spatiale négative (resp. positive).

Il ne faut pas oublier que ces tests n'ont de valeur qu'asymptotique. Il ne faut donc les pratiquer que si l'on dispose d'un nombre d'observations suffisamment élevé.

Ainsi, nous disposons avec les tests de Moran et Geary de tests qui permettront de trancher chaque fois qu'on soupçonne la présence d'autocorrélation dans une série quantitative. Dans la pratique, le test de Moran, dont l'expérience a montré qu'il était plus robuste et puissant que le test de Geary, est de loin le plus utilisé.

[7] Pour une démonstration de ces propriétés, on pourra, comme plus haut, consulter Cliff et Ord (1973 ou 1981).

On peut, là encore, illustrer l'usage des tests de Moran et de Geary à partir des données utilisées pour la carte de la figure 2. Cette fois, plutôt que de la réduire à une dimension dichotomique, nous allons utiliser telle quelle la série des densités cantonales pour calculer les statistiques de Moran et Geary. De plus, pour illustrer les effets d'une variation dans la définition de la proximité, nous allons utiliser trois matrices d'interaction.

La première est, comme pour les tests *NN* et *NB* faits plus haut, la matrice de contiguïté standard ; dans la deuxième, les coefficients de contiguïté ne sont plus égaux à l'unité, mais proportionnels à la longueur de la frontière commune entre deux cantons. Enfin, dans la troisième, les coefficients d'interaction sont inversement proportionnels à la distance entre les centroïdes des deux cantons, qu'ils soient contigus ou non. Toutes ces matrices sont normalisées. Le tableau suivant récapitule les résultats des calculs :

Tableau 2. **Tests de Moran et Geary sur les densités de population des cantons de la région Rhône-Alpes**

		Hypothèse *N*			Hypothèse *R*		
Test de Moran	*M*	*E(M)*	*V(M)*	T_M	*E(M)*	*V(M)*	T_M
Contiguïté d'ordre 1	0,475	-0,00325	0,00123	13,62	-0,00325	0,00113	14,23
Longueur de la frontière	0,494	-0,00325	0,00160	12,47	-0,00325	0,00150	13,03
Inverse de la distance	0,102	-0,00325	0,00022	22,19	-0,00325	0,00021	23,17
Test de Geary	*G*	*E(G)*	*V(G)*	T_G	*E(G)*	*V(G)*	T_G
Contiguïté d'ordre 1	0,588	1	0,00141	-10,99	1	0,00362	-6,84
Longueur de la frontière	0,493	1	0,0018	-12,02	1	0,0042	-7,83
Inverse de la distance	1,016	1	0,00063	2,07	1	0,0059	0,68

On notera que, dans tous les cas, le test de Moran conduit à un rejet beaucoup plus net de l'hypothèse nulle que le test de Geary. C'est l'illustration du fait, signalé plus haut, que la pratique a montré qu'en général le test de Moran est plus puissant que le test de Geary. Ce dernier accepte même l'hypothèse nulle ou ne la rejette que faiblement avec une matrice d'interactions proportionnelles à l'inverse de la distance entre centroïdes.

Comme cette matrice pondère beaucoup moins fortement les zones les plus proches et les petites zones que les matrices fondées sur la contiguïté, on peut penser que la similarité concerne surtout les cantons les plus proches. Cette observation simple montre bien le type de leçon qu'on peut tirer de l'usage simultané de plusieurs matrices d'interactions spatiales.

AUTOCORRÉLATION ET AUTORÉGRESSION SPATIALES : L'EXEMPLE DU MODÈLE LINÉAIRE

Formuler des modèles avec autocorrélation et autorégression spatiales

Quelles sont les conséquences économétriques de la présence d'interaction spatiale ? Pour en donner une première idée, nous allons nous intéresser au modèle économétrique le plus simple et le plus usuel, le modèle linéaire. Le raisonnement mené à cette occasion peut être réutilisé pour d'autres modèles plus complexes, par exemple les modèles de panels ou les modèles à variables dépendantes limitées.

