

Cet article est disponible en ligne à l'adresse :

[http://www.cairn.info/article.php?ID\\_REVUE=ECOP&ID\\_NUMPUBLIE=ECOP\\_162&ID\\_ARTICLE=ECOP\\_162\\_0151](http://www.cairn.info/article.php?ID_REVUE=ECOP&ID_NUMPUBLIE=ECOP_162&ID_ARTICLE=ECOP_162_0151)

---

## Hétérogénéité spatiale. Principes et méthodes

par Julie LE GALLO

| La Documentation française | Économie et Prévision

2004/1 - n° 162

ISSN 0249 - 4744 | pages 151 à 172

---

Pour citer cet article :

– Le Gallo J., Hétérogénéité spatiale. Principes et méthodes, Économie et Prévision 2004/1, n° 162, p. 151-172.

---

Distribution électronique Cairn pour La Documentation française.

© La Documentation française. Tous droits réservés pour tous pays.

La reproduction ou représentation de cet article, notamment par photocopie, n'est autorisée que dans les limites des conditions générales d'utilisation du site ou, le cas échéant, des conditions générales de la licence souscrite par votre établissement. Toute autre reproduction ou représentation, en tout ou partie, sous quelque forme et de quelque manière que ce soit, est interdite sauf accord préalable et écrit de l'éditeur, en dehors des cas prévus par la législation en vigueur en France. Il est précisé que son stockage dans une base de données est également interdit.

# Hétérogénéité spatiale

## *Principes et méthodes*

Julie Le Gallo<sup>(\*)</sup>

*L'objectif de cet article est double. D'une part, nous présentons les principales spécifications économétriques permettant de capter le phénomène de l'hétérogénéité spatiale, qui se traduit sous la forme d'une instabilité des paramètres dans l'espace et/ou d'une hétéroscédasticité des termes d'erreurs. Seules les spécifications valables en coupe transversale sont examinées. D'autre part, nous explicitons les liens entre hétérogénéité et autocorrélation spatiales, l'autre grande spécificité des données localisées qui est définie par l'absence d'indépendance entre observations géographiques. En particulier, nous examinons dans quelle mesure les tests traditionnels d'hétéroscédasticité ou d'instabilité doivent être amendés pour tenir compte de l'autocorrélation spatiale. Une application, destinée à illustrer les principales modélisations et les différents tests, est également proposée.*

(\*) LEG UMR-CNRS 5118, Université de Bourgogne.  
e-mail : julie.legallo@u-bourgogne.fr ou legallo@u-bordeaux.4.fr

L'auteur remercie C. Baumont, C. Ertur, R. Guillain, M.-C. Pichery ainsi que deux rapporteurs anonymes pour leurs commentaires et suggestions. L'auteur reste seule responsable des insuffisances que pourrait comporter ce texte.

L'étude des espaces urbains fait souvent apparaître une segmentation des marchés immobiliers : les caractéristiques et les prix des logements diffèrent substantiellement selon leurs localisations. Cette segmentation conduit à des variations persistantes et significatives des caractéristiques des logements et de leurs prix dans les différents sous-marchés. Dans ces conditions, estimer une relation " globale " entre le prix du logement et ses caractéristiques, relation s'appliquant de la même façon sur toute l'aire urbaine étudiée, ne permet pas de capter les différences importantes de prix dans l'espace.

L'instabilité dans l'espace des relations économiques illustrée par cet exemple est appelée *hétérogénéité spatiale*. Ce phénomène se retrouve à plusieurs échelles : les comportements et les phénomènes économiques ne sont pas les mêmes dans le centre d'une ville et dans sa périphérie, dans une région urbaine et dans une région rurale, dans le " Nord " de l'Europe et dans le " Sud ", etc.

En pratique, ces différences peuvent se traduire dans une régression économétrique de deux façons : par des coefficients différents ou par des variances des termes d'erreurs différentes selon la localisation. Dans le premier cas, on parle d'*instabilité spatiale des paramètres* de la régression qui varient systématiquement avec la localisation. Dans le second cas, on est confronté à un problème d'*hétéroscédasticité*, problème par ailleurs fréquent en coupe transversale. Ces deux cas, instabilité structurelle et hétéroscédasticité, peuvent apparaître conjointement.

L'objectif de cet article est double. D'une part, nous présentons les principales spécifications économétriques permettant de capter ces phénomènes d'hétérogénéité spatiale. D'autre part, nous examinons dans quelle mesure les tests traditionnels d'hétéroscédasticité ou d'instabilité doivent être amendés pour tenir compte de l'autocorrélation spatiale, l'autre grande spécificité des données localisées et qui est définie par l'absence d'indépendance entre observations géographiques (Anselin, 1988a ; Le Gallo, 2002a, 2002b). Nous nous limitons ici aux spécifications valables pour les données en coupe transversale.

Au contraire de l'autocorrélation spatiale, l'hétérogénéité spatiale est habituellement traitée par les méthodes économétriques standards. Ainsi, les modélisations en coupe transversale relatives aux paramètres variables, aux coefficients aléatoires ou aux changements de régimes, peuvent être adaptées aux spécifications particulières nécessitées par l'utilisation de données localisées.

Cet article s'organise ainsi de la façon suivante. Tout d'abord, dans la première partie, sont présentées les différentes spécifications permettant de prendre en compte l'instabilité des paramètres dans l'espace ainsi que l'hétéroscédasticité. Ensuite, dans une deuxième partie, nous explicitons les liens entre autocorrélation et hétérogénéité spatiales. En effet, ces deux aspects sont souvent présents conjointement dans une régression, conduisant à une nécessaire adaptation des tests d'hétéroscédasticité et d'instabilité structurelle. Enfin, dans une dernière partie, nous présentons une application destinée à illustrer les principales modélisations et les différents tests exposés dans cet article.

---

## L'instabilité spatiale des paramètres

---

Le point de départ est le modèle de régression linéaire classique en coupe transversale :

$$(1.1) \quad y = X\beta + \varepsilon$$

où  $y$  est le vecteur  $(N, 1)$  des observations de la variable dépendante,  $N$  est le nombre total d'observations,  $X$  est la matrice  $(N, K)$  des observations des  $K$  variables explicatives,  $\beta$  est le vecteur  $(K, 1)$  des  $K$  coefficients inconnus à estimer. Sauf indication contraire,  $\varepsilon$  désigne le vecteur  $(N, 1)$  des termes d'erreurs tel que  $\varepsilon \sim \text{iid}(0, \sigma^2 I_N)$ . Par la suite, nous considérons que les observations sont localisées, c'est-à-dire qu'elles portent sur des zones géographiques, comme des régions <sup>(1)</sup>.

L'*instabilité spatiale des paramètres* se réfère à l'absence de stabilité dans l'espace des comportements ou des relations économiques. Ce phénomène nécessite des modélisations permettant la prise en compte des caractéristiques particulières de chaque zone géographique : les paramètres de la régression (1.1) varient selon leurs localisations. Dans le cas le plus général, on suppose une relation différente pour chaque zone géographique  $i$  de l'échantillon :

$$(1.2) \quad y_i = x_i' \beta_i + \varepsilon_i \quad i = 1, \dots, N$$

où  $y_i$  représente l'observation de la variable dépendante pour la zone  $i$ ,  $x_i'$  représente le vecteur de dimension  $(1, K)$  comportant les observations des  $K$  variables explicatives pour la zone  $i$ . Ce vecteur est associé à un vecteur de dimension  $(K, 1)$  :  $\beta_i$ . Il existe alors un vecteur de paramètres différent pour chaque observation  $i$ .

Compte tenu de cette spécification, il n'est pas possible d'estimer un ensemble de  $N$  vecteurs de  $K$  paramètres inconnus  $\beta_i$  avec un échantillon de  $N$  observations. Il faut donc imposer une structure spatiale aux données, c'est-à-dire fournir une spécification parcimonieuse afin d'étudier la variation dans l'espace. La variabilité spatiale de la moyenne d'une variable ou des coefficients d'une régression peut être *discrète*, lorsque des différences systématiques entre différents régimes sont observées, ou *continue* sur l'ensemble de la zone étudiée.

## Variation spatiale discrète

### ANOVA spatiale

Dans un premier temps, il est possible de se demander dans quelle mesure la moyenne d'une variable varie entre différents sous-groupes d'observations localisées. Par exemple, on voudrait savoir si la moyenne des prix des logements est la même à l'est et à l'ouest d'une ville. Une réponse possible passe par une régression où les variables explicatives sont des variables muettes représentant l'appartenance aux différents sous-groupes.

Formellement, considérons sans perte de généralité un échantillon d'observations partitionné en 2 sous-groupes de localisations<sup>(2)</sup>,  $G_1$  et  $G_2$ . Supposons  $N_1$  observations dans le premier sous-groupe et  $N_2$  observations dans le second sous-groupe avec  $N_1 + N_2 = N$ . Le but est de tester l'hypothèse si  $\alpha_1$ , l'espérance de  $y$  dans  $G_1$  est égale à  $\alpha_2$ , l'espérance de  $y$  dans  $G_2$ . Il suffit alors d'estimer par les Moindres Carrés Ordinaires (MCO) le modèle suivant :

$$(1.3) \quad y_i = \alpha_1 d_{i1} + \alpha_2 d_{i2} + \varepsilon_i \quad i = 1, \dots, N$$

où  $d_{ij}$  sont les variables muettes définies pour  $i = 1, \dots, N$  et  $j = 1, 2$ . Elles sont égales à 1 si la région  $i$  appartient au sous-groupe  $G_j$  et 0 sinon. Sous l'hypothèse de normalité des erreurs, un test d'égalité des moyennes s'effectue alors en testant l'égalité des coefficients  $\alpha_1$  et  $\alpha_2$ .

### Régimes spatiaux

Dans le modèle (1.3), seule la constante varie entre les différents sous-groupes spatiaux. Plus généralement, si l'on considère l'ensemble des coefficients d'une régression, l'hétérogénéité spatiale peut être présente sous la forme de différentes constantes et/ou de différentes pentes. Dans ce cas général, on parle alors d'instabilité structurelle dans l'espace ou de "régimes spatiaux" (Anselin, 1988a).

Formellement, considérons le cas où deux régimes, indiqués par 1 et par 2, sont considérés et où tous les paramètres varient selon le régime :

$$(1.4) \quad \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

où  $y_1$  et  $y_2$  sont les vecteurs des observations des variables dépendantes,  $X_1$  et  $X_2$  sont les matrices des observations des variables explicatives,  $\beta_1$  et  $\beta_2$  sont les vecteurs de coefficients et  $\varepsilon_1$  et  $\varepsilon_2$  sont les vecteurs des erreurs respectivement des sous-groupes 1 et 2. On suppose à nouveau que le nombre total d'observations dans le sous-groupe 1 est  $N_1$  et que celui du sous-groupe 2 est  $N_2$  ( $N_1 + N_2 = N$ ). Les observations dans chaque régime sont supposées être suffisamment nombreuses pour estimer  $\beta_1$  et  $\beta_2$ . Enfin, soit  $\varepsilon' = [\varepsilon_1' \quad \varepsilon_2']$ .

Dans cette première partie, nous supposons que la variance des termes d'erreurs est constante sur l'ensemble des observations :  $E[\varepsilon\varepsilon'] = \sigma^2 I_N$ . Dans ces conditions, le test d'homogénéité des coefficients, c'est-à-dire le test de changement structurel spatial  $\beta_1$  et  $\beta_2$ , peut s'effectuer grâce au test de Chow traditionnel.

