ANALYSE ET MODELISATION SPATIALES AVEC SAS

Aristide Elysée HOUNDETOUNGAN

Doctorant en Economique

à l'Université Laval

3 mai 2017

UNIVERSITÉ LAVAL

Sommaire

Plan de la présentation

Le modèle linéaire standard

Spécification du modèle linéaire standard

On dispose d'une variable Y qu'on veut régresser sur K variables contenues dans une matrice X

$$Y = X\beta + \varepsilon$$

Y est de dimension $(N, 1)$, X de dimension (N, K), β un vecteur colonne $(K, 1)$ des coefficients des variables explicatives.

On fait parfois l'hypothèse que ε suit $\mathcal{N}(0, \sigma^2 I_N) \Longrightarrow$ Les erreurs sont stationnaires : homoscédastiques, non autocorrélées.

Problème général : Cette hypothèse n'est généralement pas vérifiée.

Le modèle linéaire standard

Quelques solutions apportées aux problèmes

En présence d'autocorrélation des erreurs :

Méthode itérative de Cochrane Orcutt

Moindres Carrés Quasi Généralisés : Prais-Watson.

En présence d'hétéroscédasticité des erreurs :

Moindres Carrés Généralisés

Moindres Carrés Quasi Généralisés

Estimation de plusieurs modèles sur des différents groupes d'individus homogènes.

Les limites du modèle linéaire

Lorsqu'on travaille sur des données localisées (géographiques ou spatiales), il peut avoir une liaison entre l'information d'une coordonnée géographique et les autres.

Dans ces conditions, le modèle linéaire standard n'est plus le modèle adapté aux données. Les estimateurs MCO peuvent être inefficaces et/ou non convergents.

D'où l'utilisation d'autres méthodes plus complexes utilisant l'information géographique sur les différentes observations : **Les régressions spatiales**.

Plan de la présentation

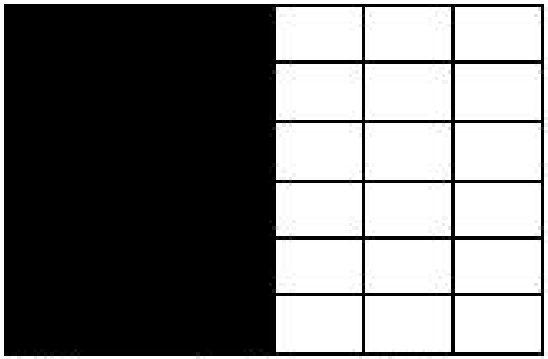
Autocorrélation spatiale

L'autocorrélation spatiale positive se traduit par une tendance à la concentration dans l'espace de valeurs faibles ou élevées d'une variable aléatoire.

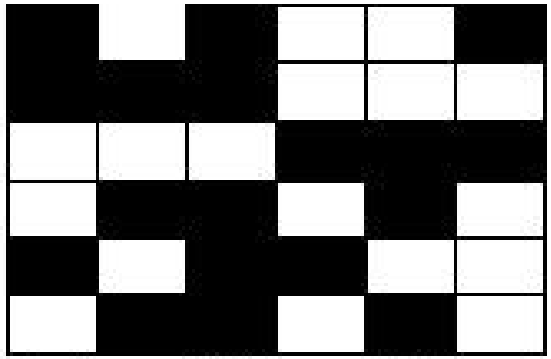
En revanche, l'autocorrélation spatiale négative signifie que chaque localisation tend à être entourée par des localisations voisines pour lesquelles la variable aléatoire prend des valeurs très différentes.

Autocorrélation spatiale

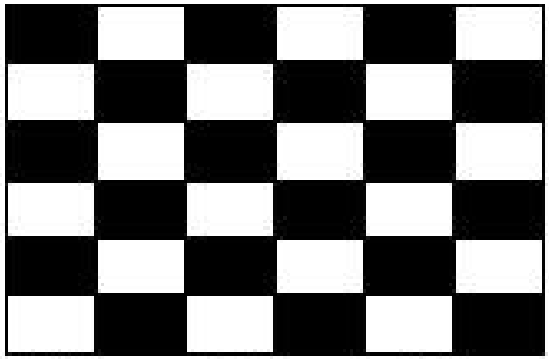
Autrement dit, il y a une relation fonctionnelle entre ce qui se passe en un point de l'espace et ce qui se passe ailleurs.