Examinons d'abord le cas d'une variable unique dont il apparaît que ses valeurs sont corrélées entre elles dans l'espace (voir également Jayet, 1993 ou Anselin, 1988). Le modèle le plus simple pouvant expliquer cette autocorrélation est l'existence d'un processus autorégressif spatial, de la forme :

$$y = Ay + \varepsilon \Leftrightarrow (I - A)^{-1}y = \varepsilon \qquad (6)$$

où A est la matrice des effets d'autorégression spatiale, ε est un vecteur d'aléas indépendants, non nécessairement homoscédastiques :

$$E(\varepsilon) = 0 \text{ et } V(\varepsilon) = \sigma^2 V$$

et V est une matrice diagonale. L'introduction de V permet de tenir compte de l'hétérogénéité des observations, dont on a vu que, dans le domaine spatial, elle était la règle plutôt que l'exception.

Il faut noter un point important : un processus autorégressif spatial est toujours stationnaire. On le voit bien en partant du fait que, si $(I - A)y = \varepsilon$, sachant que $(I - A)$ doit être inversible pour que le processus ait un sens, alors $y = (I - A)^{-1}\varepsilon$. Dans ce cas,

$$E(\varepsilon) = 0 \text{ et } y = (I - A)^{-1}\varepsilon \Rightarrow E(y) = 0 \qquad (7)$$

Toutes les observations du processus sont d'espérance nulle. En conséquence, si une variable spatialisée est d'espérance non nulle, il faudra raisonner sur l'écart à son espérance. Qui plus est, un processus autorégressif spatial ne peut être soumis à aucune tendance. On retrouve ici une difficulté bien connue des spécialistes des données temporelles, qui les conduit à tester la stationnarité d'une série avant de faire des estimations et des tests économétriques. Et, une fois de plus, la double dimension des séries spatiales et l'absence d'un analogue à la flèche du temps rend difficile la formulation d'hypothèses simples de tendances, analogues aux *trends* linéaire ou exponentiel des séries temporelles. Il arrivera cependant que des données présentent un ordre naturel (direction du vent, éloigne-

ment d'un point ou d'une ligne remarquable,…) utilisable pour formuler une tendance.

La matrice des effets d'autocorrélation spatiale peut elle-même être paramétrée. Par exemple, on prend fréquemment une combinaison linéaire des matrices de contiguïté d'ordre 1 à K:

$$A = \rho_1 C^{(1)} + \cdots + \rho_K C^{(K)} \tag{8}$$

où les $\rho_1,\ldots,\rho_K$ sont des coefficients à estimer qui mesurent l'intensité de l'interaction. Par analogie avec l'économétrie des séries temporelles, on parle alors d'un processus spatial autorégressif d'ordre K, ou SAR(K). Les processus SAR(1) sont de loin les plus utilisés.

En présence de variables explicatives, le point de départ est le modèle linéaire usuel,

$$y = X\beta + \varepsilon \tag{9}$$

où X est la matrice des variables explicatives, β le vecteur des paramètres à estimer et ε la partie aléatoire du modèle. L'introduction de l'interaction spatiale peut prendre deux formes.

Dans la première, on considère que l'interaction spatiale porte sur la variable expliquée. On aboutit alors au modèle spatial autorégressif:

$$y = Ay + X\beta + \varepsilon \Leftrightarrow (I - A)\, y = X\beta + \varepsilon \tag{10}$$

où, comme plus haut, on fait sur ε les hypothèses standard des moindres carrés,

$$E(\varepsilon) = 0 \text{ et } V(\varepsilon) = \sigma^2 V \tag{11}$$

Le modèle spatial autorégressif s'impose en particulier dès qu'on n'a aucune raison de penser que la variable expliquée est d'espérance nulle partout dans l'espace. Il faut alors utiliser des variables explicatives permettant de rendre compte de la valeur que prend cette espérance, sous peine de formuler un modèle incohérent. C'est ainsi que la manière la plus simple d'analyser une variable d'espérance constante, mais non nulle, est d'introduire une constante parmi les variables explicatives du modèle, ce qui permet d'estimer la moyenne. Si l'espérance n'est pas constante, il faudra introduire d'autres variables explicatives.