### Fonctions spline

L'hétérogénéité spatiale est enfin susceptible de se manifester sous la forme de paramètres de la régression qui diffèrent, non pas selon le régime spatial, mais selon l'intervalle des valeurs prises par une ou plusieurs variables explicatives dans l'aire géographique. Par exemple, conformément aux modèles de la nouvelle économie urbaine, on peut considérer que la densité de population dans une aire urbaine décroît lorsqu'on s'éloigne du

centre de cette aire urbaine. Cependant, cette décroissance n'est pas forcément monotone mais peut varier selon la distance au centre. Il est alors préférable de considérer qu'elle peut décroître de façon différente dans certains intervalles de valeurs de cette distance au centre.

La technique des variables muettes peut à nouveau être mobilisée pour modéliser une fonction dans laquelle les paramètres de la régression diffèrent par intervalle. L'estimation d'une telle fonction, appelée *fonction spline*, nécessite de déterminer le nombre de segments qui divisent l'axe de l'une des variables explicatives  $x$  et d'estimer une fonction linéaire pour chaque segment. Supposons que l'axe des  $x$  soit divisé en  $M$  segments. Les valeurs seuils entre les segments, notées  $\bar{x}_j, j=0, \dots, M$ , sont appelées "nœuds" avec  $x_0 = \min(x)$ .

Pour une observation  $i$ , la fonction *spline*, ou fonction linéaire par morceaux, s'écrit de la façon suivante :

$$(1.5) \quad y_i = \sum_{j=1}^M (a_j + b_j (x_i - \bar{x}_{j-1})) D_j + \varepsilon_i \quad i = 1, \dots, N$$

où  $y$  est la variable dépendante,  $x$  est la variable explicative divisée en  $M$  segments,  $D_j$  est la variable muette prenant la valeur 1 pour les observations telles que  $\bar{x}_{j-1} \leq x < \bar{x}_j$  et 0 sinon. Cette spécification permet d'estimer une fonction linéaire différente pour chaque intervalle défini par les  $M$  nœuds  $\bar{x}_j$ . Ainsi, pour les valeurs de la variable explicative comprises entre les deux premiers nœuds,  $\bar{x}_0$  et  $\bar{x}_1$ , l'équation (1.5) prend la forme suivante :  $y_i = a_1 + b_1 (x_i - \bar{x}_0)$ . Sous les hypothèses habituelles sur les erreurs, cette fonction est estimée par les MCO. En général, la fonction linéaire par morceaux (1.5) est discontinue aux valeurs seuils. Pour s'assurer de la continuité aux nœuds, il est nécessaire d'imposer la contrainte suivante sur les coefficients :  $a_{j+1} = a_j + b_j (\bar{x}_j - \bar{x}_{j-1}), \forall j=1, \dots, M-1$ .

Suits *et alii* (1978) généralisent la formulation (1.5) en posant une fonction polynomiale dont les paramètres diffèrent par intervalle. Utiliser une fonction polynomiale par intervalle permet d'assurer la continuité des dérivées de la fonction aux nœuds et d'employer une forme fonctionnelle plus flexible que des fonctions linéaires. Pour des fonctions cubiques, la fonction est appelée *spline cubique*. Lorsque les segments sont de longueurs égales et pour assurer la continuité aux nœuds, Suits *et alii* (1978) montrent que cette fonction peut être obtenue en estimant la régression suivante par les MCO :

$$(1.6) \quad y_i = a_1 + b_1 (x_i - \bar{x}_0) + c_1 (x_i - \bar{x}_0)^2 + d_1 (x_i - \bar{x}_0)^3 + \sum_{j=1}^{M-1} (d_{j+1} - d_j) (x_i - \bar{x}_j)^3 D_j^* + \varepsilon_i$$

où  $D_j^*$  est la variable muette égale à 1 pour les observations telles que  $x \geq \bar{x}_j$ .

Cette procédure peut être étendue au cas où plusieurs variables explicatives sont segmentées (Suits *et alii*, 1978). Un exemple d'une telle régression où deux variables explicatives sont divisées en 5 intervalles est fourni dans McMillen et McDonald (1998).

Les différentes approches présentées dans ce paragraphe sur la variation spatiale discrète nécessitent la définition préalable des sous-groupes ou des intervalles qui, dans l'idéal, doivent être définis pour correspondre à un ou plusieurs changements structurels géographiques du modèle, connus *a priori*. Ainsi, dans le cas des régimes spatiaux, ces modélisations sont particulièrement adaptées pour des découpages spatiaux bien définis ("Centre" *versus* "Périphérie", rural *versus* urbain, etc.).

En l'absence d'informations précises sur les changements structurels, Anderson (1982, 1985), McMillen (1994) et McMillen et McDonald (1998) choisissent d'estimer des fonctions *spline cubique* avec des segments de longueur égale et des intervalles relativement courts entre les nœuds. Brueckner (1981, 1986) détermine le nombre de ruptures en effectuant un test du rapport de vraisemblance et en calculant des critères d'information pour chaque nouvelle rupture ajoutée.

Il est à noter qu'il n'existe pas à l'heure actuelle de techniques permettant à la fois une estimation endogène des régimes et une prise en compte de l'autocorrélation spatiale (voir *infra*, section 2). En revanche, l'analyse exploratoire des données spatiales peut alors se révéler utile (Haining, 1990 ; Anselin, 1998). Ces techniques, en exploitant explicitement la nature spatiale des données, permettent en effet de caractériser la forme de l'hétérogénéité spatiale en détectant les concentrations locales de valeurs similaires. Par exemple, sur un échantillon de 138 régions européennes, Le Gallo et Ertur (2003) montrent à l'aide de ces techniques que la répartition régionale des PIB par tête est caractérisée par une forte polarisation Nord-Sud. Cette dernière peut alors être interprétée comme une indication de l'existence de 2 régimes spatiaux au sein des régions européennes.

Il est à noter que les techniques présentées dans cette section relèvent de l'économétrie générale et ne tiennent pas explicitement compte de la nature spatiale des données, même si leur adaptation dans un cadre spatial permet de mettre en évidence des interprétations intéressantes. Ainsi, les variables muettes et les régimes sont habituellement utilisés pour modéliser des variations discrètes de la fonction dans un modèle de régression. Ces variations peuvent être de toute nature, telles que des variations temporelles, des variations par âge, par niveau de revenu, etc. (Greene, 2000). Par exemple, les modèles de régression *spline* (1.5) et (1.6) peuvent être facilement adaptés à la modélisation des ruptures temporelles si  $x$  représente le temps et  $\bar{x}_j$  des points représentant des dates particulières (voir par exemple Pindyck et Rubinfeld, 1998, et Marsch et Cormier, 2001, pour des exemples d'application).

Lorsqu'on ne dispose pas d'information sur d'éventuels régimes spatiaux, ou si l'on pense que la moyenne d'une variable ou les coefficients de la régression ne changent pas brutalement d'un régime à l'autre, il est préférable d'utiliser des spécifications permettant de capter une variation spatiale continue sur l'ensemble de la zone étudiée.

### Variation spatiale continue

Deux types d'analyses sont possibles : étude de la variabilité de la moyenne ou de la variabilité des coefficients de régression. Dans le premier cas, on utilise la méthode TSA (*Trend Surface Analysis*) alors que dans le second cas, c'est la méthode VE (*Variable Expansion*) et son extension, la méthode non-paramétrique GWR (*Geographically Weighted Regression*), qui sont les plus adaptées.

#### La méthode *Trend Surface Analysis*

Dans cette méthode, on régresse la variable étudiée sur une expansion polynomiale des coordonnées de chaque localisation, telles que la latitude et la longitude<sup>(3)</sup>. Formellement, le modèle suivant est estimé par les MCO :

$$(1.7) \quad y_i = \sum_{r+s \leq p} a_{rs} u_i^r v_i^s + \varepsilon_i$$

où  $p$  est l'ordre de la TSA,  $u_i$  et  $v_i$  sont les coordonnées de la  $i^{\text{ème}}$  observation,  $a_{rs}$  sont les coefficients inconnus à estimer dont le nombre dépend de l'ordre de la TSA et est égal à  $(p+1)(p+2)/2$ . Ainsi, il y a instabilité structurelle sur les paramètres si tout ou partie des coefficients  $a_{rs}$  sont significatifs.

Par exemple, les deux fonctions associées à l'ordre 1 et à l'ordre 2, respectivement la fonction linéaire ( $p=1$ ) et la fonction quadratique ( $p=2$ ), s'écrivent de la façon suivante :

$$(1.8a) \quad y_i = a_{00} + a_{10} u_i + a_{01} v_i + \varepsilon_i$$

$$(1.8b) \quad y_i = a_{00} + a_{10} u_i + a_{01} v_i + a_{11} u_i v_i + a_{02} u_i^2 + a_{20} v_i^2 + \varepsilon_i$$

Cette méthode permet de décrire les grandes tendances caractéristiques de la surface de régression, comme des tendances simples "Nord-Sud" ou "Est-Ouest" pour la fonction linéaire (représentées par les coefficients  $\alpha_{10}$  et  $\alpha_{01}$  dans (1.8a) et (1.8b)) ou des tendances plus complexes pour les fonctions d'ordre supérieur (représentées par les coefficients  $\alpha_{11}$ ,  $\alpha_{02}$  et  $\alpha_{20}$  dans (1.8b)). Elle sert donc à "lisser" les données (Ripley, 1981 ; Agterberg, 1984). Par ailleurs, comme la variable expliquée est uniquement une fonction des coordonnées des points, des valeurs prévues pour cette variable peuvent facilement être obtenues pour chaque localisation et représentées sur une carte (Johnson et Ragas, 1987).

Ce type de modélisation a, entre beaucoup d'autres applications, servi dans l'étude des fonctions de densité urbaine (Schroeder et Sjoquist, 1976) et dans les modèles hédoniques de prix immobiliers (Olmo, 1995 ; Des Rosiers et Thériault, 1999). En effet, elle permet de représenter graphiquement les surfaces de régression, leurs "pics" ou leurs "vallées", à l'aide de représentations en 3 dimensions dans lesquelles les coordonnées des points constituent l'abscisse et l'ordonnée et la valeur estimée (ou prévue), la cote. Cependant, cette méthode présente plusieurs limites (Ripley, 1981 ; Pace *et alii*, 1998). Tout d'abord, l'autocorrélation spatiale des erreurs n'est pas systématiquement éliminée dans ce type de modèle<sup>(4)</sup>. De plus, les variables des polynômes sont très souvent fortement corrélées, ce qui est source de multicollinéarité. Un enjeu important est donc celui du choix de l'ordre de la TSA. En l'absence de guide théorique, une possibilité est la méthode *stepwise* ascendante qui consiste à estimer d'abord des TSA d'ordres peu élevés, de rajouter ensuite des polynômes d'ordres plus élevés et d'analyser la significativité des coefficients ajoutés (Des Rosiers et Thériault, 1999).



### La méthode Variable Expansion (VE)

La méthode VE, développée par Casetti (1972, 1997), généralise la méthode précédente dans la mesure où l'on suppose que l'hétérogénéité spatiale se traduit par des coefficients de la régression variables pour chaque localisation  $i$ . Cette différenciation est spécifiée par une fonction dépendant d'un certain nombre de variables auxiliaires.

