Econométrie spatiale : une introduction pratique*

Jean-Michel Floch (INSEE),

Ronan Le Saout (INSEE-CREST)

Abstract

Ce document de travail décrit la conduite d'une étude d'économétrie spatiale, à travers une modélisation descriptive du taux de chômage par zone d'emploi. Il se concentre sur l'étude de l'autocorrélation spatiale, et donc sur les interactions régionales, plus que sur l'hétérogénéité spatiale, à savoir des phénomènes différenciés spatialement. Plusieurs formes d'interactions existent, relatives à la variable à expliquer, aux variables explicatives ou aux variables inobservées. De nombreux modèles se retrouvent donc en concurrence, à partir d'une définition préalable des relations de voisinage. Une méthodologie pas à pas de choix de modèle (estimation et tests) est détaillée. Des effets de rétroaction entraînent une interprétation particulière (et plus complexe) des résultats.

La théorie économique caractérise de nombreux cas d'interactions entre agents (produits, entreprises, individus), en dehors de la sphère géographique. Les modèles classiques de l'économétrie, notamment lorsque les observations sont supposées indépendantes, tiennent compte de manière parcellaire de ces interactions. Ces modèle spatiaux ont donc une application plus large, l'approche étant cohérente avec tout problème où des relations de "voisinage" interviennent.

Mots Clés : Econométrie spatiale, Modèles d'interaction, Taux de chômage.
Codes JEL : C10, C21, R12.

*Nous remercions Pauline Givord pour ses commentaires sur de premières versions de cette communication. Cette étude ne reflète pas les opinions de l'INSEE.

1 Introduction

La première loi de la géographie (dite de Tobler) affirme "Tout interagit avec tout, mais deux objets proches ont plus de chance de le faire que deux objets éloignés". La disponibilité grandissante de données géolocalisées pose la question de la modélisation de cette proximité lors d'études économiques. Une première étape reste bien sûr de caractériser cette proximité à l'aide d'indicateurs descriptifs et de tests (Floch 2012). Une fois l'autocorrélation spatiale des données détectée vient l'étape de la modélisation dans un cadre multivarié. L'objet de ce document de travail est d'aborder la conduite pratique d'une étude d'économétrie spatiale: quel modèle retenir ? comment interpréter les résultats ? quelles en sont les limites ?

Nous nous appuierons sur l'exemple de la modélisation localisée du taux de chômage à l'aide de quelques variables explicatives décrivant les caractéristiques de la population active, de la structure économique, de l'offre de travail et du voisinage géographique. L'objectif ne sera pas de détailler les résultats d'une étude économique[1] mais d'illustrer les techniques mises en oeuvre : la définition d'une matrice de voisinage qui décrit les relations de proximité, les tests de corrélation spatiale et de spécification, l'estimation et l'interprétation de modèles d'économétrie spatiale.

Les techniques présentées s'adaptent à des domaines qui dépassent le cadre strictement géographique. Plusieurs types de données relationnelles existent en effet: des points (une station service dont on connaît l'adresse), des données par aires géographiques ou administratives (le taux de chômage localisé), des réseaux physiques (des routes interconnectés) ou relationnels (les élèves d'une même classe) ou des données continues. Ce dernier type de données est essentiellement physique, par exemple la hauteur du sol, la température... et ne sera pas abordé par la suite. Un point important à noter est qu'on considère ici des structures de proximité pré-existantes, qui n'évoluent pas ou peu et qui sont de dimension faible, limitée à quelques milliers de zones. C'est clairement le cas pour les découpages par aires géographiques, cela n'est pas le cas pour les réseaux notamment relationnels. Les relations amicales dans Facebook sont par exemple mouvantes et font intervenir des millions d'individus. On ne se pose ainsi pas la question de la caractérisation de réseaux de grande dimension ou de leur formation. On cherche ici à caractériser dans quelle mesure la proximité spatiale (ou relationnelle) influence un résultat, en contrôlant de multiples caractéristiques : le taux de chômage dépend-t-il des régions voisines ? les prix des carburants des stations proches ? la non-réponse à une enquête peut-elle se diffuser spatialement ? Au

[1] Blanc et Hild (2008) traitent cette question de manière détaillée à l'aide d'un modèle d'économétrie spatiale pour la France, Lottmann (2013) pour l'Allemagne.

sein de l'INSEE, ces méthodes ont été utilisées pour étudier la relation entre les prix immobiliers et les risques industriels (Grislain-Letrémy et Katossky 2013), les changements de lieux d'habitation (Guymarc 2015), ou la non-réponse dans l'enquête emploi (Loonis 2012).

Les logiciels commerciaux n'incluent pas de procédures standards d'estimation de modèles d'économétrie spatiale. Des outils spécifiques ont été développés. LeSage et Pace mettent à disposition des programmes MatLab. Luc Anselin a également initié le projet "GeoDa", qui est un logiciel libre d'analyses spatiales. Le logiciel le plus complet pour l'estimation de modèles d'économétrie spatiale reste néanmoins R. Les exemples et les codes seront donc présentés à l'aide de ce logiciel.

La suite est organisée comme suit. La partie 2 présente les raisons économiques et statistiques à la mise en place de ces modèles. La partie 3 rappelle quelques notations de statistique spatiale, notamment les matrices de voisinage. La partie 4 décrit les étapes de l'estimation d'un modèle d'économétrie spatiale. La partie 5 traite de points techniques plus avancés. La partie 6 détaille la mise en oeuvre sous R à travers la modélisation du taux de chômage par zone d'emploi. La partie 7 conclut. Les lecteurs intéressés par approfondir ces méthodes pourront notamment se référer à LeSage et Pace (2009), Elhorst (2013) ou Le Gallo (2002) pour une présentation en langue française.

2 Pourquoi tenir compte de la proximité spatiale ?

2.1 Les raisons économiques

L'interaction spatiale, organisationnelle ou sociale des agents économiques est classique en économie. Anselin (2002) liste ainsi plusieurs dénominations économiques telles que les effets de voisinage, de pair, les interactions stratégiques, la copie par mimétisme ou par les normes sociales ("copy-catting"), la concurrence par comparaison ("yardstick competition")... Il met en avant 2 théories générales justifiant le recours à un modèle spatial ou d'interaction.

Le premier cas est celui où la décision d'un agent économique (une entreprise par exemple) dépendra de la décision des autres agents (ses concurrents). Prenons l'exemple de firmes qui se font concurrence par les quantités (concurrence à la Cournot). La firme i cherche donc à maximiser sa fonction de profit $\Pi(q_i, q_{-i}, x_i)$ en tenant compte de la production de ses concurrents q_{-i} et des ses caractéristiques x_i qui déterminent ses coûts. La solution de ce problème de maximisation est une fonction de réaction de la forme $q_i = R(q_{-i}, x_i)$.

Le deuxième cas est celui où la décision d'un agent économique dépend d'une ressource rare. Si nous reprenons notre exemple d'une firme industrielle, la fonction de profit s'écrit $\Pi(q_i, s_i, x_i)$ avec s_i une ressource rare (qui peut être naturelle, par exemple de l'uranium, ou non, par exemple un composant électronique fabriqué par une seule firme). La quantité s_i qui sera consommée par la firme dépendra alors des quantités consommées par les autres firmes et donc de leur production q_{-i}. On retrouve la fonction de réaction précédente.

Un point important souligné par Anselin (2002) est que ces deux théories amènent à implémenter un même modèle spatial ou d'interaction. Ils sont équivalents d'un point de vue observationnel. Les processus générateurs des données (DGP) sont différents mais fournissent les mêmes observations. De simples données en coupe ne permettront donc pas d'identifier la source de l'interaction (une concurrence stratégique par les quantité ou une concurrence sur les ressources dans notre exemple), seulement de confirmer sa présence et sa force.