Dans le deuxième cas, l'interaction spatiale porte sur la partie aléatoire du modèle, ε, qui suit un processus autorégressif spatial. On aboutit au modèle avec autocorrélation spatiale des résidus:

$$\left.\begin{array}{r} y = X\beta + \varepsilon \\ (I - G)\varepsilon = \eta \end{array}\right| \Leftrightarrow (I - G)\,(y - X\beta) = \eta \tag{12}$$

où G est la matrice des effets d'autocorrélation spatiale et η est un vecteur d'aléas indépendants, non nécessairement homoscédastiques:

$$E(\eta) = 0 \text{ et } V(\eta) = \sigma^2 V \tag{13}$$

Ce qui a été dit plus haut au sujet des processus autorégressifs spatiaux est évidemment vrai du processus gouvernant le vecteur des résidus, ε : toutes les composantes doivent être d'espérance nulle et, en particulier, aucune tendance ne doit être présente. Si ce n'est pas le cas, il faut reformuler le modèle pour introduire les tendances et, éventuellement, d'autres variables explicatives adaptées.

On peut enfin combiner les deux possibilités : il y a à la fois interaction spatiale sur la variable expliquée et sur la partie aléatoire du modèle. C'est le modèle le plus général, avec autocorrélation et autorégression :

$$\left.\begin{array}{r} y = Ay + X\beta + \varepsilon \\ (I - G)\varepsilon = \eta \end{array}\right| \Leftrightarrow (I - G)\left[(\mathrm{I} - \mathrm{A})\, y - X\beta\right] = \eta \qquad (14)$$

où A est la matrice des effets d'autorégression spatiale, G la matrice des effets d'autocorrélation spatiale et η est un vecteur d'aléas indépendants, non nécessairement homoscédastiques :

$$E(\eta) = 0 \text{ et } V(\eta) = \sigma^2 V \qquad (15)$$

Estimer des modèles avec autocorrélation spatiale

Quand ni V, qui permet de tenir compte de l'hétérogénéité, ni A ou G, par lesquelles transitent les effets d'autocorrélation spatiale, ne dépendent de paramètres à estimer, les modèles ci-dessus ont des matrices de variances-covariances connues et s'estiment sans difficulté par les moindres carrés généralisés :

– Pour le modèle avec autorégression spatiale, on régresse $(I - A)y$ sur la matrice X des variables explicatives, le résidu ε ayant comme matrice de variances-covariances $V(\varepsilon) = \sigma^2 V$, d'où l'estimateur des moindres carrés généralisés :

$$\hat{\beta} = (X' V^{-1} X)^{-1}\ X' V^{-1} (I - A)\, y \qquad (16)$$

– Pour le modèle avec autocorrélation spatiale des résidus, on régresse y sur la matrice X des variables explicatives, le résidu ε ayant comme matrice de variances-covariances $V(\varepsilon) = \sigma^2\ [(I - G)V^{-1}\ (I - G')]^{-1}$, d'où l'estimateur des moindres carrés généralisés :

$$\hat{\beta} = \left|\left(X' (I - G')V^{-1} (I - G)\, X\right)^{-1}\right| X' (I - G')V^{-1} (I - G)\, y \qquad (17)$$

– Pour le modèle avec autorégression et autocorrélation spatiales, on régresse $(I - A)y$ sur la matrice X des variables explicatives, le résidu ε ayant comme matrice de variances-covariances $V(\varepsilon) = \sigma^2\ [(I - G)V^{-1}(I - G')]^{-1}$, d'où l'estimateur des moindres carrés généralisés :

$$\hat{\beta} = (X' (I - G')V^{-1} (I - G)\, X)^{-1}\ X' (I - G)' V^{-1} (I - G)\, (I - A)\, y \qquad (18)$$

Cependant, en général, les matrices V, A et G dépendent de paramètres à estimer. C'est même quasiment la règle pour les matrices d'interaction spatiale, A et G, car l'économètre n'a pas de raison de déterminer à l'avance l'ampleur des effets d'autocorrélation spatiale. Dans ce cas, les moindres carrés généralisés ne sont plus utilisables. L'économètre peut recourir aux méthodes de variables instrumentales, du type double moindres carrés, ou aux méthodes de maximum de vraisemblance. Ces dernières sont actuellement les plus utilisées et ce sont elles que nous présenterons (pour les méthodes de variables instrumentales, voir Anselin, 1988). Nous nous restreindrons au cas du modèle avec autocorrélation spatiale des résidus, le traitement des autres modèles étant similaire.