Sans perte de généralité, les propriétés de cette méthode sont illustrées formellement par un exemple simple avec une seule variable explicative. Le modèle initial est, pour une observation  $i$  :

$$(1.9) \quad y_i = \beta_0 + \beta_{1i} x_i + \varepsilon_i$$

$\beta_0$  et  $\beta_{1i}$  sont les coefficients inconnus de la régression et  $x_i$  correspond à la  $i^{\text{ème}}$  observation de la variable explicative. On suppose que le second coefficient est variable pour chaque observation <sup>(5)</sup> :  $\beta_{1i}$  est une fonction exacte de variables d' "expansion" ou variables augmentées, notées  $z_i$ ,  $i = 1, \dots, P$  :

$$(1.10) \quad \beta_{1i} = f(z_i, \gamma)$$

où  $z_i$  est un vecteur de  $P$  variables augmentées,  $\gamma$  est le vecteur des  $P$  paramètres inconnus correspondants,  $f$  est la relation fonctionnelle qui exprime la forme de la variation du coefficient  $\beta_{1i}$ . En général, cette fonction  $f$  est linéaire. Ainsi, pour deux variables augmentées et une relation fonctionnelle linéaire, la relation (1.10) s'écrit de la façon suivante :

$$(1.11) \quad \beta_{1i} = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i}$$

L'introduction des variables augmentées dans le modèle initial donne le modèle terminal suivant, estimé par les MCO sous les hypothèses habituelles :

$$(1.12) \quad y_i = \beta_0 + \gamma_0 x_i + \gamma_1 (z_{1i} x_i) + \gamma_2 (z_{2i} x_i) + \varepsilon_i$$

Sous forme matricielle, le modèle (1.12) s'écrit de la façon suivante :

$$(1.13) \quad y = S\beta_0 + \gamma_0 x + \gamma_1 (z_1 x) + \gamma_2 (z_2 x) + \varepsilon$$

où  $z_1 = \text{diag}(z_{1i})$ ,  $z_2 = \text{diag}(z_{2i})$ , et  $S$  est le vecteur somme.

Si le modèle terminal est la spécification correcte, les estimateurs des paramètres du modèle initial sont biaisés. En effet, il s'agit d'un cas particulier du problème traditionnel de variables omises. Plus précisément, soient  $b' = [\beta_0 \ \gamma_0]$ ,  $X = [S \ x]$ ,  $Z = [z_1 \ x \ z_2 \ x]$  et  $\gamma' = [\gamma_1 \ \gamma_2]$ . L'espérance de l'estimateur de  $b$  du modèle initial, si le modèle final est vrai, est  $E(\hat{b}) = b + (X'X)^{-1} X'Z\gamma$ . L'estimateur de  $b$  dans le modèle initial est donc biaisé (Anselin, 1988a ; Anselin et Griffith, 1988).

Comme dans la méthode TSA, les variables  $z_i$  correspondent le plus souvent aux coordonnées du point  $i$  (latitude et longitude). Dans le cas linéaire le plus simple, seules des tendances simples dans la variation des paramètres sont captées (tendance du type "Nord/Sud" ou "Est/Ouest") et des expansions quadratiques ou d'ordre supérieur doivent être spécifiées pour capter des tendances plus complexes. A nouveau, plus l'ordre de l'expansion est grand, plus la multicolinéarité risque de poser problème <sup>(6)</sup>.

Outre les coordonnées des observations, les variables augmentées peuvent plus généralement correspondre à toutes les variables susceptibles d'être à l'origine de la différenciation spatiale des coefficients. Par exemple, LeSage (1999) utilise la distance au centre de la ville pour modéliser la variation spatiale des prix marginaux des caractéristiques dans un modèle hédonique de prix immobiliers. Toujours dans le même cadre, Can (1990, 1992) préfère utiliser une variable composite de la "qualité" du voisinage de chaque logement : revenu moyen, composition raciale de la population, taux de chômage, etc. <sup>(7)</sup>

La méthode VE permet de modéliser des paramètres variables dans l'espace mais elle souffre de deux limites principales (Fotheringham *et alii*, 2000). La première est que cette technique ne peut capter que des tendances dans les relations dans l'espace, la complexité de ces tendances étant dépendante de la complexité des équations d'expansion spécifiées (tendances linéaires, quadratiques, etc.). Les estimations des paramètres obtenues à partir de cette méthode ne permettent donc pas toujours de capter d'éventuelles importantes variations locales. La seconde est que la forme des équations d'expansion doit être supposée *a priori*. Pour pallier ces problèmes, trois auteurs, Brundson, Fotheringham et Charlton, ont développé la méthode d'estimation non-paramétrique GWR (*Geographically Weighted Regression*).

### La méthode non-paramétrique Geographically Weighted Regression (GWR)

La méthode GWR est présentée dans un ouvrage (Fotheringham *et alii*, 2000) et dans une série d'articles (Brundson *et alii*, 1996, 1998, 1999 ; Fotheringham et Brundson, 1999 ; Fotheringham *et alii*, 1996a, 1996b, 1998). Cette méthode non-paramétrique est destinée à capter au niveau de chaque observation les variations des coefficients d'une régression dans l'espace. Pour cela, un paramètre différent est estimé pour chaque observation en utilisant les valeurs des caractéristiques prises par les observations voisines.

Formellement, le point de départ est à nouveau la formulation générale (1.2) dans laquelle il existe un vecteur de  $K$  paramètres inconnus à estimer pour chaque observation  $i$  :

$$(1.14) \quad y_i = x_i' \beta_i + \varepsilon_i \Leftrightarrow y_i = \sum_k \beta_{ki} x_{ki} + \varepsilon_i$$

Pour estimer les paramètres  $\beta_{ki}$  du modèle (1.14), on suppose que les données observées près du point  $i$  ont plus d'influence dans l'estimation de  $\beta_{ki}$  que les données situées loin de ce point  $i$ . On utilise alors une pondération variable avec le point  $i$ . Soit  $\hat{\beta}_i$  l'estimateur des Moindres Carrés Pondérés du vecteur des  $K$  coefficients pour l'observation  $i$ . Il s'écrit :

$$(1.15) \quad \hat{\beta}_i = (X'V_i X)^{-1} X'V_i Y$$

où  $X$  est la matrice des observations des variables explicatives et  $Y$  représente le vecteur des observations de la variable dépendante. Enfin,  $V_i$  est une matrice diagonale ( $N, N$ ), variable pour chaque observation  $i$ , dont les éléments de la diagonale principale représentent le poids géographique accordé aux données autour du point  $i$  :

$$(1.16) \quad V_i = \text{diag}(v_{i1}, \dots, v_{in})$$

où  $v_{ij}$  dénote le poids de la  $j^{\text{ème}}$  observation sur l'estimation du modèle autour du  $i^{\text{ème}}$  point.

Cette méthodologie diffère de la régression non-paramétrique habituelle du noyau dans laquelle les pondérations se réfèrent à l'espace des *attributs* de  $X$  (Cleveland, 1979 ; Cleveland et Devlin, 1988, Cleveland *et alii*, 1988). En revanche, la méthode GWR utilise des pondérations basées sur l'espace *géographique*. Selon les auteurs, différents schémas de pondération, ou "noyaux", sont utilisés (McMillen, 1996 ; MacMillen et McDonald, 1997 ; Fotheringham *et alii*, 2000 ; Pavlov, 2000).

Les auteurs présentent la méthode GWR comme une extension de la méthode VE. Par ailleurs, elle peut également s'interpréter comme une méthode essentiellement exploratoire permettant d'identifier la nature et les schémas d'hétérogénéité spatiale sur l'ensemble de la zone étudiée. En effet, le résultat d'une GWR est un ensemble d'estimations localisées des paramètres, ainsi que des versions localisées de mesures de qualité de la régression. Ces estimations locales étant toutes associées à des localisations spécifiques, chaque ensemble de paramètres peut être cartographié pour illustrer les variations spatiales de la relation mesurée. Pour savoir si ces estimations locales des paramètres sont significativement différentes entre elles et par rapport à l'estimateur des MCO, des tests paramétriques ont été proposés par Brundson *et alii* (1999) et Leung *et alii* (2000).

Les méthodes présentées dans cette section sur la variation spatiale continue ont été principalement développées par des géographes. Par exemple, la méthode TSA a été initialement introduite pour étudier les tendances spatiales des phénomènes géologiques, comme les surfaces d'érosion. À ce titre, elles n'ont la plupart du temps pas de fondements économiques et ne sont pas reliées à des modèles théoriques. En revanche, elles s'avèrent particulièrement utiles dans une perspective d'analyse exploratoire des données car elles permettent de détecter des grandes tendances et peuvent alors constituer une première étape pour une phase ultérieure de modélisation.

### Instabilité spatiale des paramètres et hétéroscédasticité

La deuxième forme d'hétérogénéité spatiale, l'hétéroscédasticité, se manifeste par la variabilité des variances des termes d'erreurs selon la localisation. Elle constitue un problème courant en coupe transversale et provient généralement de variables manquantes ou de toute autre forme de mauvaise spécification du modèle. Elle est particulièrement présente lorsqu'on travaille sur des données localisées, les unités spatiales utilisées ne sont généralement ni de formes régulières, ni homogènes. Les méthodes de traitement de l'hétéroscédasticité sont aujourd'hui bien connues (Greene, 2000). On peut ainsi utiliser l'approche de White (1980) qui a fourni un estimateur convergent de la matrice des variances-covariances de l'estimateur des MCO en présence d'hétéroscédasticité de forme inconnue pour que l'inférence statistique basée sur les MCO soit asymptotiquement fiable. Alternativement, lorsqu'on choisit une forme particulière pour l'hétéroscédasticité, des tests d'hétéroscédasticité plus puissants que le test de White peuvent être obtenus (tel que le test de



Breusch-Pagan, 1979). Pour des données localisées, les spécifications d'hétéroscédasticité les plus utilisées sont l'hétéroscédasticité en groupes (lorsque la variance des erreurs est différente pour chaque régime spatial) ou encore l'expression de la variance comme une fonction d'un ensemble d'autres variables. En présence d'hétéroscédasticité, les estimateurs fournis par les MCO sont non biaisés mais inefficients. En d'autres termes, l'inférence statistique obtenue à partir des MCO n'est pas fiable, comme par exemple les résultats obtenus par les tests visant à détecter l'instabilité structurelle.

Dans cet article, nous considérons uniquement les cas où l'hétéroscédasticité et une certaine forme d'instabilité spatiale des paramètres sont présents conjointement dans une régression. L'instabilité spatiale peut être modélisée, par exemple par la spécification de régimes spatiaux sous la forme (1.4). Pourtant, lorsque l'instabilité spatiale n'est pas complètement prise en compte de cette façon, il peut subsister une hétéroscédasticité qui représente l'hétérogénéité non observée. Ainsi, il est possible qu'une hétéroscédasticité résiduelle se manifeste dans le cadre d'un modèle avec régimes spatiaux. Nous présentons ce cas dans un premier point. En revanche, la situation est différente dans le cadre du modèle VE dans lequel l'hétéroscédasticité des erreurs est induite par la structure même du modèle. Ce cas est abordé dans un second point.

### Régimes spatiaux et hétéroscédasticité

Reprenons les notations du modèle (1.4) où deux régimes spatiaux sont considérés :

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

Quandt (1958) suppose en outre que les termes d'erreurs ont une variance différente pour chaque sous-ensemble (hétéroscédasticité en groupes) :

$$(1.17) \quad \Omega = E[\varepsilon\varepsilon'] = \begin{bmatrix} \sigma_1^2 I_{N_1} & 0 \\ 0 & \sigma_2^2 I_{N_2} \end{bmatrix}$$

où  $\sigma_1^2$  est la variance associée au premier régime,  $\sigma_2^2$  est la variance associée au second régime,  $N_1$  et  $N_2$  sont respectivement le nombre d'observations dans le premier et dans le deuxième régime, avec  $N_1 + N_2 = N$ ,  $I_{N_i}$  étant la matrice identité d'ordre  $N_i$ ,  $i = 1, 2$ .