2.2 Les raisons économétriques

Les raisons économétriques renvoient aux insuffisances de la modélisation par la méthode des Moindre Carré Ordinaire (MCO) lorsque les hypothèses nécessaires à sa mise en œuvre ne sont plus vérifiées. Le Sage et Pace (2009) présentent ainsi plusieurs arguments techniques justifiant l'emploi de ces méthodes. Avec des données spatiales, on observe fréquemment une autocorrélation spatiale des résidus, i.e. une dépendance entre des observations proches. Cette dépendance des observations peut se traduire soit par une perte d'efficacité des MCO, soit par des estimateurs biaisés. Si le modèle omet une variable explicative spatialement corrélée à la variable d'intérêt, il y a ainsi biais de variable omise. Tenir compte de l'autocorrélation spatiale revient également à définir une forme spécifique pour l"hétérogénéité spatiale. Enfin, la confrontation de plusieurs modèles d'économétrie spatiale permet de discuter l'incertitude du processus générateur des données (DGP).

Les variables explicatives spatialement décalées peuvent intervenir dans le processus générateur des données. D'un point de vue économique, cela s'interprète comme des externalités. Il est ainsi courant en économétrie "classique" d'inclure des variables spatiales du type distance (par exemple au plus proche concurrent), voire des indicateurs agrégés par zone géographique (par exemple le nombre de concurrents). Ce type de variables peut s'interpréter comme des variables spatialement décalées, avec une définition *a priori* de relations de voisinage. L'économétrie spatiale justifie et généralise ainsi ces choix empiriques.

Les raisons économétriques de recourir aux modèles spatiaux sont nombreuses, dans la mesure où les analyses descriptives mettent en évidence des effets de proximité et des corrélations spatiales. La difficulté tient au lien à effectuer entre les raisons économiques et économétriques, et la capacité à produire à partir de ces modèles des analyses mettant en évidence des causalités de nature économique (Gibbons et Overman 2012).

3 Autocorrélation, hétérogénéité, pondérations : quelques rappels de statistique spatiale[2]

3.1 La nature des effets spatiaux dans les modèles de régression

La célèbre phrase de Waldo Tobler, citée en introduction, résume bien les choses, mais les simplifie sans doute un peu. Le problème de la nature des effets spatiaux a été traité par Anselin et Griffiths (1988) repris dans le manuel fondateur d'Anselin (1988). Ils distinguent l'autocorrélation (la dépendance spatiale) et l'hétérogénéité (la non stationnarité spatiale). Divers phénomènes, de mesure (choix du découpage territorial), d'externalités, ou de débordement peuvent conduire à rendre les observations (variable endogène, exogène ou terme d'erreur) dépendantes spatialement. Il y a alors autocorrélation (positive) lorsqu'il y a similarité entre les valeurs observées et leur localisation. L'hétérogénéité spatiale renvoie quant à elle à des phénomènes d'instabilité structurelle dans l'espace. Les variables explicatives peuvent être les mêmes mais ne pas avoir le même effet en tout point. Les paramètres du modèle sont alors variables. Le terme d'erreur peut être différent selon la zone géographique. On parle alors d'hétérogénéité spatiale. Par exemple, pour définir l'indice des prix de l'immobier ancien INSEE-Notaires, environ 300 strates sont définies selon la nature du bien (appartement ou maison) et la zone géographique. Le prix du m^2, d'une pièce complémentaire ou d'une autre caractéristique est en effet supposé différent selon ces différentes strates. Le marché est segmenté.

Ce document traite principalement des méthodes de prise en compte de l'autocorrélation spatiale dans les modèles de régression, détaillés en partie 4. La variabilité locale des paramètres peut néanmoins être appréhendée de plusieurs façons. Lorsque l'on dispose d'une bonne connaissance du territoire d'intérêt, la plus simple est l'introduction d'indicatrices

[2]D'autres théories statistiques ont été développées pour analyser les données continues (géostatistique) et les configurations de points. On pourra se référer à Floch (2012) pour un courte introduction à ces théories alternatives.

relatives aux sous-territoires, potentiellement croisées avec les variables explicatives. Des approches plus complexes venues du monde de la géographie existent, qui seront présentées en partie 5. La régression géographique pondérée (Brunsdon, Fotheringham et Charlton 1996) consiste par exemple à estimer en chaque point un modèle sur un voisinage, le poids des voisins étant décroissant avec la distance.

Ce partage "pédagogique" entre autocorrélation et hétérogénéité ne doit pas faire oublier les interactions entre les deux (Anselin et Griffith, 1988; Le Gallo 2002 & 2004). Il n'est pas toujours facile de faire le partage entre ces deux composantes, et la mauvaise spécification de l'une peut être la cause de l'autre. Les tests classiques de l'hétéroscédasticité (i.e. une forme particulière d'hétérogénéité sur le terme d'erreur) sont affectés par l'autocorrélation spatiale, et inversement les tests d'autocorrélation spatiale le sont par l'hétéroscédasticité. Il n'y a pas de solution simple pour intégrer simultanément ces deux phénomènes, en dehors du simple ajout d'indicatrices de territoires dans les modèles d'autocorrélation. De plus, la corrélation des valeurs observées entraîne que l'information apportée par les données est moins riche que celle où les données sont indépendantes. En cas d'auto-corrélation, on oberve une seule réalisation du processus générateur des données. Estimer des modèles par sous-territoires pour étudier l'hétérogénéité des comportements n'a alors pas de sens. Tout ceci plaide pour une approche exploratoire préalable des données. Selon la question, la méthodologie traitera en premier lieu l'aucorrélation des observations ou l'hétérogénéité des comportements.

3.2 La matrice des poids

Pour mesurer la corrélation spatiale entre agents ou zones géographiques, tout commence par définir *a priori* les relations de voisinage entre les agents ou les zones géographiques. Ces relations ne peuvent pas être estimées par le modèle. Si nous observons N régions, il y a $N(N-1)/2$ couples différents de régions. Il n'est donc pas possible d'identifier des relations de corrélation entre ces N régions sans faire des hypothèses sur la structure de cette corrélation spatiale. Pour N agents ou zones géographiques, cela revient à définir une matrice carré de taille $N \times N$, dont les éléments diagonaux sont nuls (on ne peut pas être son propre voisin). La valeur des éléments non diagonaux sont le fruit de l'expertise. De nombreuses matrices de voisinage ont été proposées dans la littérature :

- Les matrices de contiguité qui associent à chaque voisin immédiat la valeur 1 (et 0 dans le cas contraire). Bien que cette approche semble simple, elle laisse place à l'interprétation. Un voisin "immédiat" peut en effet être défini de plusieurs manières

selon le mode de déplacement entre les zones. A l'instar d'un jeu d'échecs où la reine se déplace librement mais où le fou ne de déplace qu'en diagonal et la tour qu'horizontalement ou verticalement, il peut exister des éléments naturels (une rivière par exemple) qui explique qu'un voisin "immédiat" ne soit pas directement accessible.

- Des matrices tenant compte de la distance d entre les zones géographique (1 si $d < d_0$,$1/d^2$, e^{-2d}... les relations devenant plus faibles avec la distance). Des distances d'arrêt sont utilisées pour limiter le nombre d'éléments non nuls. Le calcul de la distance (entre les centroïdes, les frontières...) ou la définition de cette distance d'arrêt engendre de nombreux problèmes pratiques. Une distance trop faible peut créer de nombreuses "îles" (des zones sans voisins). Choisir une distance telle que chaque zone ait au moins un voisin peut créer des zones avec un nombre très important de relations de voisinage. Griffith (1996) précise qu'il est préférable d'utiliser une matrice sous-spécifiée (distance inférieure à la distance optimale) à une matrice surspécifiée (distance supérieure).

- Des matrices tenant compte de la force des relations entre les zones, par exemple le % de frontières communes. Le poids de voisinage entre deux zones i et j peut ainsi être défini par $b_{ij}^{\alpha}/d_{ij}^{\beta}$ avec b_{ij} une mesure de la force des relations entre les zones i et j (qui n'est pas forcément symétrique) tels que le % de frontières communes, la population, la richesse et d_{ij} la distance entre les zones. Le choix de $\alpha = 1$ et $\beta = 2$ correspond à un modèle du type gravitaire.

- Les k plus "proches" voisins, une mesure de la proximité non géographique entre agents (être dans la même classe par exemple)...