Autocorrélation spatiale positive



Autocorrélation spatiale nulle



Autocorrélation spatiale négative

Les matrices de poids

La dépendance entre régions est étudiée à travers leur proximité. Pour ce faire, on définit la topologie spatiale par une matrice carrée de poids W de dimension N égale au nombre d'observations (ou de régions). L'élément w_{ij} de la matrice matérialise la façon dont les régions sont connectées.

Plusieurs spécifications peuvent être admises.

Les matrices de poids

Matrices de contiguïté

La contiguïté entre deux régions se définit par le fait qu'elles aient une frontière commune et chaque terme de la matrice de contiguïté est égal à 1 si les régions sont contiguës à l'ordre 1 et 0 sinon (par convention, une région n'est pas contiguë avec elle-même : $w_{ii} = 0, \forall i$).

Cette notion de contiguïté peut être généralisée : deux régions i et j sont contiguës à l'ordre k si k est le nombre minimal de frontières à traverser pour aller de i à j. Ces matrices de contiguïté sont souvent utilisées en raison de leur simplicité.

Les matrices de poids

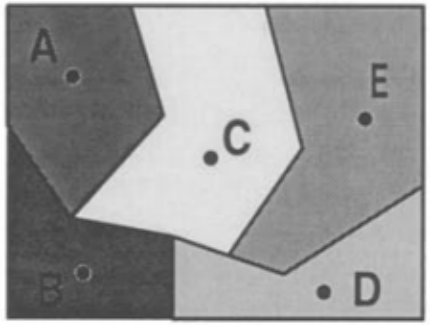
Matrices de distances

L'interaction entre deux régions i et j dépend de la distance entre les centroïdes de ces régions. La distance peut être : distance à vol d'oiseau, distance par routes ou généralisation aux temps de transport ou à des indices d'accessibilité.

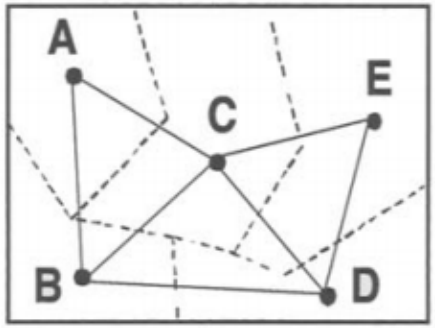
Ici également, on suppose que la distance entre une zone et elle-même est nulle.

Les matrices de poids

ZONE



SPECIFICATION



Matrice de contiguïté

	A	B	C	D	E
A	_	1	1	0	0
B	1	_	1	1	0
C	1	1	_	1	1
D	0	1	1	_	1
E	0	0	1	1	_

Matrice de distance

	A	B	C	D	E
A	0	d_{AB}	d_{AC}	d_{AD}	d_{AE}
B	d_{BA}	0	d_{BC}	d_{BD}	d_{BE}
C	d_{CA}	d_{CB}	0	d_{CD}	d_{CE}
D	d_{DA}	d_{DB}	d_{DC}	0	d_{DE}
E	d_{EA}	d_{EB}	d_{EC}	d_{ED}	0

Les matrices de poids

Les matrices de poids sont souvent standardisées : chaque élément w_{ij} de la matrice est divisé par la somme des éléments de le ligne i : $\sum_j w_{ij}$.

Les poids sont alors compris entre 0 et 1 et cette opération rend les paramètres spatiaux comparables entre les modèles économétriques.

La variable spatiale décalée

Soit N observations qui prennent des valeurs pour une variable quantitative X. Pour une matrice de poids W, on appelle variable spatiale décalée (« *spatial lag* ») de X, la variable

$$W_x = WX$$

Implicitement, la valeur prise par la variable spatiale décalée pour un individu, est une moyenne pondérée des valeurs prises par les autres individus lorsque la matrice de poids est standardisée.