Dans ce cas, l'estimation peut être basée sur la fonction de vraisemblance du modèle global qui est de la forme suivante (en ignorant le terme constant) :

$$(1.18) \quad L = -\left(\frac{N_1}{2}\right) \ln(\sigma_1)^2 - \left(\frac{N_2}{2}\right) \ln(\sigma_2)^2 - \frac{(y_1 - X_1 \beta_1)' (y_1 - X_1 \beta_1)}{2(\sigma_1)^2} - \frac{(y_2 - X_2 \beta_2)' (y_2 - X_2 \beta_2)}{2(\sigma_2)^2}$$

Cette fonction de log-vraisemblance peut être modifiée pour tenir compte de structures d'erreurs plus complexes (comme par exemple l'incorporation d'une autocorrélation spatiale des erreurs, voir *infra*, deuxième partie).

### Variation spatiale continue et hétéroscédasticité

Certains auteurs (Anselin, 1988a ; Casetti et Can, 1999) ont souligné que la méthode VE était également susceptible d'être affectée par un problème d'hétéroscédasticité. En effet, dans ce modèle (équations (1.9) à (1.13)), on a supposé que la forme exacte de l'expansion est connue *a priori*. Pourtant, en pratique, l'hypothèse d'une relation exacte (déterministe) entre les coefficients et leurs variables augmentées est difficile à tenir (Anselin, 1988a). Il est donc plus pertinent de rajouter un terme aléatoire dans l'expansion linéaire (1.11) :

$$(1.19) \quad \beta_{1i} = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + v_i$$

où  $v_i$  est un terme aléatoire normalement et indépendamment distribué d'espérance nulle et de variance  $\sigma_v^2$ . La substitution de cette expression dans le modèle initial conduit au modèle terminal suivant :

$$(1.20) \quad y = S\beta_0 + \gamma_0 x + \gamma_1 (z_1 x) + \gamma_2 (z_2 x) + vx + \varepsilon$$

Soit  $w$  le nouveau terme aléatoire :  $w = vx + \varepsilon$ . Cette erreur est normalement et indépendamment distribuée d'espérance nulle et de matrice des variances-covariances  $V(w) = \sigma_v^2 \text{diag}(x_i^2) + \sigma_\varepsilon^2 I_N$  (si les erreurs  $v$  et  $\varepsilon$  sont indépendantes).

Le modèle final induit donc une certaine forme d'hétéroscédasticité dont il faut tenir compte pour l'inférence statistique, en particulier si l'on cherche à tester la significativité des coefficients  $\gamma_1$  et  $\gamma_2$ . Cette forme particulière d'hétéroscédasticité peut être testée avec le test de Breusch-Pagan (1979). Cette approche revient alors à tester la version aléatoire contre la version déterministe de la méthode VE : sous l'hypothèse nulle d'homoscédasticité, on retrouve le modèle (1.13) alors que sous l'hypothèse alternative d'hétéroscédasticité, on retrouve le modèle (1.20).

---

## L'interaction entre autocorrélation et hétérogénéité spatiales

---

Outre l'hétérogénéité spatiale, les données localisées sont très souvent caractérisées par un phénomène d'autocorrélation spatiale. L'autocorrélation spatiale se réfère à l'absence d'indépendance entre observations géographiques. Ainsi, une autocorrélation spatiale positive se traduit par une tendance à la concentration dans l'espace de valeurs faibles ou élevées d'une variable aléatoire. En revanche, l'autocorrélation spatiale négative signifie que chaque localisation tend à être entourée par des localisations voisines pour lesquelles la variable aléatoire prend des valeurs très différentes. La présence d'autocorrélation spatiale pour une variable signifie qu'il existe une relation fonctionnelle entre ce qui se passe en un point de l'espace et ce qui se passe ailleurs et peut provenir de deux sources différentes. D'une part, elle peut provenir du fait que les données sont affectées par des processus qui relient des lieux différents et qui sont à l'origine d'une organisation spatiale particulière des activités. D'autre part, elle peut également provenir d'une mauvaise spécification du modèle.

Lorsque les deux effets spatiaux, autocorrélation et hétérogénéité spatiales, sont présents conjointement dans une régression, les méthodes d'estimation et d'inférence présentées dans la section précédente doivent être adaptées. Notre présentation s'organise en trois points. Tout d'abord, nous montrons comment l'autocorrélation spatiale et l'hétéroscédasticité sont nécessairement liées dans les principaux modèles destinés à capter l'autocorrélation spatiale. Ensuite, nous détaillons les autres liens qui unissent les deux effets. Enfin, nous présentons la façon dont les tests d'hétéroscédasticité et d'instabilité structurelle doivent être modifiés pour tenir compte de l'autocorrélation spatiale lorsqu'elle est présente.

### Autocorrélation spatiale et hétéroscédasticité

Plusieurs types de modèles permettent de capter ce phénomène (Le Gallo, 2002a, 2002b). Nous rappelons ici brièvement les deux principaux modèles utilisés dans la littérature.

Le premier est le *modèle spatial autorégressif* d'ordre 1 qui s'écrit :

$$(2.1) \quad y = \rho W y + X \beta + \varepsilon$$

où  $N$  est le nombre d'observations,  $y$  est le vecteur  $(N,1)$  des observations de la variable dépendante,  $X$  est la matrice  $(N,K)$  des observations des variables exogènes,  $K$  étant le nombre de paramètres inconnus,  $\beta$  est le vecteur  $(K,1)$  des coefficients de régression inconnus,  $\varepsilon$  est le vecteur  $(N,1)$  tel que  $\varepsilon \sim \text{iid}(0, \sigma^2 I_N)$ ,  $W$  est la matrice de poids et  $\rho$  est le paramètre spatial autorégressif indiquant l'intensité de l'interaction spatiale existant entre les observations de  $y$ .

La matrice de poids  $W$  permet de spécifier l'autocorrélation spatiale. Il s'agit d'une matrice carrée, de dimension  $N$ , où chaque terme  $w_{ij}$  représente la façon dont la région  $i$  et la région  $j$  sont connectées spatialement. Par exemple, dans le cas d'une matrice de contiguïté, chaque terme de cette matrice est égal à 1 si les régions possèdent une frontière commune et 0 sinon. D'autres spécifications font intervenir des mesures de distance entre les observations. Cette matrice de poids est souvent standardisée : chaque élément  $w_{ij}$  de la matrice est divisé par la somme totale de chaque ligne.

Le modèle (2.1) est habituellement estimé par la méthode du maximum de vraisemblance ou la méthode des variables instrumentales. Si la matrice  $(I - \rho W)$  est inversible, il se réécrit sous forme réduite de la façon suivante :

$$(2.2) \quad y = (I - \rho W)^{-1} X \beta + (I - \rho W)^{-1} \varepsilon$$

Posons  $u = (I - \rho W)^{-1} \varepsilon$ . La matrice de variances-covariances de  $u$  s'écrit alors :

$$(2.3) V(u) = \sigma^2 [(I - \rho W)'(I - \rho W)]^{-1}$$

De (2.3), il s'ensuit que les éléments de la diagonale de  $V(u)$  ne sont pas constants, ce qui implique l'hétéroscédasticité des erreurs  $u$ , que  $\varepsilon$  soit hétéroscédastique ou non.

Le second est le *modèle avec autocorrélation spatiale des erreurs* :

$$(2.4) y = X\beta + \varepsilon$$

$$\varepsilon = \lambda W\varepsilon + u$$

Le paramètre  $\lambda$  reflète l'intensité de l'interdépendance entre les résidus et  $u$  est le terme d'erreur tel que :  $u \sim \text{idd}(0, \sigma^2 I_N)$ . Le modèle (2.4) s'estime habituellement par la méthode du maximum de vraisemblance ou la méthode des moments généralisés.

Si la matrice  $(I - \lambda W)$  est inversible, alors la matrice des variances-covariances du terme d'erreur  $\varepsilon$  s'écrit de la façon suivante :

$$(2.5) V(\varepsilon) = \sigma^2 [(I - \lambda W)'(I - \lambda W)]^{-1}$$

À nouveau, cette formulation (2.5) indique que les erreurs  $\varepsilon$  sont hétéroscédastiques (McMillen, 1992). Hétéroscédasticité et autocorrélation spatiales sont donc toujours liées dans les principaux modèles destinés à capter l'autocorrélation spatiale. Plus généralement, d'autres liens unissent encore hétérogénéité et autocorrélation spatiales.

### **Les liens entre hétérogénéité et autocorrélation spatiales**

La complexité du lien entre autocorrélation spatiale et hétérogénéité spatiale se retrouve à plusieurs niveaux.

Premièrement, il n'est pas toujours facile de distinguer les conséquences de l'autocorrélation de celles de l'hétérogénéité spatiales dans une coupe transversale : une équivalence observationnelle entre ces deux effets peut exister (Anselin et Bera, 1998 ; Anselin, 2000). Par exemple, l'observation d'un regroupement spatial de valeurs très élevées des erreurs estimées peut provenir d'une hétérogénéité (qui se manifeste par exemple sous la forme d'une hétéroscédasticité en groupes) ou d'une d'autocorrélation spatiale (c'est-à-dire un processus stochastique impliquant le regroupement de points extrêmes).

Deuxièmement, les tests pour l'hétéroscédasticité et pour l'instabilité structurelle ne sont pas fiables en présence d'autocorrélation spatiale. Des simulations de Monte-Carlo effectuées par Anselin et Griffith (1988) montrent, par exemple, que la présence d'autocorrélation spatiale des erreurs affecte le seuil et la puissance des tests de White (1980) et de Breusch-Pagan (1979). En particulier, sous l'hypothèse nulle d'absence d'hétéroscédasticité, les fréquences empiriques de rejet du test de White et de Breusch-Pagan sont inférieures au niveau de significativité choisi en présence d'autocorrélation spatiale. Les puissances des tests sont également affectées : elles diminuent pour des valeurs positives du paramètre spatial. Réciproquement, les tests d'autocorrélation spatiale voient leurs propriétés affectées lorsqu'il y a hétéroscédasticité de forme inconnue. Le test de Chow, permettant de tester l'instabilité structurelle et basé sur les résidus de l'estimation par les MCO, est également affecté par la présence d'autocorrélation spatiale des erreurs, particulièrement dans le cas de régimes correspondant à deux ensembles compacts d'unités spatiales contiguës (configuration de type Est-Ouest ou Nord-Sud par exemple). Les simulations de Monte-Carlo réalisées par Anselin (1990a) montrent, en effet, que dans ce cas l'hypothèse nulle d'homogénéité régionale est trop souvent rejetée pour des valeurs positives du paramètre spatial alors que la puissance du test augmente presque exponentiellement. La présence jointe d'hétéroscédasticité et de dépendance spatiale nécessite donc des tests spécialisés et des méthodes d'estimation adaptées (Anselin, 1988a, 1988b).

Troisièmement, l'autocorrélation spatiale est parfois le résultat d'une instabilité des paramètres non modélisée (Brundson *et alii*, 1999). En d'autres termes, si des relations variables dans l'espace sont modélisées à l'aide d'une régression " globale " (spécifiée de la même façon pour toutes les observations), les termes d'erreurs peuvent être spatialement autocorrélés.