L'économètre a donc spécifié une matrice d'effets d'autocorrélation spatiale des résidus, $G(\rho)$, le plus souvent sous la forme $G(\rho) = \rho W$ où W est une matrice normalisée de poids spatiaux, par exemple une matrice de contiguïté (on a alors un SAR (1)). Eventuellement, il a également paramétré la matrice de poids spatiaux, $V = V(\mu)$. On a vu plus haut que le modèle avec autocorrélation des résidus se présentait sous la forme :

$$\eta = \left((I - G(\rho)\right)(y - X\beta) \tag{19}$$

avec $E(\eta) = 0$ et $V(\eta) = \sigma^2 V(\mu)$. Sous l'hypothèse de normalité des résidus, on en déduit l'expression de la log-vraisemblance,

$$L(\beta,\rho,\mu,\sigma^2) = -\frac{N}{2}\ln 2\pi\sigma^2 - \frac{1}{2}\ln\det\left[I - G(\rho)\right] - \frac{1}{2}\ln\det V(\mu) - \frac{1}{2\sigma^2}SRG(\rho,\mu,\beta) \tag{20}$$

où

$$SRG(\rho,\mu,\beta) = (y - X\beta')\left|\left|(I - G(\rho'))V^{-1}(\mu)\right|\right|(I - G(\rho))\left|\left|(y - X\beta)\right|\right|$$

s'interprète comme une somme de carrés de résidus généralisés. On voit facilement que pour ρ et μ donnés, l'estimateur de β est le même que celui des moindres carrés généralisés,

$$\hat{\beta}(\rho,\mu) = \left[\left|X'\left|\left|(I - G'(\rho))\right|\right|V^{-1}(\mu)\left|\left|(I - G(\rho))\right|\right|X\right|\right]^{-1}\left|X'\left|\left|(I - G'(\rho))\right|\right|V^{-1}(\mu)\left|\left|(I - G(\rho))\right|\right|y\right| \tag{21}$$

et qu'il en est de même de l'estimateur de la variance,

$$\hat{\sigma}^2(\rho,\mu) = N^{-1}SRG\left(\left|\rho,\mu,\hat{\beta}(\rho,\mu)\right|\right) \tag{22}$$

d'où, sachant que la matrice des effets d'hétéroscédasticité est diagonale, $V(\mu) = diag(v_1(\mu), \ldots, v_1(\mu))$, l'expression de la log-vraisemblance concentrée :

$$CL(\rho,\mu,\sigma^2) = -\frac{N}{2}\left[1 + \ln 2\pi + \ln SRG(\rho,\mu,\hat{\beta}(\rho,\mu))\right] + \frac{1}{2}\ln\det\left[I - G(\rho)\right] + \frac{1}{2}\sum_i \ln v_i(\mu) \tag{23}$$

C'est cette expression qu'il faut maximiser par rapport aux paramètres ρ et μ pour trouver leurs estimateurs. En général, la maximisation par rapport à μ ne pose pas de problème. Pour ce qui est de ρ, les

choses sont plus difficiles. En effet, dans le cas général, il faut calculer le déterminant d'une matrice qui peut être de taille élevée et ce à chaque itération de l'algorithme de maximisation. On notera cependant (Ord, 1975) que, quand $G(\rho) = \rho W$, on a :

$$\ln\det\,[I - G(\rho)] = \ln\det\,[I - \rho W] = \sum_i \ln\,(1 - \rho\lambda_i) \qquad (24)$$

où les λ_i sont les valeurs propres de la matrice W. Il suffit donc de les calculer une fois pour toutes au départ.