Les matrices de voisinage sont considérées exogènes dans la majorité des applications d'économétrrie spatiale (Anselin 2002). Il ne faut donc pas créer des poids de voisinage qui seraient fonctions du phénomène qu'on cherche à expliquer. Elles peuvent néanmoins inclure des paramètres à estimer, par exemple si on considère des poids de voisinage $b_{ij}^{\alpha}/d_{ij}^{\beta}$ où les paramètres α et β sont inconnus. Ce cas reste néanmoins rare en pratique.

Pour facilier l'interprétation, ces matrices de voisinage sont le plus souvent normées par ligne (i.e. que la somme des éléments par ligne vaut 1). Pour une matrice de contiguité, si une région a k voisins, chaque terme non nul de la k-ème ligne sera ainsi égal à $1/k$. Le terme Wy s'interpète alors simplement comme la moyenne du voisinage pour la variable y. En l'absence d'accord sur la "meilleure" matrice de voisinage, il est courant de vérifier que les résultats sont robustes à ce choix en testant plusieurs matrices de voisinage possibles.

3.3 Les méthodes exploratoires

Avant de mettre en place un modèle d'économétrie spatiale, il convient de vérifier qu'il y a bien un phénomène spatial à prendre en compte. Cela commence par une caractérisation de l'autocorrélation spatiale à l'aide de représentations graphiques (carte) et de tests statistiques.

Le principal indicateur[3] est celui de Moran qui mesure l'association globale , $I = \frac{N}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij}(y_i - \overline{y})(y_j - \overline{y})}{\sum_i (y_i - \overline{y})^2}$ avec w_{ij} le poids de la i-ème ligne et j-ème colone de la matrice de voisinage W. Une valeur proche de 1 sera synonyme de corrélation positive (les zones avec de hautes ou de basses valeurs pour y se regroupent), une valeur proche de -1 sera synonyme de corrélation négative (des zones géographiques proches ont des valeurs de y très différentes). Sous l'hypoyhèse H_0 d'absence d'autocorrélation spatiale ($I = 0$), la statistique $I^* = \frac{I - \mathbb{E}(I)}{\sqrt{\mathbb{V}(I)}}$ suit asymptotiquement une loi normale $\mathcal{N}(0, 1)$. Rejeter l'hypothèse nulle du test de Moran revient donc à conclure à la présence d'autocorrélation spatiale. Ce test reste bien sûr dépendant du choix de la matrice de voisinage W. De plus, il ne signifie pas qu'un modèle d'économétrie spatiale soit nécessaire mais qu'un tel modèle doit être envisagé (cf. partie 4). Il ne peut en effet refléter que la répartition spatiale d'une variable sous-jacente.

Des indicateurs locaux (par zone géographique i, dits LISA pour Local Indicators of Spatial Association) ont été définis pour mesurer la propension d'une zone à regrouper de fortes ou faibles valeurs de y ou au contraire des valeurs très diverses. Pour chaque zone, on calcule un indicateur de Moran local $I_i = N \cdot y_i \cdot \frac{\sum_j w_{ij}(y_j - \overline{y})}{\sum_i (y_i - \overline{y})^2}$ qui sous l'hypothèse nulle d'absence d'autocorrélation spatiale locale tend asymptotiquement vers une loi normale. Cet indicateur permet donc d'identifier les régions d'un intérêt particulier, des points "chauds" (regroupement de fortes valeurs de y) et "froids" (regroupement de faibles valeurs de y).

4 Estimer un modèle d'économétrie spatiale

4.1 La galaxie des modèles d'économétrie spatiale

Elhorst (2010) a établi une classification des principaux modèles d'économétrie spatiale. Le modèle de Manski (1993) est le plus général. Il distingue 3 types d'interaction spatiale :

[3] Les indicateurs de Geary et de Getis et Ord, ainsi que les autres indicateurs locaux, sont présentés dans Floch 2012.

- Une interaction endogène, i.e. que la décision économique d'un agent va dépendre de la décision de ses voisins ;
- Une interaction exogène, i.e. que la décision économique d'un agent va dépendre des caractéristiques observables de ses voisins ;
- Une corrélation spatiale des effets liée à de mêmes caractéristiques inobservées.

Ce modèle s'écrit sous forme matricielle[4] :

$$Y = \rho \cdot WY + X \cdot \beta + WX \cdot \theta + u$$

$$u = \lambda \cdot Wu + \varepsilon$$

Avec les paramètres β pour les variables explicatives exogènes, ρ pour l'effet d'interaction endogène (de dimension 1) dit autorégressif spatial, θ pour les effets d'interaction exogène (de dimension le nombre de variables exogènes K) et λ pour les effets de corrélation spatiale des erreurs dit autocorrélation spatiale.

Le modèle de Manski (1993) n'est pas identifiable dans le cas général. Une première solution est de supposer que les matrices de voisinage W ne sont pas identiques pour les 3 interactions spatiales. Il y aurait par exemple des relations de voisinage définies par W_ρ pour le paramètre autorégressif et W_λ pour l'autocorrélation spatiale, ce qui est difficile à justifier en pratique. Une autre solution est de supprimer l'un des effets d'interaction parmi les $K+2$ paramètres du modèle (ρ, θ et λ). C'est la solution privilégiée dans la littérature empirique.

La matrice de voisinage doit respecter plusieurs contraintes techniques (Lee 2004; Elhorst 2010) pour assurer notamment le caractère inversible des matrices $I - \rho W$ et $I - \lambda W$. On peut retenir que les matrices usuelles de contiguité ou de distance inverse respectent ces contraintes. Ce n'est pas forcément le cas de matrice "atypique" créée par exemple pour les relations de proximité sociale. Il n'est par exemple pas possible d'avoir des îles (un agent qui

[4] Par souci de simplification, la constante du modèle est ici incluse dans la matrice des variables explicatives X. Dans le cas d'une matrice de contiguité, $W \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ représente le nombre moyen de voisins de chaque observation. Si ce nombre de voisins est le même pour tous les individus, les deux termes ne sont pas identifiables séparemment. De plus, le nombre moyen de voisins n'a pas forcément un sens économique clair. C'est pourquoi on trouve dans la littérature une présentation des modèles où la constante n'est pas incluse dans la matrice des variables explicatives X.

n'a pas de voisin) ou qu'au contraire tout le monde soit le voisin de tout le monde. On sait de plus que $|\rho| < 1$ et $|\lambda| < 1$ (critère qu'on peut intuitivement rapprocher des conditions de stationnarité pour les solutions d'un modèle de type $ARMA$).

Deux principaux types de modèles peuvent être déduits du modèle de Manski (1993) selon la contrainte utilisée, $\theta = 0$ ou $\lambda = 0$[5].

Dans le cas $\theta = 0$, on trouve le modèle de Kelejan-Prucha (ou également nommé SAC, Spatial Autoregressive Confused):

$$Y = \rho \cdot WY + X \cdot \beta + u$$

$$u = \lambda \cdot Wu + \varepsilon$$

Les estimateurs du modèle de Kelejan-Prucha présentent le défaut d'être biaisés et non convergents si le vrai modèle inclut des interactions exogènes (LeSage et Pace 2009). Il y a en effet dans ce cas biais de variables omises. Ce n'est pas le cas pour le modèle spatial de Durbin ($\lambda = 0$, SDM, Spatial Durbin model):

$$Y = \rho \cdot WY + X \cdot \beta + WX \cdot \theta + \varepsilon$$

Ce dernier modèle aura de plus des statistiques de test valides même si le vrai modèle est à erreurs autocorrélées spatialement (SEM). Ces deux modèles incluent en effet les cas particuliers du modèle spatial autorégresssif (SAR, Spatial AutoRegression) $Y = \rho \cdot WY + X \cdot \beta + \varepsilon$ et du modèle à erreurs autocorrélées spatialement (SEM, Spatial Error Model) $Y = X \cdot \beta + u$ et $u = \lambda \cdot Wu + \varepsilon$. Pour obtenir ce dernier modèle à partir du modèle spatial de Durbin, on pose $\theta = -\rho\beta$ (hypothèse dite de facteur commun). Le modèle SDM s'écrit dans ce cas, $Y = X \cdot \beta + \rho \cdot W(Y - X \cdot \beta) + \varepsilon$. En notant $u = Y - X \cdot \beta$, on retrouve bien le modèle SEM. Le modèle à interactions exogènes (noté SLX, Spatial Lag X) correspond au cas $\lambda = \rho = 0$ et $\theta \neq 0$.