Par exemple, supposons une zone géographique pour laquelle  $y$  est régressé sur  $x$  et que le modèle global estimé qui en résulte est :

$$(2.6) \hat{y}_i = 1,5 + 2x_i$$

Supposons que la vraie relation n'est pas la même dans les deux régions de cette zone :

$$\text{région 1 : } y_1 = 1,5 + 2,5x_i \quad (2.7a)$$

$$\text{région 2 : } y_1 = 1,5 + 1,5x_i \quad (2.7b)$$

Dans ces conditions, appliquer l'équation (2.6) à la partie de la zone pour laquelle l'équation (2.7a) est valable conduit à la sous-estimation des valeurs de  $y_i$  dans cette zone et des erreurs estimées positives (pour des valeurs positives de  $x_i$ ). Réciproquement, appliquer l'équation (2.6) à la partie de la zone pour laquelle l'équation (2.7b) est valable conduit à la surestimation des valeurs de  $y_i$  dans cette région et des erreurs estimées négatives. Une forte autocorrélation spatiale des erreurs apparaît donc, résultant de l'incapacité du modèle global à traiter l'instabilité structurelle de la relation.

Dans le même ordre d'idée, la modélisation de l'hétérogénéité par l'utilisation de variables augmentées est susceptible d'éliminer l'autocorrélation spatiale des erreurs estimées du modèle initial (Jones, 1983, Casetti et Jones, 1988). Anselin (1988a) et Anselin et Griffith (1988) montrent en effet que, si des variables augmentées sont omises à tort, le test de l'autocorrélation spatiale de Moran, basé sur les erreurs estimées du modèle initial mal spécifié, peut indiquer à tort la présence d'autocorrélation spatiale.

Tous ces éléments suggèrent que les liens entre hétérogénéité et autocorrélation spatiales sont forts et complexes. Les deux aspects ne peuvent pas être considérés indépendamment l'un de l'autre et ils doivent être spécifiés précisément pour pouvoir identifier les paramètres du problème et tester chaque type de mauvaise spécification. Par la suite, nous présentons la façon dont les tests d'hétéroscédasticité et d'instabilité structurelle doivent être adaptés à la présence d'autocorrélation spatiale <sup>(8)</sup>.

### **Le test de l'hétéroscédasticité en présence d'autocorrélation spatiale**

En présence d'autocorrélation spatiale, les propriétés des tests d'hétéroscédasticité ne sont plus valables et, face à ce problème, deux stratégies sont possibles (Anselin, 1988a, 1988b).

*La première stratégie* consiste tout d'abord à faire un test joint d'hétéroscédasticité et d'autocorrélation spatiale des erreurs et ensuite, si l'hypothèse nulle est rejetée, à tester séparément ces deux effets.

Formellement, considérons le modèle avec autocorrélation spatiale des erreurs et hétéroscédasticité sur le terme d'erreur résiduel :

$$(2.8) \quad y = X\beta + \varepsilon$$

$$\varepsilon = \lambda W\varepsilon + u \quad u \sim N(0, \Omega)$$

où les éléments de la diagonale principale de  $\Omega$  sont de la forme suivante :  $\sigma_i^2 = \sigma^2 f(\alpha_0 + \alpha_1 z_{1i} + \dots + \alpha_p z_{pi})$ .

Ce modèle peut s'estimer avec la méthode du maximum de vraisemblance. L'hypothèse jointe est  $H_0 : \lambda = 0$  et  $\alpha_j = 0, j = 1, \dots, P$  dans (2.8). La statistique du multiplicateur de Lagrange est la suivante (Anselin, 1988a) <sup>(9)</sup> :

$$(2.9) \quad JLM = (1/2) f' Z (Z' Z)^{-1} Z' f + (1/T^*) [\hat{\varepsilon}' W \hat{\varepsilon} / \hat{\sigma}^2]^2$$

où  $Z$  est la matrice  $(N, P + 1)$  comportant les observations sur  $(1, z_i)$ ,  $f$  est le vecteur  $(N, 1)$  comportant les observations sur  $f_i = (\hat{\sigma}^{-1} \hat{\varepsilon}_i) - 1$ ,  $T^* = tr\{(W' + W)W\}$ ,  $\hat{\varepsilon}$  et  $\hat{\sigma}^2$  sont respectivement les erreurs estimées et la variance estimée du modèle de régression standard estimé par les MCO. Cette statistique converge asymptotiquement vers une loi du *chi-deux* à  $P + 1$  degrés de liberté.

Dans ces conditions, un rejet de l'hypothèse nulle jointe doit être suivi par un test de chacun des deux cas particuliers. Cette approche séquentielle a fait l'objet d'une simulation par Anselin et Griffith (1988). Il apparaît que la puissance du test JLM est meilleure que les tests individuels lorsque autocorrélation spatiale et hétéroscédasticité sont présentes conjointement. En revanche, la deuxième étape, consistant à faire les tests

individuels si l'hypothèse nulle du test joint est rejetée, ne fournit pas de bons résultats<sup>(10)</sup>. Une stratégie alternative s'avère donc nécessaire.

La seconde stratégie consiste à effectuer un test d'hétéroscédasticité en présence d'autocorrélation spatiale des erreurs. On teste donc  $H_0 : \alpha_j = 0, j = 1, \dots, P$  dans (2.8). Par conséquent, sous l'hypothèse nulle, ce modèle devient un modèle homoscedastique avec autocorrélation spatiale des erreurs (2.4). Sous l'hypothèse nulle, la statistique du test (noté BPS comme "Breusch-Pagan Spatial") s'écrit de la façon suivante :

$$(2.10) \text{ BPS} = (1/2) f' Z^{-1} [Z' DZ]^{-1} Z' f$$

avec :  $D_{(N,N)} = I - (1/2\hat{\sigma}^4) d\hat{V}d'$ ,  $d_{(N,2)} = [S 2\hat{\sigma}^2 w]$ ,  $w_{(N,1)}$  est le vecteur composé des éléments diagonaux de  $W(I - \hat{\lambda}W)^{-1}$ ,  $\hat{V}_{(N,N)}$  est la matrice estimée des variances-covariances pour  $\hat{\sigma}^2$  et  $\hat{\lambda}$ . Cette statistique converge asymptotiquement vers une loi du *chi-deux* à  $P$  degrés de liberté.

Ce test s'effectue après avoir estimé le modèle avec autocorrélation des erreurs (2.4) alors que le test précédent (2.9) pouvait s'effectuer à partir des erreurs estimées des MCO du modèle de régression standard. Un test de Breusch-Pagan spatialement ajusté peut aussi être dérivé pour le modèle avec variable endogène décalée et hétéroscédasticité. Cependant, les propriétés de ces deux tests en échantillon fini ne sont pas connues.

### Le test de l'instabilité des paramètres en présence de dépendance spatiale

Lorsque les hypothèses d'homoscedasticité et de non-corrélation ne sont pas vérifiées dans les modèles ANOVA (1.3), TSA (1.7) et VE (1.12), les MCO fournissent des estimateurs non convergents (cas d'une variable spatiale autorégressive) ou inefficients (cas d'une autocorrélation spatiale des erreurs). Par exemple, s'il y a autocorrélation spatiale positive des erreurs dans un modèle ANOVA spatial (1.3), la variance intra-groupe est sous-estimée et l'hypothèse nulle d'égalité entre les moyennes des différentes zones géographiques est rejetée trop souvent (Cliff et Ord, 1981 ; Legendre *et alii*, 1990 ; Sokal *et alii*, 1993). Il faut donc estimer le modèle (1.3) grâce à une méthode adaptée dans le but d'ajuster les sommes des carrés associés aux variances intra- et inter-groupe (Griffith, 1978). D'une façon générale, il faut tester dans ce modèle ainsi que dans les autres modèles d'instabilité des paramètres la présence d'hétéroscédasticité et d'autocorrélation spatiale et si nécessaire les estimer par des méthodes appropriées (Griffith, 1978, 1992).

Dans le cas des régimes spatiaux et sous l'hypothèse d'homoscedasticité, le test de Chow, basé sur une statistique  $F$ , permet de tester l'instabilité structurelle. En revanche, la présence d'autocorrélation spatiale nécessite son adaptation (Anselin, 1990a).

Formellement, reprenons les notations du modèle (1.4) avec deux régimes :

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

Soit  $\varepsilon' = [\varepsilon_1' \varepsilon_2']$  et  $\Psi$  sa matrice des variances-covariances :  $\Psi = E(\varepsilon\varepsilon')$ . On veut tester l'hypothèse d'homogénéité régionale, soit :  $H_0 : \beta_1 = \beta_2$ . La forme de la statistique du test dépend des hypothèses faites sur les erreurs. Lorsque la variance est constante sur l'ensemble des observations, le test traditionnel de Chow s'applique. En revanche, lorsque les erreurs ne sont pas homoscedastiques, le test d'instabilité structurelle doit être établi sur des bases asymptotiques. Pour  $\Psi = \sigma^2 \Omega$ , l'expression correspondante est :

$$(2.11) C_G = \frac{\hat{e}'_c \hat{\Omega}^{-1} \hat{e}'_c - \hat{e}'_L \hat{\Omega}^{-1} \hat{e}_L}{\hat{\sigma}^2}$$

où  $\hat{e}_c$  est le vecteur des erreurs estimées du modèle contraint et  $\hat{e}_L$  celui du modèle libre. Cette statistique suit asymptotiquement une loi du *chi-deux* à  $K$  degrés de liberté,  $K$  étant le nombre de paramètres du modèle. De la même manière, on peut effectuer ce test sur chacun des coefficients du modèle.

L'encadré 1 résume les principaux cas combinant instabilité structurelle, hétéroscédasticité et autocorrélation spatiale.

En pratique, Anselin (1990a) suggère de commencer par effectuer un test destiné à détecter la présence d'autocorrélation spatiale. Différents tests et différentes règles de décision ont été proposés permettant de

déterminer la forme prise par l'autocorrélation spatiale : autocorrélation des erreurs ou variable endogène décalée (Anselin *et alii*, 1996 ; Le Gallo, 2002a, 2002b). S'il existe une forte indication d'autocorrélation spatiale, plus particulièrement pour des valeurs positives et des régimes correspondant à des observations compactes et contiguës, les techniques standard ne sont pas fiables et d'autres méthodes d'estimation doivent être mobilisées (méthode du maximum de vraisemblance, méthode des variables instrumentales ou des moments généralisés). Cette approche empirique peut cependant poser des problèmes de pretest, qui invalident la distribution asymptotique usuelle des différents tests. Dans ces conditions, il est prudent de choisir des niveaux de significativité restrictifs pour les tests.

### Encadré 1 : instabilité structurelle, hétéroscédasticité et autocorrélation spatiale

*1<sup>er</sup> cas : hétéroscédasticité en groupes*

Dans ce cas, la matrice des variances-covariances de  $\varepsilon$  s'écrit :

$$[1] \Psi = E[\varepsilon\varepsilon'] = \begin{bmatrix} \sigma_1^2 I_{N_1} & 0 \\ 0 & \sigma_2^2 I_{N_2} \end{bmatrix}$$

Lorsque l'autocorrélation spatiale est présente dans les erreurs, quatre situations différentes peuvent être distinguées selon la structure de la variance  $\Psi$  dans le modèle [1.4]. Dans les deux premières, la dépendance des erreurs est liée au système spatial complet alors que dans les deux dernières, la structure de la dépendance spatiale est différente pour chaque régime.