Tests sur les modèles

La formulation d'un modèle avec autorégression et/ou avec autocorrélation spatiale conduit à des tests qui permettent de déterminer si l'introduction de l'une ou l'autre de ces formes d'interaction entre observations est pertinente, au moins au niveau des données. C'est ainsi que, dans le cadre du modèle avec autorégression spatiale,

$$y = Ay + \varepsilon \qquad (25)$$

si, comme c'est habituellement le cas, $A = \rho C$, tester l'absence d'autorégression spatiale revient à tester l'hypothèse nulle $\rho = 0$. De même, dans le cadre du modèle avec autocorrélation spatiale,

$$y = X\beta \text{ et } \varepsilon = G\varepsilon + \eta \qquad (26)$$

quand $G = \gamma C$, tester l'absence d'autocorrélation spatiale revient à tester l'hypothèse nulle $\gamma = 0$. Enfin, si l'on part du modèle général avec autorégression et autocorrélation spatiales,

$$y = Ay + X\beta \text{ et } \varepsilon = G\varepsilon + \eta \qquad (27)$$

et que $A = \rho C$ et $G = \gamma C$, on peut tester plusieurs hypothèses nulles : absence d'autorégression spatiale ($\rho = 0$), absence d'autocorrélation spatiale ($\gamma = 0$), absence simultanée d'autorégression et d'autocorrélation spatiales ($\rho = 0$ et $\gamma = 0$).

L'estimation par le maximum de vraisemblance fournit ici un cadre commode, puisqu'il suffit d'appliquer la méthodologie usuelle des tests fondés sur la vraisemblance[(8)]. On peut alors :

– Soit estimer le modèle sous l'hypothèse nulle, c'est-à-dire sans autorégression ou autocorrélation spatiale, et pratiquer un test du multiplicateur de Lagrange. Cette stratégie est bien adaptée au cas où l'on souhaite éviter les difficultés de calcul d'un estimateur du maximum de vraisemblance quand celle-ci n'est pas nécessaire et/ou quand la dépendance entre observations n'est pas une préoccupation centrale. On sou-

[(8)] Pour une présentation générale et accessible des tests fondés sur la vraisemblance, voir Greene (1997), pp. 155-168.

haite simplement vérifier son absence, car sa présence pourrait perturber les résultats. L'inconvénient est évidemment que, si les tests rejettent l'hypothèse nulle, il faudra recommencer les estimations. De plus, les statistiques sont d'un calcul malaisé.

– Soit estimer le modèle sous l'hypothèse alternative, c'est-à-dire avec autorégression et/ou autocorrélation spatiale, et pratiquer un test de Wald. Cette stratégie est bien adaptée au cas où l'on pense que la présence d'interdépendances spatiales est très vraisemblable et/ou quand ces interactions sont un aspect central du modèle. L'inconvénient est que l'on prend le risque de se lancer d'emblée dans un processus lourd d'estimation alors qu'un processus beaucoup plus simple suffit quand l'hypothèse nulle est vraie. Et, comme dans le premier cas, les statistiques sont d'un calcul malaisé.

– Soit estimer les deux modèles et faire un test du rapport de vraisemblance. L'estimation des deux modèles peut être un processus lourd, alors qu'un seul des deux sera retenu. En contrepartie, le calcul de la statistique du rapport de vraisemblance est immédiat.

Parce qu'ils cherchent en général à éviter l'estimation d'un modèle complet quand ce n'est pas nécessaire, la plupart des économètres spatiaux tendent à privilégier la première méthode. Ils ont, dans ce cadre, développé des tests complémentaires au test usuel du rapport de vraisemblance, dont on trouvera la présentation en annexe (voir également Anselin *et al.*, 1996).

Avant de conclure, il nous faut souligner deux points. Le premier est que, pour simplifier l'exposé, on a volontairement laissé de côté dans cette présentation le problème posé par l'interaction entre les observations à l'intérieur de la zone géographique sur laquelle on fait l'estimation et les observations à l'extérieur de cette zone. Ce faisant, on suppose implicitement que la zone est fermée. Pour reprendre un exemple antérieur, estimer une série sur les communes ou les cantons de la région Rhône-Alpes suppose implicitement que les localités de cette région n'interagissent pas avec celles des régions voisines, ce qui n'est valide en toute rigueur que si la région est fermée. Cette hypothèse implicite de fermeture est sans doute acceptable quand on travaille sur le découpage exhaustif d'un territoire national. Dans des espaces plus petits, elle est plus difficilement acceptable.

Vouloir traiter cette question, qui est l'analogue du problème des premières observations d'un processus autorégressif temporel, pose cependant des problèmes beaucoup plus délicats que nous n'aborderons pas ici. Notons cependant que, si les spécialistes des données temporelles peuvent raisonner conditionnellement aux premières observations (considérées comme purement exogènes), ce n'est pas possible ici puisque toutes les observations s'influencent réciproquement. Pour plus de précisions sur ces questions, on pourra consulter Griffith (1988).