Des versions plus générales de ces modèles ont été developpées, qui autorisent les effets de voisinage à varier selon l'ordre de voisinage ou selon les interactions prises en compte. Ils correspondent à des versions spatiales des modèles temporelles ARMA(p,q). Ils restent peu employés dans la littérature empirique.

[5]Le cas $\rho = 0$ (Modèle SDEM, Spatial Durbin Error Model) peut également être envisagé mais est d'un usage moins courant car, à la différence des 2 autres contraintes, il n'inclut pas les 2 principaux modèles d'économétrie spatiale SAR et SEM.

4.2 Critères statistiques du choix de modèle

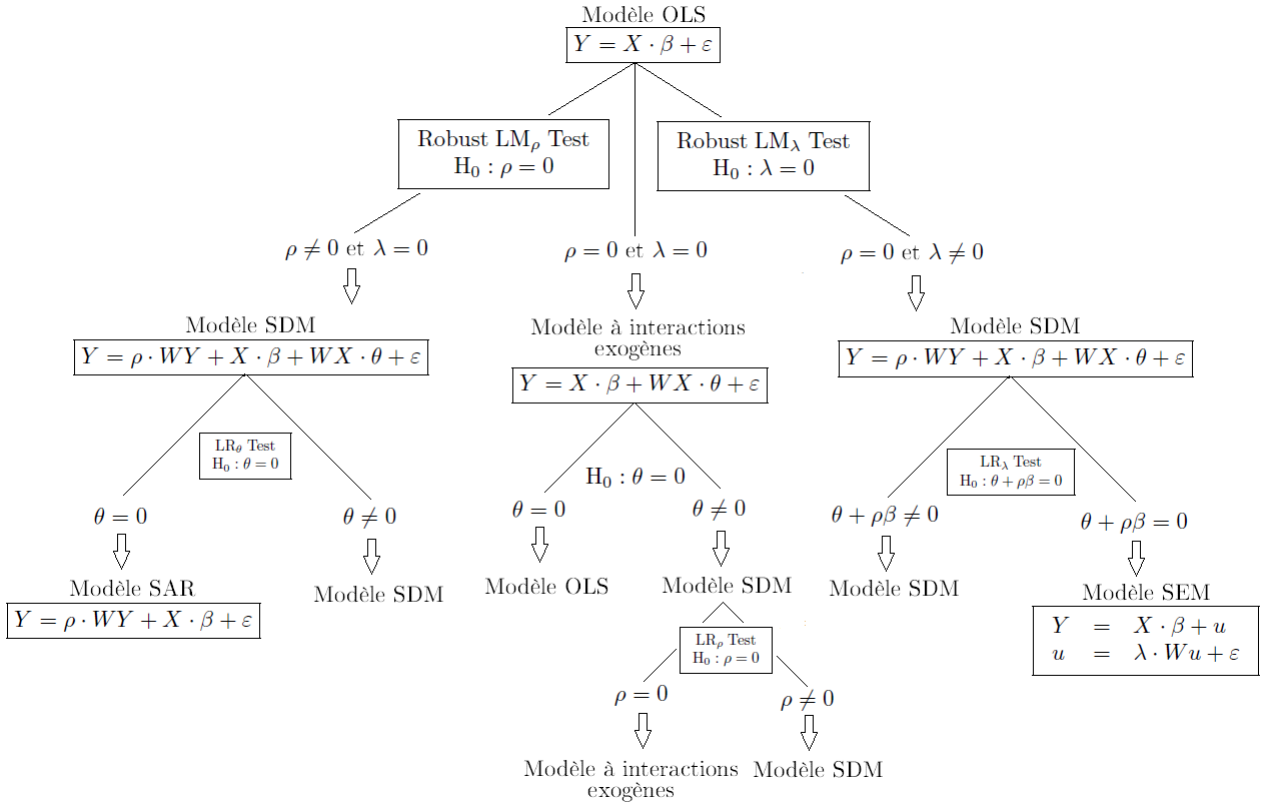
Deux approches principales ont été utilisées pour le choix des modèles (cf. annexe 1 pour une représentation graphique). La première dite "approche ascendante" ou bottom-up consiste à partir du modèle non spatial (Le Gallo 2002 pour une synthèse). Des tests du multiplicateur de Lagrange (Anselin et al. 1996 pour des tests de spécification des modèles SAR et SEM, robustes à la présence d'autres types d'interactions spatiales) permettent ensuite de trancher entre le modèle SAR, SEM ou le modèle non spatial. Cette approche a été celle plébiscitée jusqu'aux années 2000 car les tests développés par Anselin et al. (1996) s'appuient sur les résidus du modèle non spatial. Ils sont donc peu coûteux d'un point de vue computationnel. Florax et al. (2003) a également montré, à l'aide de simulations, que cette procédure était la plus performante dans le cas où le vrai modèle est un modèle SAR ou SEM. La deuxième dite "approche descendante" ou top-down consiste à partir du modèle spatial de Durbin. Avec des tests du rapport de vraisemblance, on en déduit le modèle le plus adapté aux observations. L'amélioration des performances informatiques a permis de rendre aisée l'estimation de ces modèles plus complexes, dont le modèle spatial de Durbin pris comme référence dans le livre de LeSage et Pace (2009).

Elhorst (2010) propose une approche "mixte" représentée en figure 1. Elle consiste à commencer par l'approche ascendante mais, en cas d'interactions spatiales ($\rho \neq 0$ ou $\lambda \neq 0$), au lieu de choisir directement un modèle SAR ou SEM, on étudie le modèle spatial de Durbin. Cela permet de confirmer à l'aide de plusieurs tests (Multiplicateur de Lagrange, Rapport de Vraisemblance) la pertinence du modèle choisi. Cela permet également d'intégrer les interactions exogènes dans l'analyse. Enfin, en cas d'incertitude, c'est le modèle a priori le plus robuste (le modèle spatial de Durbin) qui est choisi. Prenons le cas où, à partir des résidus du modèle OLS, les tests du multiplicateur de Lagrange (LM_ρ et LM_λ)[6] concluent à la présence d'autocorrélation endogène, i.e. $\rho \neq 0$ et $\lambda = 0$ (branche de gauche du graphique 1). On estime alors le modèle SDM. A l'aide d'un test du rapport de vraisemblance ($\theta = 0$), on peut alors choisir entre le modèle SAR et le modèle SDM. Dans le cas où les tests concluent à la pésence d'autocorrélation résiduelle, i.e.$\rho = 0$ et $\lambda \neq 0$ (branche de droite du graphique 2), un test du rapport de vraisemblance de l'hypothèse de facteur commun ($\theta = -\rho\beta$) permet de choisir entre le modèle SEM et le modèle SDM. Dans le cas où les tests soulignent l'absence d'autocorrélation, i.e. $\rho = 0$ et $\lambda = 0$, le modèle à interactions exogènes (SLX) est estimé. Des tests du rapport de vraisemblance permettent de choisir entre les modèles OLS, SLX

[6] Il existe deux versions de ces tests, l'une robuste à la présence d'autres formes d'autocorrélation spatiale, l'autre non (Anselin *et al.* 1996).

et SDM. Enfin, dans le cas où les tests concluent à la présence simultanée d'autoccorélation endogène et résiduelle, i.e. $\rho \neq 0$ et $\lambda \neq 0$, le modèle SDM est estimé.

Graphique 1 : Approche d'Elhorst (2010) pour le choix d'un modèle d'économétrie spatiale



Ces approches "pratiques" supposent que la matrice de voisinage soit connue et que les variables explicatives soient exogènes. Elles s'appuient sur une estimation par maximum de vraisemblance des modèles. D'autres méthodes d'estimation existent. Dans le cas de variables explicatives endogènes, Fingleton et le Gallo (2007, 2008 et 2012) proposent une estimation par variables instrumentales et la méthode des moments généralisée. Pour calculer la précision des effets directs et indirects (cf. 4.3), LeSage et Pace (2009) proposent une estimation bayésienne par MCMC (Monte-Carlo Markov Chain).