Plan de la présentation

Indice de Moran

La statistique de Moran (1948) est celle la plus utilisée pour mesurer l'autocorrélation spatiale. Pour une variable X, et une matrice de poids W, la statistique de Moran est définie par :

$$I = \frac{\sum_i \sum_j w_{ij}(x_j - \bar{x})(x_i - \bar{x})}{S_0} \Big/ \frac{\sum_i (x_i - \bar{x})^2}{N}$$

Avec $\bar{x} = \frac{1}{N} \sum_i x_i$ et $S_0 = \sum_{i\,j} w_{ij}$

Indice de Moran

Lorsque la variable X est centrée, on peut écrire

$$I = \frac{X^t W X}{S_0} \Big/ \frac{X^t X}{N}$$

Si la matrice W est standardisée, $S_0 = N$. Donc

$$I = \frac{X^t W X}{X^t X}$$

Test d'autocorrélation spatiale

La statistique I est comprise entre -1 et 1. Lorsqu'elle est proche de 1, on peut soupçonner la présence d'une autocorrélation spatiale positive. Inversement une forte valeur négative présume une autocorrélation spatiale négative. Dans le cas d'une absence de dépendance spatiale, $I = 0$.

Le test d'autocorrélation spatiale est généralement basé sur la statistique I tel que :

$$\begin{cases} H_0 : \ I = 0 \\ H_1 : \ I \neq 0 \end{cases}$$

L'espérance et la variance de I seront données plus loin.

Modèles d'autocorrélation spatiale

Considérons le modèle linéaire standard.

$$Y = X\beta + \varepsilon \tag{1}$$

Sauf indication contraire, supposons également les hypothèses habituelles de ce modèle.

La prise en compte de l'autocorrélation dans un modèle peut être effectuée de trois manières différentes.

Modèles d'autocorrélation spatiale

1-Variable endogène décalée (SAR)

On peut prendre en compte l'autocorrélation spatiale au niveau de la variable endogène. Dans ce cas, la variable endogène décalée est utilisée comme un facteur explicatif dans le modèle (Spatial Autoregresive model SAR).

$$Y = \rho W_Y + X\beta + \varepsilon \tag{2}$$

$W_Y = WY$: variable endogène spatiale décalée. ρ est le coefficient de cette variable dans le modèle.

Modèles d'autocorrélation spatiale

2-Variable exogène décalée (SLX)

On peut également prendre en compte l'autocorrélation spatiale au niveau de la variable exogène. Dans ce cas, on introduit une variable exogène décalée comme un facteur explicatif dans le modèle.

$$Y = X\beta + W_z\delta + \varepsilon \tag{3}$$

$W_z = WZ$: variable exogène spatiale décalée associée à un bloc de variables Z choisies parmi X ou non.

3-Autocorrélation spatiale des erreurs (SEM)

Enfin, on peut également préciser l'autocorrélation spatiale au niveau des erreurs (Spatial Error Model SEM).

$$\begin{cases} Y = X\beta + \varepsilon \\ \varepsilon = \lambda W\varepsilon + \mu \end{cases} \tag{4}$$

avec $\mu \rightsquigarrow iid(0, \sigma^2 I_N)$

Contrairement aux modèle SAR et SEM, le modèle SLX peut être estimé par la méthode des MCO. Il ne pose pas de problème d'endogénéité ni d'autocorrélation. Les estimateurs MCO sont BLUE.

4-Autres modèles

D'autres modèles sont également utiles tels que :

- General Spatial Model (GSM) qui est une combinaison de SAR et SEM ; Spatial Autoregressive Moving Average (SARMA) très proches des GSM ;
- Spatial Durbin Model (SDM), qui n'est rien d'autre qu'une extension du modèle SEM avec prise en compte d'une liaison entre les erreurs et les variables explicatives ;
- Spatial Durbin and Error Model (SDEM), extension du modèle SEM avec prise en compte d'une liaison entre les erreurs et les explicatives et la variable endogène.

Estimation de modèles spatiaux

Modèle général

Puisque le modèle SLX ne pose pas de problème d'estimation, le modèle général le plus souvent étudié est celui qui combine les modèles SAR et SEM : General Spatial Model (SGM).

$$\begin{cases} Y = \rho W_1 Y + X\beta + \varepsilon \\ \varepsilon = \lambda W_2 \varepsilon + \mu \end{cases} \tag{5}$$

avec $\mu \rightsquigarrow iid(0, \sigma^2 I_N)$

W_1 et W_2 sont les matrices de poids associées à la variable endogène et aux erreurs respectivement .