Notre deuxième remarque est que les modèles présentés relèvent de méthodes statistiques paramétriques globales, le même modèle paramétrique étant valable pour l'ensemble des observations. Il est possible d'utiliser des méthodes non-paramétriques ou semi-paramétriques, ainsi que des modèles flexibles capables de s'adapter à des situations locales. Ces modèles sont adaptés à de grands espaces où l'on a de bonnes raisons de penser que les processus connaissent des variations locales importantes dont ni les variables explicatives utilisées, ni les matrices d'interactions spatiales ne peuvent rendre compte de manière suffisante. Mais, comme toutes les formes flexibles, ils ont un prix : difficulté d'incorporation des variables explicatives, coût des calculs, faible puissance des tests… Le lecteur intéressé pourra utilement consulter l'ouvrage de Upton et Fingleton (1985).

C0NCLUSI0N

Que conseiller au praticien confronté à des données spatiales ?

En premier lieu, de tenir compte de leur hétérogénéité. Ce qui se traduira de deux manières. D'une part, en introduisant des variables permettant de tenir compte des différences de taille, de structure, voire de forme entre les observations spatiales utilisées. D'autre part, en testant et en corrigeant une possible hétéroscédasticité.

En deuxième lieu, de tester systématiquement l'existence d'interactions spatiales sous leurs différentes formes, autorégression des variables expliquées et autocorrélation sur les parties aléatoires des modèles utilisés. Le plus souvent, pour réaliser ces tests, on ne se cantonnera pas à une forme unique d'interaction et donc à une seule famille de matrices d'interaction. En effet, chaque famille de matrices d'interactions implique des restrictions particulières sur le type d'interaction. Aucune n'est suffisamment générale. De plus, la confrontation des résultats obtenus avec des matrices différentes peut s'avérer très révélatrice de la structure spatiale des données.

En troisième lieu, de formuler explicitement les modèles économétriques de base et la manière dont on y introduit l'interaction spatiale. C'est ce que nous avons fait dans cet article, sur le modèle linéaire en partant de sa version standard. C'est ce qu'il faudra faire dans le cas d'autres modèles, comme par exemple les modèles à variables qualitatives ou plus généralement à variables dépendantes limitées. On introduit alors l'autorégression et l'autocorrélation spatiales dans la partie latente du modèle pour en tirer ensuite les conséquences sur le modèle observable.

BIBLIOGRAPHIE

ANSELIN (L.), 1988 — *Spatial Econometrics: Methods and Models*, Dordrecht, Kluwer.

ANSELIN (L.), BERA (A. K.), FLORAX (R.), YOON (M. J.), 1996 — Simple diagnostic tests for spatial dependance, *Regional Science and Urban Economics*, 26, pp. 77-104.

ARBIA (G.), 1989 — *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*, Dordrecht, Kluwer.

CLIFF (A.), ORD (J. K.), 1973 — *Spatial Autocorrelation*, Londres, Pion.

CLIFF (A.), ORD (J. K.), 1981 — *Spatial Processes. Models and Applications*, Londres, Pion.

GEARY (R. C.), 1954 — The contiguity ratio and statistical mapping, *The Incorporated Statistician*, 5, pp. 115-145.

GREENE (W. H.), 1997 — *Econometric Analysis*, 3rd edition, London, Prentice-Hall.

GRIFFITH (D.), 1988 — *Advanced Spatial Statistics: Special Topics in the Exploration of Quantitative Spatial Data Series*, Dordrecht, Kluwer.

JAYET (H.), 1993 — *Analyse spatiale quantitative: une introduction*, Paris, Economica.

MORAN (P. A. P.), 1950 — A test for serial dependence of residuals, *Biometrika*, 37, pp. 178-181.

ORD (J. K.), 1975 — Estimation methods for models of spatial interaction, *Journal of the American Statistical Association*, 70, pp. 120-126.

UPTON (G. J. G.), FINGLETON (B.), 1985 — *Spatial Data Analysis by Example*, New York, Wiley.