Elles ne doivent pas être prises comme des règles intangibles, mais plutôt comme de bonnes pratiques. Il ne sert en effet à rien d'estimer directement un modèle SAR, complexe à interpréter, si ni l'analyse économique, ni l'analyse statistique ne le justifie.

4.3 L'interprétation des résultats : attention aux rétroactions

L'économétrie spatiale s'écarte du cadre habituel des MCO. Pour expliquer les différents indicateurs associés à l'interprétation du modèle, nous reprenons le cadre de LeSage et Pace (2009).

Le modèle SAR est $Y = \rho \cdot WY + X\beta + \varepsilon$ avec $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Il peut se ré-écrire de plusieurs manières, en notant r l'indice pour une variable explicative et S_r des matrices carrés de la taille du nombre d'observations:

$$\begin{aligned} Y &= (1-\rho W)^{-1} X\beta + (1-\rho W)^{-1} \varepsilon \\ &= \sum_{r=1}^{k} S_r(W) X_r + (1-\rho W)^{-1} \varepsilon \\ &= \sum_{r=1}^{k} (1-\rho W)^{-1} \beta_r X_r + (1-\rho W)^{-1} \varepsilon \end{aligned}$$

$$\text{Avec } Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \text{ et } S_r(W) = \begin{pmatrix} S_r(W)_{11} & S_r(W)_{12} & \cdots & S_r(W)_{1n} \\ S_r(W)_{21} & S_r(W)_{22} & & \\ \vdots & \vdots & \ddots & \\ S_r(W)_{n1} & S_r(W)_{n2} & \cdots & S_r(W)_{nn} \end{pmatrix}$$

La valeur prédite est donc $\hat{y} = (1-\hat{\rho}W)^{-1} X\hat{\beta}$ et non $X\hat{\beta}$ comme dans un modèle MCO classique.

On a de plus $\mathbb{E}(y) = (1-\rho W)^{-1} X\beta$. L'effet marginal (pour une variable quantitative) d'une modification de la variable X_{ir} (pour l'individu i) n'est pas β_r mais $S_r(W)_{ii}$, la valeur diagonale de rang i de la matrice S_r. A la différence des séries temporelles où il n'y a qu'une direction à prendre en compte (y_t dépend de y_{t-1} qui n'est expliquée que par des valeurs passées), l'économétrie spatiale est multidirectionnelle. Une modification de mon territoire impacte mes voisins, ce qui m'impacte en retour. Il faut en tenir compte pour l'analyse globale des résultats.

Par ailleurs, l'effet marginal apparait différent pour chaque zone (comme pour un modèle Logit ou Probit par exemple). Les termes diagonaux de la matrice S_r sont les effets directs, pour chaque zone, d'une modification de la variable X_r dans la même zone. Les autres termes représentent des effets indirects, i.e. l'impact de la modification de la variable X_r dans une zone sur une autre zone. Au niveau global, plusieurs indicateurs peuvent donc être calculés pour synthétiser les résultats (LeSage et Pace 2009) :

- L'effet direct moyen correspond à la moyenne des termes diagonaux de la matrice S_r, i.e. $\frac{1}{n}\text{trace}\,(S_r)$. C'est cet indicateur qui est le plus proche de l'interprétation des coefficients β calculés par MCO (ou des effets marginaux moyens d'un modèle Logit ou Probit);

- L'effet total moyen sur une observation représente l'effet sur la zone i d'une modification d'une unité de la variable X_r dans toutes les zones, i.e. $\sum_k S_r\,(W)_{ik}$. Il y a donc n valeurs possibles. L'indicateur moyen est la moyenne $\frac{1}{n}\sum_i\sum_k S_r\,(W)_{ik}$.

- L'effet total moyen d'une observation représente l'effet d'une modification d'une unité de la variable X_r dans la zone j (X_{jr}) sur l'ensemble des zones, i.e. $\sum_k S_r\,(W)_{kj}$. Il y a donc n valeurs possibles. L'indicateur moyen est la moyenne $\frac{1}{n}\sum_j\sum_k S_r\,(W)_{kj}$.

- Ces deux derniers indicateurs donnent le même résultat, bien que leur construction soit différente. Ils permettent de calculer l'effet indirect moyen comme la différence entre l'effet direct moyen et l'effet total moyen d'une observation.

Une autre proposition est de se baser sur la décomposition en séries entières $(1-\rho W)^{-1} = (I_n + \rho W + \rho^2 W^2 + \cdots)$. Abreu, de Groot et Florax (2005) adopte ainsi la décomposition suivante:

$$\frac{\partial y}{\partial x_r} = \underbrace{I_n\beta_r}_{\text{Effets Directs}} + \underbrace{\rho W\beta_r}_{\text{Effets Indirects}} + \underbrace{\left[\rho^2W^2 + \rho^3W^3 + \cdots\right]\beta_r}_{\text{Effets Induits}}$$

Cette présentation peut néanmoins être trompeuse car les effets induits incluent des termes diagonaux expliquant que l'effet marginal ne soit pas β_r.

Dans tous les cas, le calcul de la précision de ces estimateurs est complexe. LeSage et Pace (2009) s'appuient ainsi sur des simulations bayésiennes MCMC[7]. Il est clair également que ces effets dépendent en premier lieu du voisinage proche. On peut noter que l'effet direct moyen est supérieur en valeur absolue à l'effet marginal du modèle MCO, $|S_r| > |\beta_r|$. Les termes diagonaux de la matrice de voisinage W sont en effet nuls. Le premier terme de rétroaction (et qui domine les autres termes d'ordre supérieur) est donc proportionnel à ρ^2.

[7]Les méthodes de Monte-Carlo par Chaîne de Markov sont des algorithmes d'échantillonnage permettant de générer des échantillons d'une loi de probabilité complexe (pour en déduire par exemple la précision d'une statistique). Elles s'appuient sur un cadre bayésien et une chaîne de Markov dont la loi limite est la distribution à échantilloner.

L'analyse des effets par ordre de voisinage (distinguer l'effet direct, l'effet des voisins, des voisins des voisins...) est également proposée dans la littérature.

Pour l'interprétation globale du modèle, il est utile de calculer pour chaque variable, l'effet direct moyen ($\frac{1}{n}\text{trace}\,(S_r)$) et l'effet indirect moyen ($\frac{1}{n}\left[\sum_j\sum_k S_r\,(W)_{kj} - \text{trace}\,(S_r)\right]$). Les effets indirects demeurent plus complexes à interpréter. Calculer l'effet induit par l'espace ($\frac{1}{n}trace\,(S_r) - \hat{\beta}_r$) permet également d'illustrer la force des effets de rétroaction.

5 Limites et difficultés économétriques

5.1 Que faire des données manquantes ?

En économétrie classique, on observe un échantillon de n individus. Si quelques individus présentent des valeurs manquantes, ils sont exclus de l'analyse. Cela réduit la taille de l'échantillon mais pas la mise en oeuvre des méthodes économétriques.

En économétrie spatiale, on observe une seule réalisation du processus générateur des données (Une analogie peut être effectuée avec les séries temporelles, les paramètres d'un modèle $ARMA$ étant estimés à l'aide d'une seule trajectoire temporelle.). Si l'observation de la distribution spatiale est incomplète (il y a des valeurs manquantes), il n'est pas possible d'estimer le modèle. Une solution est alors d'interpoler les valeurs manquantes à l'aide de techniques de géostatistique (Anselin 2001).

Une autre implication est qu'il n'est pas aisé de mettre en place ces techniques sur données individuelles d'enquête. On observe en effet dans ce cas uniquement des relations de voisinage partielles, pour les seules individus enquêtés (et sous réserve que le questionnaire inclut une interrogation de ce type). Il faut alors faire une hypothèse complémentaire et très forte que les observations des voisins non enquêtés sont exogènes, i.e. qu'elles ne modifient pas les effets de voisinage pour les seuls individus enquêtés. Cela pourrait être vérifié par exemple si le plan de sondage ne dépend pas des zones géographiques, ce qui est rarement le cas.