*2<sup>ème</sup> cas : homoscedasticité et autocorrélation spatiale*

Supposons une autocorrélation spatiale autorégressive pour le terme d'erreur complet  $\varepsilon$  :

$$[2] \varepsilon = \lambda W\varepsilon + u$$

où  $\lambda$  est le paramètre autorégressif et  $u$  un vecteur de termes d'erreurs indépendants. Soit  $B = (I - \lambda W)$ . On a alors  $\varepsilon = B^{-1}u$ . Si les erreurs  $u$  sont homoscedastiques, soit  $E(uu') = \sigma^2 I$ , la matrice des variances-covariances de  $\varepsilon$  prend alors la forme :

$$[3] \Psi = E[\varepsilon\varepsilon'] = \sigma^2 (B' B)^{-1}$$

*3<sup>ème</sup> cas : hétéroscédasticité en groupes et autocorrélation spatiale*

Reprenons la structure [2] pour  $\varepsilon$  mais supposons le modèle hétéroscédastique où chaque régime possède une variance des erreurs différentes. Dans ce cas, la matrice des variances-covariances pour  $u$  devient :

$$[4] \Omega = E[uu'] = \begin{bmatrix} \sigma_1^2 I_{N_1} & 0 \\ 0 & \sigma_2^2 I_{N_2} \end{bmatrix} \Rightarrow \Psi = E[\varepsilon\varepsilon'] = B^{-1} \Omega (B')^{-1}$$

*4<sup>ème</sup> cas : processus spatiaux différents entre sous-groupes indépendants*

Lorsque la structure de dépendance spatiale est différente pour chaque sous-régime, les erreurs  $\varepsilon_1$  et  $\varepsilon_2$  dans [1.4] suivent chacune un processus spatial différent, reflétés par les matrices de poids  $W_1$  et  $W_2$  respectivement :

$$[5] \varepsilon_1 = \lambda_1 W_1 \varepsilon_1 + u_1$$

$$\varepsilon_2 = \lambda_2 W_2 \varepsilon_2 + u_2$$

où  $\lambda_1$  et  $\lambda_2$  sont les coefficients spatiaux pour chaque régime.

Supposons tout d'abord que les erreurs des deux sous-régimes sont indépendantes :  $E[u_{h1} u_{k2}] = 0$  pour tout  $(h, k)$  appartenant aux deux sous-régimes. Alors la matrice des variances-covariances de chaque sous-régime est de la forme :  $\Psi_i = \sigma_i^2 (B_i' B_i)^{-1}$  pour  $i = 1, 2$ . La matrice des variances-covariances totale s'écrit :

$$[6] \Psi = E[\varepsilon\varepsilon'] = \begin{bmatrix} \sigma_1^2 (B_1' B_1)^{-1} & 0 \\ 0 & \sigma_2^2 (B_2' B_2)^{-1} \end{bmatrix}$$

*5<sup>ème</sup> cas : processus spatiaux différents entre sous-groupes dépendants*

Reprenons la structure [5] pour les erreurs. Lorsqu'il y a autocorrélation spatiale à la fois à l'intérieur et entre les différents régimes, alors les erreurs entre les deux sous-régimes ne sont plus indépendantes :  $E[u_{h1} u_{k2}] = \sigma_{12}$ . Dans ce cas, la variance totale est de la forme :

$$[7] \Psi = \begin{bmatrix} B_1^{-1} & 0 \\ 0 & B_2^{-1} \end{bmatrix} \begin{bmatrix} \sigma_1^2 I_1 & \sigma_{12} S S' \\ \sigma_{12} S S' & \sigma_2^2 I_2 \end{bmatrix} \begin{bmatrix} (B_1')^{-1} & 0 \\ 0 & (B_2')^{-1} \end{bmatrix} \quad \text{où } S \text{ est le vecteur somme.}$$



## Une illustration

Le modèle étudié ici est destiné à illustrer certaines des techniques présentées dans cet article. Ce modèle relie le nombre de délits (vols et cambriolages) au revenu et aux valeurs immobilières. Une partie des résultats est disponible dans Anselin (1988a, 1990a) mais nous avons produit d'autres résultats à partir des logiciels SpaceStat 1.90 (Anselin, 1999) et Arc-View ©3.2<sup>(11)</sup>. Comme le souligne Anselin (1988a), il ne s'agit pas de fournir une explication aux schémas spatiaux des délits mais d'illustrer l'interaction entre autocorrélation et hétérogénéité spatiales.

L'échantillon est composé de 49 observations pour 1980 sur des *census tracts* à Columbus, une ville de l'Ohio. Ces zones sont représentées sur la figure 1. La variable DELIT est le nombre de cambriolages et de vols de voiture par millier de ménages dans la zone. La variable de revenu (REV) et celle de valeurs immobilières (IMMO) sont exprimées en milliers de dollars :

$$(3.1) \text{ DELIT} = a + b\text{REV} + c\text{IMMO} + \varepsilon$$

Dans cet exemple, la dépendance spatiale est prise en compte à travers une matrice de contiguïté d'ordre 1 standardisée.

### Les effets spatiaux dans le modèle simple

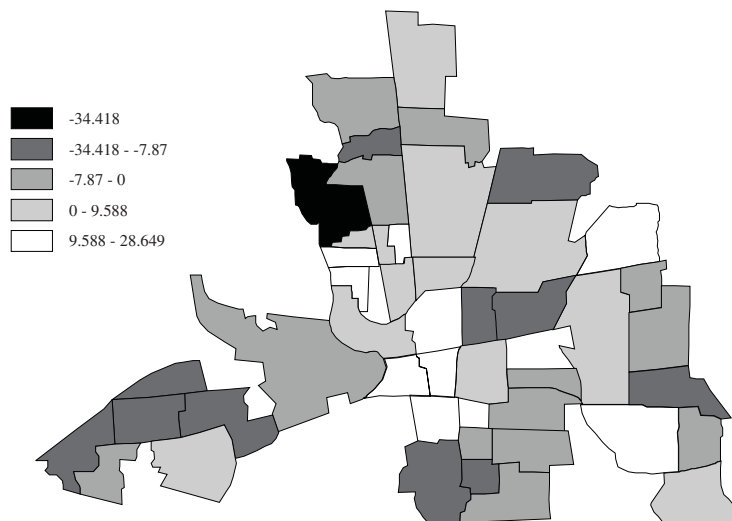
La première étape consiste à estimer le modèle simple reliant le crime au revenu et aux valeurs immobilières. Les résultats de l'estimation de (3.1) sont donnés dans la première colonne du tableau 1. Tous les coefficients sont significatifs, à l'exception de  $c$  qui n'est pas significatif lorsqu'on applique la correction de White. Le test de Jarque-Bera ne permet pas de rejeter l'hypothèse de normalité. Par conséquent, la validité de l'estimation par le maximum de vraisemblance et du calcul des tests du multiplicateur de Lagrange sont assurés. Les tests suivants concernent des diagnostics pour l'autocorrélation spatiale des erreurs et l'hétéroscédasticité. Le test joint *JLM* indique la présence jointe de ces deux effets et donc des sources multiples de mauvaise spécification du modèle sont donc présentes :

– concernant l'autocorrélation spatiale, si l'on applique la règle de décision de Anselin et Florax (1995)<sup>(12)</sup>, l'étude de la significativité des différents tests nous conduit à choisir un modèle autorégressif avec variable endogène décalée  $W\_DELIT$  :

$$(3.2) \text{ DELIT} = a + b\text{REV} + c\text{IMMO} + \rho W\_DELIT + \varepsilon$$

– concernant l'hétéroscédasticité, le test de White indique une hétéroscédasticité générale. Pour chercher à préciser la cause de cette hétéroscédasticité, nous avons effectué le test de Breusch-Pagan associé à plusieurs variables : les variables initiales REV et IMMO, la variable associée à l'aire de chaque voisinage, des variables dichotomiques Est-Ouest, Nord-Sud, centre-périphérie. De toutes ces variables, seul le test de Breusch-Pagan associé à la variable Est-Ouest (EO) permet de rejeter l'hypothèse nulle d'homoscédasticité. Ce résultat est confirmé par l'examen de la carte des résidus du modèle (3.1) où l'on observe que les résidus positifs se situent principalement au centre et à l'est de la ville.

Figure 1 : résidus du modèle [3.1]



Ces résultats indiquent l'existence d'une autocorrélation spatiale sous la forme d'une variable endogène décalée ainsi que l'existence d'une hétérogénéité spatiale sous la forme d'une hétéroscédasticité en groupes et/ou d'une instabilité structurelle associée à deux régimes spatiaux Est-Ouest. Par conséquent, toute l'inférence statistique basée sur l'estimateur des MCO n'est pas fiable. Ainsi, les tests d'hétéroscédasticité doivent être interprétés avec prudence compte tenu de la présence d'autocorrélation spatiale. Inversement, les tests d'autocorrélation spatiale ne sont pas fiables en présence d'hétéroscédasticité.

**Tableau 1 : estimation du modèle simple**

Modèle		Modèle simple	Modèle autorégressif
Estimation		<i>MCO-White</i>	<i>MV</i>
Constante		68,619 (0,000)	45,265 (0,000)
REV		-1,597 (0,001)	-1,036 (0,000)
IMMO		-0,274 (0,123)	-0,259 (0,003)
W_DELIT		-	0,423 (0,000)
LIK		-187,377	-182,518
AIC		380,754	373,035
BIC		386,430	380,603
$\hat{\sigma}^2$		130,759	95,723
<b>Test de normalité</b>	<i>JB</i> ( $\chi^2_2$ )	1,836 (0,399)	-
<b>Test joint d'hétéroscédasticité et d'autocorrélation spatiale</b>	<i>JLM</i> ( $\chi^2_3$ )	13,715 (0,003)	-
<b>Tests d'autocorrélation spatiale</b>	Moran	2,938 (0,003)	-
	<i>LM<sub>ERR</sub></i> ( $\chi^2_1$ )	5,815 (0,016)	-
	<i>RLM<sub>ERR</sub></i> ( $\chi^2_1$ )	0,127 (0,721)	-
	<i>LM<sub>LAG</sub></i> ( $\chi^2_1$ )	8,760 (0,003)	-
	<i>RLM<sub>LAG</sub></i> ( $\chi^2_1$ )	3,072 (0,002)	-
	<i>SARMA</i> ( $\chi^2_2$ )	8,887 (0,012)	-
	<i>LM<sub>ERR</sub>*</i> ( $\chi^2_1$ )	-	0,505 (0,477)
<b>Tests d'hétéroscédasticité</b>	<i>White</i> ( $\chi^2_5$ )	19,947 (0,001)	-
	<i>BP</i> (EO) ( $\chi^2_1$ )	7,055 (0,008)	5,796 (0,016)
	<i>BPS</i> (EO) ( $\chi^2_1$ )	-	5,797 (0,016)

Notes : les probabilités critiques sont entre parenthèses. MCO-White indique l'utilisation de la correction de White (1980). LIK est la valeur de la fonction de log-vraisemblance à l'optimum. AIC est le critère d'information d'Akaike. BIC est le critère d'information de Schwartz. JB est le test de Jarque-Bera de normalité des résidus. JLM est le test joint pour l'autocorrélation spatiale des erreurs et l'hétéroscédasticité liée aux variables initiales IMMO et REV (Anselin, 1988a, 1988b). Moran est le test I de Moran (Cliff et Ord, 1981). *LM<sub>ERR</sub>* et *LM<sub>LAG</sub>* sont les tests du multiplicateur de Lagrange respectivement pour l'autocorrélation spatiale des erreurs et la variable endogène décalée. *LM<sub>ERR</sub>\** et *LM<sub>LAG</sub>\** sont leurs équivalents robustes (Anselin *et alii*, 1996). SARMA est le test du multiplicateur de Lagrange pour la présence des deux effets spatiaux (Anselin, 1988a, 1988b). *LM<sub>ERR</sub>\** est le test du multiplicateur de Lagrange d'une autocorrélation des erreurs dans un modèle spatial autorégressif (Anselin, 1988a). White est le test d'hétéroscédasticité de White (1980). BP (EO) est le test de Breusch-Pagan lié à la variable de régime  $D_1$ . BPS est sa version ajustée (Anselin, 1988a).