ANNEXE 1

Moments des statistiques *NN*, *NB* et des statistiques de Moran et de Geary

On adopte les notations suivantes :

N Nombre d'observations = nombre de nœuds du graphe
N_1 Nombre d'observations avec $x_i = 1$ (nœuds noirs)
N_2 Nombre d'observations avec $x_i = 0$ (nœuds blancs)

$$N_1 + N_2 = N, p = N_1/N, q = N_2/N, p + q = 1$$

$$W = \sum_{i \neq j} c_{ij}$$ $W/2$ est le poids total des arcs.

$$\bar{x} = \frac{1}{N} \sum x_i$$

$$b_2 = N \frac{\sum (x_i - \bar{x})^4}{\left(\sum (x_i - \bar{x})^2\right)^2}$$

$$c_{i.} = \sum_j c_{ij}, \quad c_{.i} = \sum_j c_{ji}$$

$$Z_1 = \frac{\left[\sum_{i \neq j} (c_{ij} + c_{ji})^2\right]}{2W}, \quad Z_2 = \frac{\left[\sum_i (c_{i.} + c_{.j})^2\right]}{W}$$

Espérances et variances de *NN* et *NB* sous l'hypothèse *N*

$$E(NN) = \frac{1}{2} Wp^2 = \mu_{NN}$$

$$V(NN) = \frac{1}{2} \mu_{NN} \left[Z_1 + (Z_2 - 2Z_1)p + (Z_1 - Z_2)p^2\right]$$

$$E(NB) = Wpq = \mu_{NB}$$

$$V(NB) = \frac{1}{2} \mu_{NB} \left[Z_1 + (Z_2 - 2Z_1) \frac{p + q}{2} + 2(Z_1 - Z_2)\, pq\right]$$

Espérances et variances de *NN* et *NB* sous l'hypothèse *R*

$$E(NN) = \frac{W}{2} \frac{N_1(N_1 - 1)}{N(N - 1)} = \bar{\mu}_{NN}$$

$$V(NN) = \frac{1}{2}\,\bar{\mu}_{NN}\left[Z_1 + \frac{N_1 - 2}{N-2}\left[Z_2 - 2Z_1 + \frac{N_1 - 3}{N-3}\,(W + Z_1 - Z_2)\right] - 2\bar{\mu}_{NN}\right]$$

$$E(NB) = W\,\frac{N_1\,N_2}{N(N-1)} = \bar{\mu}_{NB}$$

$$V(NB) = \frac{1}{2}\,\bar{\mu}_{NB}\left[Z_1 + (Z_2 - 2Z_1)\,\frac{N_1 + N_2 - 2}{2(N-2)} + 2(W + Z_1 - Z_2)\,\frac{(N_1 - 1)(N_2 - 1)}{(N-2)(N-3)} - 2\bar{\mu}_{NB}\right]$$

Espérances et variances des statistiques de Moran et Geary sous l'hypothèse *N*

$$E(M) = -\frac{1}{N-1}$$

$$V(M) = \frac{N(NZ_1 - Z_2) + 3W}{(N^2 - 1)W} - \frac{1}{(N-1)^2}$$

$$E(G) = 1$$

$$V(G) = \frac{(N-1)\,(2Z_1 + Z_2) - 4W}{2(N+1)\,W}$$

Espérances et variances des statistiques de Moran et Geary sous l'hypothèse *R*

$$E(M) = -\frac{1}{N-1}$$

$$V(M) = \frac{N\left[\left(N^2 - 3N + 3\right)Z_1 - nZ_2 + 3W\right] - b_2\left[\left(N^2 - N\right)Z_1 - 2NZ_2 + 6W\right]}{(N-1)(N-2)\,(N-3)W}$$

$$E(G) = 1$$

$$V(G) = \frac{Y_1 - Y_2 + Y_3}{N(N-2)(N-3)\,W}$$

$$Y_1 = (N-1)\left[(N^2 - 3N + 3) - (N-1)\,b_2\right]Z_1$$

$$Y_2 = \left[(N^2 + 3N - 6) - (N^2 - N + 2)\,b_2\right]\frac{(N-1)Z_2}{4}$$

$$Y_3 = \left[N^2 - 3 - (N-1)^2\,b_2\right]W$$

ANNEXE 2

Tests sur le modèle linéaire

Dans le cas homoscédastique ($V = I$), on dispose des tests suivants :