5.2 Le choix de la matrice de poids

Pour définir une matrice de voisinage, les contraintes sont fortes, puisque l'on recherche une description simple (afin que le modèle soit identifiable), mais adéquate des relations entre territoires. Des conceptions différentes sont exposées par ceux qui cherchent à mettre

plus d'économie dans l'économétrie (Corrado et Fingleton 2012, Harris 2011) et ceux qui considèrent la matrice de poids comme "le plus grand mythe" de l'économétrie spatiale (Lesage et Pace 2012). La majorité des auteurs souligne la sensibilité des résultats au choix de cette matrice, alors que Lesage et Pace (2012) souligne que ces conclusions proviennent d'une mauvaise interprétation des modèles. Les effets directs et indirects seraient plus robustes au choix de W que les estimateurs des paramètres, qui n'ont eux pas d'interprétation immédiate. On peut néanmoins souscrire à la remarque de Harris (2011) "The spatial econometrics point out the importance of W choice but tells us little about adequate foundations for these choice", difficultés qui ont contribué au scepticisme de plusieurs économistes (Gibbons et Overman 2012). Ces considérations montrent la complexité de la détermination de W qui reste un sujet de controverses, sur lesquelles il n'y a pas de consensus scientifique.

On a vu que les modèles traitent en général la matrice W comme exogène. D'autres méthodes (Alstadt et Getis 2006) s'appuient néanmoins sur les données utilisées pour déterminer la matrice des poids. Un exemple en est fourni par l'utilisation des indices de Getis et Ord (Floch 2012), qui calculent l'autocorrélation en fonction de la distance. On cherche alors les distances qui correspondent à des ruptures dans la valeur des indicateurs. Il est également possible d'estimer les poids à travers les modèles économétriques avec des contraintes fonctionnelles *a priori* faibles (Bhattacharjee et Jensen-Butler 2013). Ces dernières approches sont souvent lourdes en calcul et plus difficiles à implémenter. De plus, on constate facilement qu'une description plus réaliste et plus conforme à la réalité économique risque d'introduire de l'endogénéité. Dans notre exemple, une matrice pourrait être construite à partir des migrations domicile-travail. Mais ces mobilités sont en rapport avec le niveau de l'emploi et du chômage (et servent même à construire le découpage territorial). L'emploi d'une telle matrice est donc plus délicate. Des travaux faisant intervenir des matrices endogènes ont été récemment proposés (Kelejian et Piras 2014).

5.3 Et si le phénomène est hétérogène spatialement ?

Deux formes d'hétérogénéité existent. La première est l'hétéroscédasticité. Les paramètres du modèle sont les mêmes mais pas sa variabilité. Les modèles d'économétrie spatiale du type SEM correspondent à une forme particulière d'hétérogénéité (Le Sage et Pace 2009). Une autre solution serait de définir la forme de l'hétéroscédasticité spatiale de la matrice de variance-covariance (Dubin 1998). La deuxième correspond à la variabilité spatiale des paramètres ou de la forme fonctionnelle du modèle. Elle est souvent traitée dans la littérature empirique en ajoutant des indicatrices de zones géographiques dans le modèle. Des méthodes

plus complexes ont été développées (Le Gallo 2004). Elles restent en grande partie descriptive et exploratoire, car leur comportement théorique n'est pas complètement connu, notamment la convergence et la prise en compte des ruptures géographiques. Il existe des méthodes de lissage géographique où la constante (voire chaque variable explicative) est croisée avec des polynômes des coordonnées géographiques. Il existe également des régressions locales dont l'application spatiale est la régression géographique pondérée[8] (Brunsdon *et al.* 1996).

Le principe est simple. Pour chaque observation, on estime le modèle sur son voisinage en pondérant les observations selon leur distance. Le modèle linéaire classique peut être vu comme un cas particulier où les coefficients sont stables dans l'espace.

Formellement, le modèle s'écrit de la manière suivante:

$$Y_i = \sum_{j=0}^{p} \beta_j (u_i, v_i) X_{ij} + \epsilon_i$$

Avec (u_i, v_i) le couple désignant les coordonnées du point i et $\beta_j (u_i, v_i)$ les paramètres pour chaque observation de la variable j.

Les estimateurs sont obtenus en minimisant (à l'instar des MCO)

$$\sum_{k=1}^{n_h} w_k (u_i, v_i) \left[Y_i - \sum_{j=0}^{p} \beta_j (u_i, v_i) X_{ij} \right]^2$$

Les poids utilisés sont construits à partir de fonctions décroissantes de la distance au point d'estimation (u_i, v_i), dont les plus courants sont:

- La fonction qualifiée de gaussienne $w_k(u_i, v_i) = exp\left[- \left(d_{ik}/h \right)^2 \right]$;
- La fonction biweight$w_k(u_i, v_i) = \begin{cases} \left[1 - \left(d_{ik}/h \right)^2 \right]^2 & si \;\; d_{ik} \leq h \\ 0 \;\; si \;\; d_{ik} > h \end{cases}$.

Comme dans toutes les méthodes utilisant des noyaux, la fenêtre h a un impact beaucoup plus important que la forme fonctionnelle. Il est donc important de déterminer cette fenêtre. Plusieurs méthodes ont été proposées, dont la plus utilisée est celle de la validation croisée. On cherche la valeur de h qui minimise la quantité$\Delta (h) = \sum_{i=1}^{n} \left(y_i - \widehat{y}_{\neq i} (h) \right)^2$ où $\hat{y}_{\neq i} (h)$ est l'estimateur obtenu en utilisant la fenêtre h et en éliminant l'observation i.

[8] GWR, Geographically Weighted Regression

5.4 Le risque de régression "écologique"

Les méthodes présentées dans ce document s'appuie sur des zonages géographiques prédéfinis (une zone d'emploi dans notre exemple). De nombreuses variables économiques sont ainsi définies en cohérence avec une division administrative (région, département, canton). Or ce découpage administratif ne correspond pas forcément à la réalité économique des relations entre agents. Ce phénonomène géographique est connue sous l'acronyme MAUP ("Modifiable Areal Unit problem"). Il a plusieurs conséquences (Floch 2012). Avec des découpages ou des échelles différentes, les résultats des modèles et les interactions entre agents ne sont pas identiques. Il faut également tenir compte de l'étendue spatiale des zones, 1000 agents économiques n'interagissent pas de la même manière dans 1 km^2 ou dans 10000 km^2. Il n'y a pas de solution pour résoudre le problème du MAUP. Lorsque des données individuelles sont disponibles (par exemple les caractéristiques d'emploi issues du recencement de la population plutôt que les taux de chômage par zone d'emploi), il est en effet possible de faire abstraction de ce découpage administratif ou de construire le niveau géographique *a priori* le plus pertinent. Mais si de tels traitements sont effectués sur les variables d'intérêt, d'autres variables explicatives ne resteront disponibles qu'à des niveaux géographiques prédéfinis.

De plus, les données utilisées sont souvent agrégées, au sens où elles représentent la moyenne de nos variables d'intérêt sur une zone géographique. En économétrie "classique", l'utilisation de données agrégées entraîne des problèmes d'identification et d'hétéroscédasticité connue sous le nom de *régression écologique*. Anselin (2002) donne l'exemple d'un modèle où les décisions d'un individu i, y_{ik}, s'expliquent par ces caractéristques x_{ik} mais également par les caractéristiques du groupe k auquel il appartient $\overline{x}_k = \sum_i x_{ik}/n_k$. Le modèle s'écrit $y_{ik} = \alpha + \beta \cdot x_{ik} + \gamma \cdot \overline{x}_k + \varepsilon_{ik}$ où β représente l'effet individuel et γ l'effet de contexte. Si les données sont agrégées, le modèle devient $\overline{y}_k = \alpha + (\beta + \gamma) \cdot \overline{x}_k + \overline{\varepsilon}_k$. Il n'est alors plus possible d'identifier séparément les paramètres β et γ. Le modèle est hétéroscédastique car $\mathbb{V}(\overline{\varepsilon}_k) = \sigma^2/n_k$ dans le cas de perturbations initiales i.i.d. de variance σ^2.