Estimation de modèles spatiaux

Ce modèle est un modèle d'autocorrélation d'ordre 1. Il existe également d'autres spécifications de type autocorrélation d'ordre supérieur (Voir Brandsma et Kelletaper (1979), Huang (1984), Jayet (1993) ou Anselin et Florax (1995b)).

Le modèle 5 (GSM) est équivalent à

$$Y = \rho W_1 Y + \lambda W_2 Y - \rho\lambda W_2 W_1 Y + X\beta - \lambda W_2 X\beta + \mu$$

Estimation de modèles spatiaux

Conséquences sur les estimateurs MCO

Le problème d'autocorrélation des erreurs et d'endogénéité a d'importantes conséquences sur les estimateurs MCO. Les estimateurs ne sont pas convergents (modèle AR) et non efficaces (autocorrélation des erreurs)

Estimation par maximum de vraisemblance

En supposant que $\mu \rightsquigarrow \mathcal{N}(0, \sigma^2 I_N)$, on peut écrire la log-vraisemblance du modèle.

$$Ln(L) = -\frac{N}{2} ln(2\pi) - \frac{N}{2} ln(\sigma^2) + ln|I_N - \rho W_1| + ln|I_N - \lambda W_2| - \frac{1}{2\sigma^2} u^t u$$

Estimation de modèles spatiaux

Estimation par maximum de vraisemblance

On trouvera dans Anselin (1988) que le programme n'admet pas toujours de solution (quand $|I_N - \rho W_1|$ et $|I_N - \lambda W_2|$ sont tous positifs) pour le modèle général (SGM).

En effet, dans la pratique le modèle SGM n'est pas souvent utilisé. On restreint le modèle général soit au modèle avec variable endogène spatiale décalée (SAR) ou soit au modèle avec erreurs autocorrélées (SEM).

Estimation de modèles spatiaux

Autres méthodes d'estimation

Le modèle SAR pose un problème d'endogénéité. Ainsi Anselin (1988a), Kelejian et Robinson (1993) ou Kelejian et Prucha (1998) ont proposé la méthode de variables instrumentales (VI) pour l'estimation du modèle. Pour la variable spatiale décalée W_Y, on peut prendre W_X comme *VI*.

Pour le modèle SEM, Kelejian et Prucha (1998, 1999) ont également développé une approche basée sur la méthode des moments généralisés (GMM).

Tests d'économétrie spatiale

Principes des tests

Les trois grands principes des tests du modèle standard sont toujours valides : test de Wald, test de rapport de vraisemblance et test de multiplicateur de Lagrange.

Le test de Lagrange est le plus utilisé, vu sa simplicité. En effet, le test de Wald exige l'estimation du modèle adéquat. Le test de rapport de vraisemblance contraint également l'estimation de deux modèles (sous l'hypothèse nulle et sous l'alternative). La statistique de ce dernier test est juste la différence des log-vraisemblances des modèles.

Tests d'économétrie spatiale

Principes des tests

Or le test de Multiplicateur de Lagrange nécessitera uniquement l'estimation du modèle sous l'hypothèse alternative.

Tests d'autocorrélation spatiale

Avant d'estimer un modèle spatial, on part d'abord d'un modèle linéaire standard. C'est après estimation de ce dernier qu'on pourra tester si la prise en compte de l'autocorrélation spatiale est importante.

Tests d'autocorrélation spatiale

Test d'autocorrélation spatiale (de Moran) des résidus

Pour vérifier si les résidus du modèle MCO sont spatialement autocorrélés, on peut utiliser le test de Moran :

$$I = \frac{N}{S_0}\left(\frac{\hat{\varepsilon}^t W \hat{\varepsilon}}{\hat{\varepsilon}^t \hat{\varepsilon}}\right)$$

Sous l'hypothèse nulle d'indépendance spatiale on a :

$$E(I) = \frac{tr(MW)}{N-K}$$

$$V(I) = \frac{tr(MWMW^t) + tr(MW)^2 + [tr(MW)]^2}{(N-K)(N-K+2)} - [E(I)]^2$$

Tests d'autocorrélation spatiale

Test d'autocorrélation spatiale (de Moran) des résidus

$M = I_N - X(X'X)^{-1}X'$

Sous H_0 $\left(\frac{I-E(I)}{\sqrt{V(I)}}\right) \rightsquigarrow \mathcal{N}(0,1)$.