- Tests d'autorégression spatiale

Multiplicateur de Lagrange classique : $\left(\frac{\hat{\varepsilon}' W y}{\hat{\sigma}^2}\right)^2 \Big/ \Theta \rightarrow \chi^2(1)$

Bera et Yoon : $\left(\frac{\hat{\varepsilon}' W y}{\hat{\sigma}^2} - \frac{\hat{\varepsilon}' W \hat{\varepsilon}}{\hat{\sigma}^2}\right)^2 \Big/ [\Theta - T] \rightarrow \chi^2(1)$

- Tests d'autocorrélation spatiale

Moran : $\frac{\hat{\varepsilon}' W \hat{\varepsilon}}{\hat{\varepsilon}' \hat{\varepsilon}} \rightarrow N(0,1)$

Multiplicateur de Lagrange classique : $\left(\frac{\hat{\varepsilon}' W \hat{\varepsilon}}{\hat{\sigma}^2}\right)^2 \Big/ T \rightarrow \chi^2(1)$

Bera et Yoon : $\left(\frac{T}{\Theta} \frac{\hat{\varepsilon}' W y}{\hat{\sigma}^2} - \frac{\hat{\varepsilon}' W \hat{\varepsilon}}{\hat{\sigma}^2}\right)^2 \Big/ [T - T^2/\Theta] \rightarrow \chi^2(1)$

Kelejian-Robinson : $\frac{\hat{\gamma}' Z' Z \hat{\gamma}}{\hat{\alpha}' \hat{\alpha}/b} \rightarrow \chi^2(K)$

- Test simultané d'autorégression et autocorrélation spatiales

Multiplicateur de Lagrange classique :

$$\left(\frac{\hat{\varepsilon}' W y}{\hat{\sigma}^2} - \frac{\hat{\varepsilon}' W \hat{\varepsilon}}{\hat{\sigma}^2}\right)^2 \Big/ [\Theta - T] + \left(\frac{\hat{\varepsilon}' W \hat{\varepsilon}}{\hat{\sigma}^2}\right)^2 \Big/ T \rightarrow \chi^2(2)$$

Pour tous ces tests, les notations suivantes sont utilisées :

W est la matrice des coefficients d'interaction spatiale, avec $A(\rho) = \rho W$ et $G(\gamma) = \gamma W$. Elle est normalisée.

y est le vecteur des observations de la variable expliquée.

$\hat{\varepsilon}$ est le vecteur des résidus de la régression sans interaction spatiale.

$\hat{\sigma}^2$ est l'estimateur de la variance.

$T = \text{trace}(W'W + W^2)$ et $\Theta = T + (WX\hat{\beta}')M\,(WX\hat{\beta})/\hat{\sigma}^2$ où $\hat{\beta}$ est le vecteur des coefficients estimés et $M = I - X(X'X)^{-1}X'$.

$\hat{\gamma}$ et $\hat{\alpha}$ sont les estimateurs du vecteur des coefficients et du vecteur des résidus de la régression auxiliaire $e = Z\gamma + \alpha$ dont chaque observation correspond à un couple de zones contiguës (le coefficient correspondant de W est non nul). La valeur de e pour cette observation est le produit des valeurs correspondantes du vecteur des résidus estimés, $\hat{\varepsilon}$. Les K colonnes de Z sont formées de manière analogue à partir des valeurs des variables explicatives.

Le test de Moran est une adaptation au cas des résidus d'une régression du test de Moran présenté plus haut. Les tests du multiplicateur de Lagrange classiques sont ceux qu'on obtient à partir de la vraisemblance du modèle linéaire sous l'hypothèse de normalité des résidus. Les tests de Bera et Yoon, connus aussi sous le nom de tests du multiplicateur de Lagrange robustes, sont robustes à une mauvaise spécification locale du terme autorégressif ou de la forme de l'autocorrélation des résidus. Enfin, le test de Kelejian Robinson est un test robuste qui reste valide avec des résidus non normaux et pour des modèles non linéaires. On notera cependant que ce test est très peu puissant, ce qui le rend peu attractif malgré sa robustesse.