Avant de considérer le problème de l'hétéroscédasticité, nous traitons tout d'abord le problème de l'autocorrélation spatiale. Nous avons estimé (3.2) par la méthode du maximum de vraisemblance. Les résultats se trouvent dans la deuxième colonne du premier tableau. Tous les coefficients sont hautement significatifs, y compris la variable endogène décalée, ce qui indique une forte concentration spatiale du nombre de crimes. Les coefficients structurels sont similaires à ceux estimés par les MCO mais plus petits en valeur absolue. Le modèle est meilleur que le précédent en termes des critères d'information (critère AIC d'Akaike et BIC de Schwartz) et le test  $LM_{ERR}^*$  ne permet pas de rejeter l'absence d'une autocorrélation spatiale des erreurs supplémentaires. En revanche, le test de Breusch-Pagan (BP) et sa version ajustée de l'autocorrélation spatiale (BPS) indiquent qu'il subsiste une certaine forme d'hétéroscédasticité liée à la localisation des zones dans l'est ou dans l'ouest de Columbus.

### Régimes spatiaux et hétéroscédasticité en groupes

Les résultats précédents indiquent la présence d'une hétérogénéité spatiale discrète, sous la forme de deux régimes spatiaux est-ouest. Cet aspect peut être considéré sous la forme d'une hétéroscédasticité en groupes et/ou d'une instabilité structurelle entre les deux régimes.

Formellement, soit les deux variables muettes suivantes :

$$(3.3a) \quad D_{1i} = \begin{cases} 1 & \text{si la zone } i \text{ est à l'est} \\ 0 & \text{si la zone } i \text{ est à l'ouest} \end{cases}$$

$$\text{et } (3.3b) \quad D_{2i} = \begin{cases} 1 & \text{si la zone } i \text{ est à l'ouest} \\ 0 & \text{si la zone } i \text{ est à l'est} \end{cases}$$

On estime le modèle suivant incorporant à la fois des régimes spatiaux et une hétéroscédasticité en groupes :

$$(3.4a) \quad DELIT = a_1 D_1 + a_2 D_2 + b_1 D_1 REV + b_2 D_2 REV + c_1 D_1 IMMO + c_2 D_2 IMMO + \varepsilon$$

avec

$$(3.4b) \quad \varepsilon \sim \begin{bmatrix} \sigma_1^2 I_1 & 0 \\ 0 & \sigma_2^2 I_2 \end{bmatrix}$$

Les résultats sont fournis dans les deux premières colonnes du tableau 2. Tous les coefficients sont significatifs à l'exception de celui associé à IMMO à l'est. Le test du ratio de vraisemblance portant sur l'égalité des variances indique que l'hypothèse de variances différentes par régimes ne peut pas être rejetée. En revanche, le test de Chow asymptotique (test de Wald) portant sur l'instabilité structurelle ne permet pas de rejeter l'hypothèse de coefficients identiques pour les deux régimes. Concernant les tests individuels, seul le coefficient associé à IMMO peut être considéré comme différent selon le régime. Pourtant, l'inférence statistique faite sur ce modèle n'est pas fiable puisque les derniers tests fournis dans le tableau indiquent l'omission à tort de l'autocorrélation spatiale sous la forme d'une variable endogène décalée.

Le dernier modèle estimé est alors le modèle (3.5) incluant une variable autorégressive :

$$(3.5a) \quad DELIT = \rho W\_DELIT + a_1 D_1 + a_2 D_2 + b_1 D_1 REV + b_2 D_2 REV + c_1 D_1 IMMO + c_2 D_2 IMMO + \varepsilon$$

avec

$$(3.5b) \quad \varepsilon \sim \begin{bmatrix} \sigma_1^2 I_1 & 0 \\ 0 & \sigma_2^2 I_2 \end{bmatrix}$$

La prise en compte de l'autocorrélation spatiale modifie substantiellement certains des résultats précédents. En particulier, le coefficient associé à REV dans l'ouest n'est plus significatif. Certains tests d'instabilité structurelle basculent également : le test global devient significatif ainsi que le test individuel sur REV (à 10%). On obtient donc un schéma spatial distinct dans les deux régimes : à l'est, c'est le revenu qui influence le plus le nombre de crimes commis alors qu'à l'ouest, ce sont les valeurs immobilières qui jouent le rôle le plus important. Les variances sont toujours différentes d'un régime à l'autre et le coefficient spatial autorégressif est hautement significatif. Enfin, ce modèle est le meilleur du point de vue des critères d'information.

**Tableau 2 : régimes spatiaux et hétéroscédasticité en groupes**

Modèle		Régimes spatiaux et hétéroscédasticité en groupes		Régimes spatiaux, hétéroscédasticité en groupes et autocorrélation spatiale	
Estimation		MV		MV	
		Ouest	Est	Ouest	Est
constante		76,650 (0,000)	67,294 (0,000)	50,058 (0,000)	45,337 (0,000)
REV		-1,455 (0,016)	-2,014 (0,000)	-0,652 (0,203)	-1,650 (0,000)
IMMO		-0,545 (0,004)	-0,064 (0,499)	-0,546 (0,000)	-0,011 (0,893)
W_CRIME			-	0,431 (0,000)	
$\hat{\sigma}^2$		182,805 (0,001)	57,932 (0,000)	129,07 (0,001)	42,609 (0,000)
LIK			-180,472		-173,826
AIC			372,943		351,652
SC			384,294		374,895
<b>Test d'égalité des variances</b>	LR ( $\chi_1^2$ )		7,944 (0,005)		7,137 (0,007)
<b>Test d'instabilité structurelle</b>	Chow (Wald) ( $\chi_3^2$ )		5,35 (0,148)		10,053 (0,018)
<b>Tests de stabilité des coefficients individuels</b>	Constante ( $\chi_1^2$ )		0,776 (0,378)		0,278 (0,598)
	REV ( $\chi_1^2$ )		0,672 (0,412)		3,003 (0,083)
	IMMO ( $\chi_1^2$ )		5,162 (0,023)		8,942 (0,003)
<b>Tests d'autocorrélation spatiale</b>	$LM_{ERR}(\chi_1^2)$		5,972 (0,014)		-
	$LM_{LAG}(\chi_1^2)$		11,267 (0,000)		-

Notes : les probabilités critiques sont entre parenthèses. LIK est la valeur de la fonction de log-vraisemblance à l'optimum. AIC est le critère d'information d'Akaike. BIC est le critère d'information de Schwartz. LR est le test du ratio de vraisemblance destiné à tester l'égalité des variances dans les régimes. Le test global d'instabilité structurelle et les tests de stabilité des coefficients sont basés sur une statistique de Wald asymptotique.  $LM_{ERR}$  et  $LM_{LAG}$  sont les tests du multiplicateur de Lagrange respectivement pour l'autocorrélation spatiale des erreurs et la variable endogène décalée (Anselin, 1988a).

---

## Conclusion

---

L'objectif de cet article était d'examiner les méthodes permettant de prendre en compte l'hétérogénéité spatiale sous la forme d'une instabilité des paramètres et/ou d'une hétéroscédasticité et d'examiner dans quelle mesure les tests traditionnels d'hétéroscédasticité ou d'instabilité structurelle doivent être amendés pour tenir compte d'une autocorrélation spatiale.

La présence simultanée d'autocorrélation spatiale et d'hétérogénéité spatiale dans un modèle est très fréquente compte tenu des nombreux liens qui unissent ces deux effets. Ce cas de figure nécessite des méthodes d'estimation et d'inférence adaptées car les tests traditionnels d'hétéroscédasticité et d'instabilité structurelle ne peuvent plus être appliqués. D'une part, si l'on utilise une forme particulière d'hétéroscédasticité, il est possible de tester conjointement ou séparément les deux effets spatiaux. D'autre part, le test de Breusch-Pagan ainsi que le test de Chow traditionnel doivent tous deux être amendés pour tenir compte d'une éventuelle autocorrélation spatiale des erreurs.

Cependant, au contraire du cas où seule l'autocorrélation spatiale est envisagée, il n'existe pas à l'heure actuelle de règles de décision permettant de choisir entre les différents modèles combinant les deux effets spatiaux. En outre, les propriétés de plusieurs de ces tests en échantillon fini (biais et puissance) ne sont pas encore bien connues.

---

## Notes

---

- (1) On entend par zone géographique un point ou encore une aire géographique, comme des régions ou des pays.
- (2) La généralisation à  $G$  sous-groupes est immédiate.
- (3) Tout autre système de coordonnées est également valable. Il est à noter que l'utilisation de Systèmes d'Information Géographique permet maintenant d'obtenir des coordonnées d'une grande précision.
- (4) Dans le cadre des modèles hédoniques, Des Rosiers et Thériault (1999) ont développé une stratégie d'estimation et de tests combinant autocorrélation spatiale et variables polynomiales tout en évitant les problèmes de multicollinéarité.
- (5) Comme précédemment, on peut également supposer la constante variable pour chaque observation.
- (6) Casetti et Jones (1988) proposent dans ce cas de remplacer les variables augmentées initiales par leurs composantes principales (méthode VE orthogonale).
- (7) Voir Casetti et Jones (1992) pour les différentes applications de cette méthode dans un cadre spatial ou non.
- (8) Des tests d'autocorrélation spatiale en présence d'hétéroscédasticité de forme connue ou inconnue ont été proposés par Anselin (1990a, 1990b) mais sont rarement utilisés.
- (9) Ce test est la somme de la statistique de Breusch-Pagan et de la statistique  $LM_{ERR}$ , la statistique du multiplicateur de Lagrange associée au test de l'autocorrélation spatiale des erreurs.
- (10) Kelejian et Robinson (1998) proposent une autre formulation de ce test qui n'est pas basé sur le principe du maximum de vraisemblance et qui n'exige pas *a priori* un modèle linéaire, des erreurs normales et une forme d'hétéroscédasticité connue. Des simulations de Monte-Carlo effectuées par les deux auteurs sur les performances de ces différents tests suggèrent que le test  $JLM$  et le test de Kelejian et Robinson sont relativement puissants qu'il y ait ou non de l'hétéroscédasticité, pour détecter au moins un des deux problèmes : autocorrélation spatiale ou hétéroscédasticité. Si l'hypothèse nulle est rejetée, un test pour chaque cas doit être effectué afin de déterminer l'origine du rejet.
- (11) Cette base de données a été très souvent utilisée pour illustrer les différentes propriétés des modèles d'économétrie spatiale (Anselin, 1988a, 1990a ; Getis, 1995 ; LeSage, 1999 ; Baltagi et Li, 2001). Par rapport aux études citées, l'illustration se distingue par une étude complète de l'hétérogénéité spatiale. Notons cependant que la taille réduite de l'échantillon et les possibles problèmes d'endogénéité de la variable explicative représentant les valeurs immobilières pourraient conduire à remettre en cause les résultats. Cependant, cet exemple reste purement illustratif des techniques présentées dans l'article et n'a pas pour objectif une étude des déterminants économiques du taux de criminalité.
- (12) Une présentation détaillée de ces règles de décision se trouve dans Le Gallo (2002a, 2002b).