Test LM d'autocorrélation spatiale des résidus

On peut également utiliser le test LM sur les résidus. Ce test permet de vérifier si $\lambda = 0$ dans le modèle (GSM).

$$LM_{ERR} = \frac{\hat{\varepsilon}^t W \hat{\varepsilon} / \hat{\sigma}^2}{T} \quad \text{avec} \quad T = tr[(W^t + W)W]$$

LM_{ERR} converge asymptotiquement vers une loi du $\chi^2_{(1)}$.

$\hat{\varepsilon}$ et $\hat{\sigma}^2$ sont calculés à partir du modèle MCO standard.

Tests d'autocorrélation spatiale

Test LM de variable endogène spatiale décalée

On peut aussi tester si $\rho \neq 0$ afin de vérifier s'il n'est pas pertinent d'introduire la variable endogène spatiale décalée dans le modèle.

$$LM_{LAG} = \frac{\hat{\varepsilon}^t WY}{T_1}$$

avec $T_1 = (WX\hat{\beta})^t M(WX\hat{\beta}) + T\hat{\sigma}^2$.

$\hat{\beta}$, $\hat{\varepsilon}$ et $\hat{\sigma^2}$ sont calculés à partir du modèle MCO standard.

LM_{ERR} converge asymptotiquement vers une loi du $\chi^2_{(1)}$.

Tests d'autocorrélation spatiale

Tests LM pour une forme de dépendance spatiale

Les deux tests *LM* précédents ont pour hypothèse nulle le modèle linéaire standard. Autrement dit, ces tests vérifient une mauvaise spécification du modèle linéaire standard contre le modèle SAR ou le modèle SEM.

On pourrait voir également une mauvaise spécification du modèle GSM contre le modèle SAR (test $LM_{ERR}*$) et une mauvaise spécification du modèle GSM contre le modèle SEM (test $LM_{LAG}*$). Ces tests sont donc construits sur le modèle de maximum de vraisemblance.

Tests d'autocorrélation spatiale

Test SARMA et Test du facteur commun

Le test SARMA est construit sur le modèle linéaire. Contrairement aux tests LM_{ERR}, pour le modèle SEM et $LM_{LAG}*$ pour le modèle SAR, le test SARMA vérifie une mauvaise spécification du modèle linéaire standard contre l'un ou l'autre des modèles SAR et SEM. Il est donc plus général que LM_{ERR} et $LM_{LAG}*$.

On a également le test du facteur commun qui permet de choisir entre un modèle SEM et un modèle SAR avec variable exogène décalée (combinaison de SAR + SLX).

Recherche d'une meilleure spécification

- Estimer d'abord le modèle linéaire général et tester à l'aide du test SARMA la présence éventuelle d'une autocorrélation.
- Si le test SARMA notifie une autocorrélation, on peut d'abord inclure certaines variables dans le modèle. En effet la mauvaise spécification indiquée par le test SARMA peut être due à une omission de variable pertinente dans le modèle.
- Si l'autocorrélation persiste, utiliser les tests LM_{ERR} et LM_{LAG} pour détecter s'il s'agit d'une autocorrélation des résidus ou de la variable endogène.

Recherche d'une meilleure spécification

- Si LM_{ERR} seul est significatif, on garde un modèle SEM. Si LM_{LAG} seul est significatif, on garde un modèle SAR. Si les deux sont significatifs, on maintient celui le plus significatif.
- Après avoir choisi entre SAR et SEM, il faut tester si en présence de SAR (resp. de SEM), il faut combiner avec un modèle SEM (resp. SAR) pour avoir un modèle GSM. on pourrait utiliser les tests $LM_{ERR}*$ et $LM_{LAG}*$.
- Il est toujours important de chercher d'autres variables explicatives pour atténuer l'effet de l'autocorrélation. Les critères d'information permettent aussi de choisir le meilleur modèle.

Plan de la présentation

Application sous SAS

Application sous SAS

- **Proc gmap**
- **Macros « régressions spatiales » de Elysée Aristide HOUNDETOUNGAN**