Le problème est encore plus complexe dans le cas de modèles spatiaux. Il n'est en effet pas possible d'agréger une matrice de voisinage W définie au niveau individuel. Avec des données individuelles, un individu i du groupe k peut avoir des voisins parmi le groupe k mais également parmi un autre groupe k'. Si on considère désormais une matrice de voisinage agrégée au niveau groupe, les relations intra-groupes ne seront plus pris en compte (la diagonale est nulle par hypothèse). De plus, il peut y avoir de nombreux individus du groupe k voisins d'individus du groupe k' mais très peu voisins d'un autre groupe k''. Avec

une matrice de contiguité agrégée au niveau groupe, la force des relations individuelles ne sera plus pris en compte (chaque voisin a le même poids). Au delà des problèmes d'identification d'une *régression écologique*, un modèle SAR défini au niveau individuel ne peut pas être agrégé pour correspondre à un modèle SAR défini à un niveau supérieur. Il n'y a pas de relations simples entre les paramètres.

Pour bien comprendre cette question, prenons l'exemple du marché immobilier. On observe des villes dont les prix sont très élevés au centre et diminuent ensuite progressivement. Il existe également des niveaux de prix très différents entre les villes. Si on ne considère que des prix moyens par centre urbain (regroupant des villes proches), la disparité des prix au sein des villes sera cachée. Ces emboîtements d'échelle peuvent engendrer des résultats à première vue paradoxaux.

En pratique, cela signifie que l'interprétation des résultats n'est valable que pour le découpage géographique choisi. Si on étudie des relations économiques à un niveau agrégé avec un modèle spatial, on ne peut rien dire des relations individuelles entre agents. Pour tenir compte de cette imbrication des zones géographiques (régions, départements, cantons, individus) et rendre les analyses cohérentes entre elles, une solution est alors de mener des analyses multi-niveaux. Dans le cas d'études macro-économiques telles que la croissance régionale, ce problème est moins présent. Le niveau agrégé est en effet celui pertinent.

6 Mise en pratique sous R

Dans cette partie, nous détaillons la mise en pratique d'une étude d'économétrie spatiale, en modélisant le taux de chômage localisé (par zone d'emplois) à l'aide de caractéristiques structurelles relatives aux caractéristiques de la population active (% des peu diplômés et des moins de 30 ans dans la population active), de la structure économique (% des emplois dans le secteur industriel) et du marché du travail (taux d'activité), ainsi que par le dynamisme du marché du travail mesuré par la variation d'emploi. L'objectif de ce modèle est descriptif et illustratif. Il ne sera pas de détailler les résultats d'une étude économique mais d'illustrer les techniques mises en oeuvre : la définition d'une matrice de voisinage qui décrit les relations de proximité entre agent, les tests de corrélation spatiale et de spécification, l'estimation, et l'interprétation de modèles d'économétrie spatiale. D'autres variables peuvent bien sûr expliquer les taux de chômage locaux. Les variables économiques sont supposées structurelles et peu variables à court terme (hormis la variation de l'emploi). Pour limiter les problèmes

d'endogénéité, le taux de chômage est calculé sur l'année 2013 et les variables explicatives correspondent au millésime 2011 de CLAP (Connaissance locale de l'appareil productif) et du RP (Recensement de la Population). Une interprétation causale reste néanmoins impossible. De nombreuses variables ont en effet été omises de l'analyse, par exemple sur l'offre d'emploi. Les variables explicatives prises en compte peuvent ainsi intégrer l'effet de ces variables omises et non leur seul effet propre. Enfin, le décalage temporel entre les variables explicatives et le taux de chômage ne supprime pas complètement le caractère simultané des phénomènes (par exemple entre le taux d'activité et le taux de chômage), structurellement stables à court terme.

Les logiciels commerciaux n'incluent pas de procédures standards d'estimation de modèles d'économétrie spatiale. Des outils spécifiques ont été développés. LeSage et Pace mettent à disposition des programmes MatLab. Luc Anselin a également initié le projet "GeoDa", qui est un logiciel libre d'analyses spatiales. Le logiciel le plus complet pour l'estimation de modèles d'économétrie spatiale reste néanmoins R. Les exemples et les codes seront donc présentés à l'aide de ce logiciel. Nous listons ci-dessous quelques packages utiles dans R :

- L'importation et la représentation de cartes : *sp* et *rgdal* pour la définition des objets spatiaux, *maptools* pour la définition de cartes;
- Des fonctions similaires à celles de SIG (Système d'Information Géographique) du type calcul de distance ou des méthodes de géostatistique: *fields*, *raster* et *gdistance*;
- L'économétrie spatiale : *spdep* pour l'ensemble des modèles classiques, *splm* pour les modèles de panel, *spatialprobit* pour le modèle Probit et *spgwr* pour la régression géographique pondérée.

Ces packages s'installent à l'aide de la commande *install.packages('Nom Package',dependencies=TRUE)*, l'attribut *dependencies=TRUE* permettant d'installer les autres packqges associés. Ils s'appellent ensuite à l'aide la commande *library('Nom Package')*. R est un logiciel open source. Des packages plus spécialisés ou nouveaux sont décrit sur le site

http://cran.r-project.org/web/views/Spatial.html.

6.1 La gestion des données géographiques

Le logiciel R fournit des commandes permettant de mettre en œuvre les modèles d'économétrie spatiale de manière simple. En pratique, toute étude sur des données géographiques nécessite

un travail préalable de mise en forme de ces données. Deux types de bases de données sont en effet utilisées :

- Le fichier de définition des zones géographiques et de ses attributs (points, lignes, polygones...) sous format *shapefile* (principaux SIG), *MIF/MID* (MapInfo) ou *KML* (Google Earth) ;
- La base de données classique des variables explicatives, avec un identifiant pour les zones géographiques.

Le premier fichier correspond donc à une carte qui permettra de visualiser les résultats obtenus (valeurs prédites par les différents modèles, résidus...) mais également de constituer la matrice de poids à l'aide d'outils disponibles sous R. Il est donc nécessaire de disposer d'un fonds cartographique correspondant au maillage territorial utilisé dans l'étude. L'Insee dispose de nombreux fonds de carte destinés à produire des cartes à l'aide du logiciel MapInfo. L'importation de ces cartes est efffectuée sous R à l'aide du package *rgdal* et de la commande *readOGR*. Le package *rgdal* permet la lecture de données vectorielles (sous la forme de points, de lignes ou de polygones). Tous les fichiers de définition de la carte (fichiers MapInfo Z10F.tab, Z10F.map.. dans notre exemple) doivent être placés dans le répertoire de travail. L'importation des données statistiques se fait à l'aide des outils classiques d'importation *read.table*, *read.csv*.

```
### Importation des données
> donnees_ze=read.csv("donnees_dt_eco_spatiale.csv",
colClasses=c('character','character',rep('numeric',32)))
### Importation du fonds de carte
> file<-"Z10F.tab"
> carte<-readOGR(file,layer="Z10F")
### Affichage de la carte de France
> plot(carte,cex=.01)
```

L'une des principales difficultés est que les formats des données géographiques ne sont pas toujours les mêmes selon les logiciels. Les exemples présentés à l'aide de fichiers MapInfo sont volontairement simplifiés. L'importation et la lecture de données géo-localisées peut néanmoins être complexe, ce qui dépasse le cadre de ce document. On pourra se référer à Bivand *et al.* 2013 pour des détails sur l'utilisation de données géographiques sous R.