---

## Bibliographie

---

- Agterberg F. (1984).** "Trend Surface Analysis", in Gaile G.L. et Wilmot C.J. (eds.), *Spatial Statistics and Models*, Reidel, Boston.
- Anderson J.E. (1982).** "Cubic Spline Urban Density Functions", *Journal of Urban Economics*, vol. 12, pp. 55-67.
- Anderson J.E. (1985).** "The Changing Structure of a City: Temporal Changes in Cubic Spline Urban Density Patterns", *Journal of Regional Science*, vol. 25, pp. 413-425.
- Anselin L. (1988a).** *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers, Dordrecht.
- Anselin L. (1988b).** "Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity", *Geographical Analysis*, vol. 20, pp. 1-17.
- Anselin L. (1990a).** "Spatial Dependence and Spatial Structural Instability in Applied Regression Analysis", *Journal of Regional Science*, vol. 30, pp. 185-207.
- Anselin L. (1990b).** "Some Robust Approach to Testing and Estimating in Spatial Econometrics", *Regional Science and Urban Economics*, vol. 20, pp. 141-163.
- Anselin L. (1998).** "Exploratory Spatial Data Analysis in a Geocomputational Environment", in Longley P.A., Brooks S.M., McDonnell R. et Mcmillan B. (eds.), *Geocomputation, a Primer*, Wiley, New York.
- Anselin L. (1999).** "SpaceStat, a Software Package for the Analysis of Spatial Data, Version 1.90", Ann Arbor, BioMedware.
- Anselin L. (2000).** "Spatial Econometrics", in Baltagi B. (ed.), *Companion to Econometrics*, Basil Blackwell, Oxford.
- Anselin L., Bera A.K., Florax R. et Yoon M.J. (1996).** "Simple Diagnostic Tests for Spatial Dependence", *Regional Science and Urban Economics*, vol. 26, pp. 77-104.
- Anselin L. et Bera A. (1998).** "Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics", in Ullah A. et Giles D.E.A. (eds.), *Handbook of Applied Economics Statistics*, Springer-Verlag, Berlin.
- Anselin L. et Florax (1995).** "Small Sample Properties of Tests for Spatial Dependence in Regression Model", in Anselin L. et Florax R.J.G.M. (eds.), *New Directions in Spatial Econometrics*, Springer-Verlag, Berlin.
- Anselin L. et Griffith D.A. (1988).** "Do Spatial Effects Really Matter in Regression Analysis?", *Papers of the Regional Science Association*, vol. 65, pp. 11-34.
- Baltagi B.H. et Li D. (2001).** "LM Tests for Functional Form and Spatial Error Correlation", *International Regional Science Review*, vol. 24, pp. 194-225.
- Breusch T. et Pagan A. (1979).** "A Simple Test for Heteroskedasticity and Random Coefficient Variation", *Econometrica*, vol. 47, pp. 1287-1294.
- Brueckner J.K. (1981).** "A Vintage Model of Urban Growth", *Journal of Urban Economics*, vol. 8, pp. 389-402.
- Brueckner J.K. (1986).** "A Switching Regression Analysis of Urban Population Densities", *Journal of Urban Economics*, vol. 19, pp. 174-189.
- Brundson C., Fotheringham A.S. et Charlton M. (1996).** "Geographically Weighted Regression: a Method for Exploring Spatial Nonstationarity", *Geographical Analysis*, vol. 28, pp. 281-298.
- Brundson C., Fotheringham A.S. et Charlton M. (1998).** "Spatial Nonstationarity and Autoregressive Models", *Environment and Planning A*, vol. 30, pp. 957-973.
- Brundson C., Fotheringham A.S. et Charlton M. (1999).** "Some Notes on Parametric Significance Tests for Geographically Weighted Regression", *Journal of Regional Science*, vol. 39, pp. 497-524.
- Can A. (1990).** "The Measurement of Neighborhood Dynamics in Urban House Prices", *Economic Geography*, vol. 66, pp. 254-272.
- Can A. (1992).** "Specification and Estimation of Hedonic Housing Price Models", *Regional Science and Urban Economics*, vol. 22, pp. 453-474.
- Casetti E. (1972).** "Generating Models by the Expansion Method: Applications to Geographical Research", *Geographical Analysis*, vol. 4, pp. 81-91.
- Casetti E. (1997).** "The Expansion Method, Mathematical Modeling, and Spatial Econometrics", *International Regional Science Review*, vol. 20, pp. 9-33.

- Casetti E. et Can A. (1999).** “The Econometric Estimation and Testing of DARP Models”, *Journal of Geographical Systems*, vol. 1, pp. 91-106.
- Casetti E. et Jones J.P. III (1988).** “Spatial Parameter Variation by Orthogonal Trend Surface Expansions: an Application to the Analysis of Welfare Program Participation Rates”, *Social Science Research*, vol. 16, pp. 285-300.
- Casetti E. et Jones J.P. III (1992).** *Applications of the Expansion Method*, Routledge, Londres.
- Cleveland W.S. (1979).** “Robust Locally Weighted Regression and Smoothing Scatterplots”, *Journal of the American Statistical Association*, vol. 74, pp. 829-836.
- Cleveland W.S. et Devlin S.J. (1988).** “Locally Weighted Regression: an Approach to Regression Analysis by Local Fitting”, *Journal of the American Statistical Association*, vol. 83, pp. 596-610.
- Cleveland W.S., Devlin S.J. et Grosse E. (1988).** “Regression by Local Fitting, Methods, Properties, and Computational Algorithms”, *Journal of Econometrics*, vol. 37, pp. 87-114.
- Cliff A.D. et Ord J.K. (1981).** *Spatial Processes: Models and Applications*, Pion, Londres.
- Des Rosiers F. et Thériault M. (1999).** “House Prices and Spatial Dependence: Towards an Integrated Procedure to Model Neighborhood Dynamics”, Document de travail, n°1999-002, Faculté des Sciences de l'Administration, Université Laval, Québec.
- Fotheringham A.S. et Brundson C. (1999).** “Local Forms of Spatial Analysis”, *Geographical Analysis*, vol. 31, pp. 340-358.
- Fotheringham A.S., Brundson C. et Charlton M. (2000).** *Quantitative Geography, Perspectives on Spatial Data Analysis*, Sage Publications, Londres.
- Fotheringham A.S., Charlton M. et Brundson C. (1996a).** “The Geography of Parameter Space: an Investigation of Spatial Non-stationarity”, *International Journal of Geographical Information Systems*, vol. 10, pp. 605-627.
- Fotheringham A.S., Charlton M. et Brundson C. (1996b).** “Two Techniques for Exploring Non-stationarity in Geographical Data”, *Geographical Systems*, vol. 28, pp. 281-298.
- Fotheringham A.S., Charlton M. et Brundson C. (1998).** “Geographically Weighted Regression: a Natural Evolution of the Expansion Method for Spatial Data Analysis”, *Environment and Planning A*, vol. 30, pp. 1905-1927.
- Getis A. (1995).** “Spatial Filtering in a Regression Framework: Examples using Data on Urban Crime, Regional Inequality, and Government Expenditures”, in Anselin L. et Florax R.J.G.M. (eds.), *New Directions in Spatial Econometrics*, Springer-Verlag, Berlin.
- Greene W.H. (2000).** *Econometric Analysis*, Prentice-Hall, Londres.
- Griffith D.A. (1978).** “A Spatially Adjusted ANOVA Model”, *Geographical Analysis*, vol. 10, pp. 296-301.
- Griffith D.A. (1992).** “A Spatially Adjusted N-way ANOVA Model”, *Regional Science and Urban Economics*, vol. 22, pp. 347-369.
- Haining R. (1990).** *Spatial Data Analysis in the Social and Environmental Sciences*, Cambridge University Press, Cambridge.
- Johnson M.S. et Ragas W.R. (1987).** “CBD Land Values and Multiple Externalities”, *Land Economics*, vol. 63, pp. 335-347.
- Jones J.P. III (1983).** “Parameter Variation Via the Expansion Method with Tests for Autocorrelation”, *Modeling and Simulation*, vol. 14, pp. 853-857.
- Kelejian H.H. et Robinson D.P. (1998).** “A Suggested Test for Spatial Autocorrelation and/or Heteroskedasticity and Corresponding Monte Carlo Results”, *Regional Science and Urban Economics*, vol. 28, pp. 389-417.
- Le Gallo J. (2002a).** “Économétrie spatiale : l'autocorrélation spatiale dans les modèles de régression linéaire”, *Économie et Prévision*, vol. 155, pp. 139-158.
- Le Gallo J. (2002b).** “Disparités géographiques et convergence des régions européennes : une approche par l'économétrie spatiale”, *Thèse de Doctorat en Sciences Économiques*, Université de Bourgogne, Dijon.
- Le Gallo J. et Ertur C. (2003).** “Exploratory Spatial Data analysis of the Distribution of Regional per Capita GDP in Europe, 1980-1995”, *Papers in Regional Science*, vol. 82, pp. 175-201.
- Legendre P., Oden N.L., Sokal R.R., Vaudor A. et Kim J. (1990).** “Approximate Analysis of Variance of Spatially Autocorrelated Regional Data”, *Journal of Classification*, vol. 7, pp. 53-75.
- LeSage J.P. (1999).** *Spatial Econometrics*, WebBook of Regional Science.
- Leung Y., Mei C. et Zhang W. (2000).** “Statistical Tests for Spatial Non-stationarity Based on the Geographically Weighted Regression Model”, *Environment and Planning A*, vol. 32, pp. 9-32.
- Marsch L.C., Cormier D.R. (2001).** *Spline Regression Models*, Sage Publications, Thousand Oaks.

- McMillen D.P. (1992).** “Probit with Spatial Autocorrelation”, *Journal of Regional Science*, vol. 32, pp. 335-348.
- McMillen D.P. (1994).** “Vintage Growth and Population Density: an Empirical Investigation”, *Journal of Urban Economics*, vol. 36, pp. 333-352.
- McMillen D.P. (1996).** “One Hundred Fifty Years of Land Values in Chicago: a Nonparametric Approach”, *Journal of Urban Economics*, vol. 40, pp. 100-124.
- McMillen D.P. et McDonald J.F. (1997).** “A Nonparametric Analysis of Employment Density in a Polycentric City”, *Journal of Regional Science*, vol. 37, pp. 591-612.
- McMillen D.P. et McDonald J.F. (1998).** “Population Density in Chicago: a Bid Rent Approach”, *Urban Studies*, vol. 7, pp. 1119-1130.
- Olmo J.C. (1995).** “Spatial Estimation of Housing Prices and Locational Rents”, *Urban Studies*, vol. 32, pp. 1331-1344.
- Pace R. K., Barry R. et Sirmans C.F. (1998).** “Spatial Statistics and Real Estate”, *Journal of Real Estate Finance and Economics*, vol. 17, pp. 5-13.
- Pavlov A.D. (2000).** “Space-varying Regression Coefficients: a Semi-parametric Approach applied to real Estate Markets”, *Real Estate Economics*, vol. 2, pp. 249-283.
- Pindyck R.S., Rubinfeld D.L. (1998).** *Econometric Models and Economic Forecasts*, McGraw-Hill, New York.
- Quandt R. (1958).** “The Estimation of the Parameters of a Linear Regression System obeying two Separate Regimes”, *Journal of the American Statistical Association*, vol. 53, pp. 873-880.
- Ripley B. (1981).** *Spatial Statistics*, New York, Wiley.
- Schroeder L.D. et Sjoquist D.L. (1976).** “Investigation of Population Density Gradients using Trend Surface Analysis”, *Land Economics*, vol. 52, pp. 382-392.
- Sokal R.R., Oden N.L., Thomson B.A. et Kim J. (1993).** “Testing for Regional Differences in Means: Distinguishing Inherent from Spurious Spatial Autocorrelation by Restricted Randomization”, *Geographical Analysis*, vol. 25, pp. 199-209.
- Suits D.B., Mason A. et Chan L. (1978).** “Spline Functions Fitted by Standard Regression Methods”, *Review of Economics and Statistics*, vol. 60, pp. 132-139.
- White H. (1980).** “A Heteroskedastic-consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity”, *Econometrica*, vol. 48, pp. 817-838.