Les données statistiques et le fonds cartographique sont appariés selon l'ordre de tri des zones d'emploi. Il est donc utile de vérifier que ces zones d'emploi sont triées de la même

façon dans les 2 fichiers et que l'appariement a été correctement effectué. Dans le cas présent, l'instruction *isTRUE* vérifie que les codes des zones d'emploi sont les mêmes dans le fichier de données statistiques (variable ze2010) et le fichier cartographique (variable codgeo) en vérifiant que les types des objets sont les mêmes.

```
### Vérification de la validité de l'appariement entre les données statistiques et la carte
> isTRUE(all.equal(donnees_ze$ze2010,carte@data$codgeo))
[1] TRUE
```

6.2 Définir une matrice de voisinage

Pour définir une matrice de contiguité, l'objet géographique, de type polygone dans notre exemple, doit être transformé en un objet de type .nb. Cette opération est effectuée à l'aide de la commande *poly2nb*, du package *spdep*, lorsqu'on prend un critère de contiguïté. La relation de contiguité esr du type "Tour" par défaut.

```
### Définition des voisins
> carte.nb<-poly2nb(carte)
> summary(carte.nb)
# Le paramètre Queen=True permet de définir une relation de contiguité du type "Reine".
```

L'objet carte.nb contient 1606 liens. Parmi les zones d'emploi, il y a en moyenne un peu plus de cinq liens. 68 zones ont 5 voisins. Quatre ont un seul lien. Les objets .nb fournissent un moyen de stockage de l'information. Les liens peuvent être visualisés sous la forme d'un graphe, construit autour des centroïdes des zones.

Ces objets de type .nb permettent ensuite de constituer les matrices de proximité, à l'aide de la commande *nb2listw* du package *spdep*. L'objet produit est un objet de type liste (et se termine par .w), et non une matrice. Par rapport à l'objet .nb, on a en plus création des poids, avec plusieurs possibilités. Le paramètre *style* permet de calculer des matrices de proximité brutes, codifiées en 0 ou 1, des matrices standardisées en ligne (les plus fréquemment utilisées dans la littérature) ou des matrices standardisées en ligne ou en colonne.

```
### Matrice de contiguité standardisé en ligne, méthode par défaut
> carte.w<-nb2listw(carte.nb,style="W")
### Matrice de contiguité binaire
> carte.w<-nb2listw(carte.nb,style="B")
### Matrice de contiguité standardisé de manière globale
> carte.w<-nb2listw(carte.nb,style="C")
```

D'autres critères que la contiguité peuvent être utilisés pour définir une matrice de voisinage. On peut la construire à partir des plus proches voisins, à l'aide de plusieurs commandes. Il faut d'abord récupérer à partir de la carte les coordonnées des centroïdes des zones. Ces coordonnées sont transformées en distances à l'aide de la commande *knearneigh* (du package *rann*). La commande *knn2nb* permet ensuite de créer des listes de voisins avec la même logique que la commande *poly2nb*. Par la suite, nous utiliserons 3 matrices de voisinage des plus proches voisins: une définition restreinte à 2 voisins seulement, une définition cohérente avec le nombre moyen de voisins de la matrice de contiguité (5), et une définition large avec 10 voisins.

```
### Récuperation des centroides
> coor<-coordinates(carte)
### Définition des 10 plus proches voisins
#   k indique le nombre de voisins
> carte.knn<-knearneigh(coor,k=10)
> kn10.nb<- knn2nb(carte.knn)
### Visualisation des relations de voisinage
> plot(kn10.nb, coor, col="red",cex=0.1,add=TRUE)
### Matrice de voisinage des 10 plus proches voisins
> kn10.w<-nb2listw(kn10.nb,style="W")
```

Dans le cas de pondérations décroissantes avec la distance, l'exemple ci-dessous détaille le cas simple où la pondération décroît comme le carré de la distance dans un rayon de 100 kms. La matrice de poids est créée à l'aide de la commande *mat2listw*. Il y a en moyenne 15 liens de voisinage avec cette matrice.

```
### Matrice de voisinage fondée sur la distance
> distance<-rdist(coor,coor)
> diag(distance)<-0
> distance[distance>=100000]<-0
> dist<- (distance/100000*distance/100000)
> dist.w<-mat2listw(dist, row.names = NULL, style="W")
```

Dernier exemple, il est possible de créer une matrice de voisinage spécifique à la question posée. C'est par exemple le cas pour des données non géographiques (proximité sociale, entre produits industriels...). Nous définissons une dernière matrice de voisinage, dont la pondération est proportionnelle aux déplacements domicile-travail entre les zones d'emploi. Il y a en moyenne 9 liens de voisinage avec cette matrice. Le code de création de cette matrice est fourni en annexe 2. Nous la nommerons matrice endogène par la suite car les poids de voisinage sont a priori corrélés aux variables explicatives du modèle.

6.3 Cartographie et tests

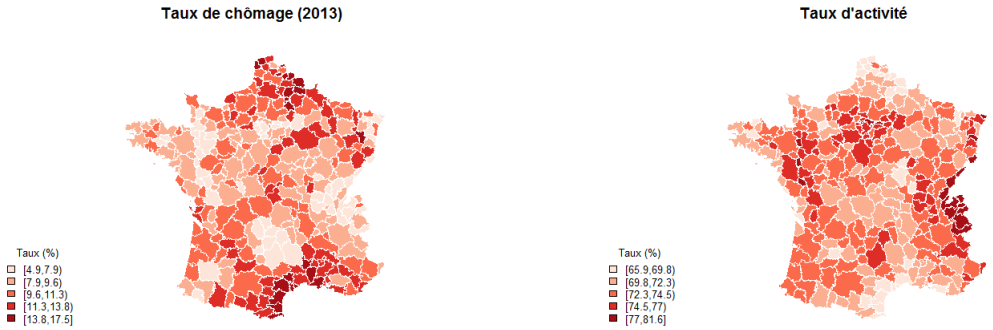
Avant de mettre en place un modèle d'économétrie spatiale, il convient de vérifier qu'il y a bien un phénomène spatial à prendre en compte. Cela commence par une caractérisation de l'autocorrélation spatiale à l'aide de représentations graphiques (carte) et de tests statistiques (Moran). Il est utile, lors du travail exploratoire sur les données, ou lors de la validation des modèles de cartographier les résultats, en complément des analyses descriptives. Cela permet d'apprécier de façon visuelle s'il y a une spatialisation forte des phénomènes.

6.3.1 Cartographie

L'objectif n'est pas ici de présenter les nombreuses techniques cartographiques disponibles dans R, mais de donner quelques exemples utiles. Ces instructions nécessitent les packages *rcolorBrewer* (pour la gestion des couleurs) et *classInt* (pour le partage en classes de la variable cartographiée). Pour une introduction à la cartographie, on peut utilement se référer à R pour les géographes du groupe ElementR. Les codes des cartes présentées sont fournies en annexe 2.

La carte 1 représente les taux de chômage par zone d'emploi en 2013. On constate des zones polarisées, ce qui pourrait être le signe d'une hétérogénéité spatiale. Le Nord de la France et le Languedoc-Roussillon présentent ainsi des taux de chômage plus élevés, les zones frontalières de la Suisse plus faibles. Les zones contigues de ces régions ont des taux de chômage proches également, ce qui est caractéristique d'une autocorrélation spatiale. Pour les variables explicatives, on constate notamment une polarisation forte du pourcentage d'emploi industriel. Les taux d'activité présentent une structuration spatiale proche du taux de chômage.

Carte 1 : Distribution du taux de chômage et d'activité, par zone d'emploi



Le tableau 1 décrit la distribution des variables. Le taux de chômage moyen est de 10%, pour un taux d'activité de 73%. Il y a 22% d'actifs peu diplômés et de jeunes actifs de moins de 30 ans. Le nombre d'emploi diminue de 1%. Hormis pour le pourcentage d'emploi industriel, les écarts inter-quartiles sont faibles, inférieurs à 5%. Le pourcentage d'emploi industriel apparaît comme la variable la plus polarisée.

Tableau 1 : Descriptif de l'échantillon

	N	Moyenne	Ecart-Type	Min	Q25	Médiane	Q75	Max
Taux de chômage	297	10.0	2.37	4.9	8.3	9.6	11.4	17.5
Taux d'activité	297	72.8	2.65	65.9	71.3	72.8	74.2	81.6
% Actifs Peu Diplômés	297	22.1	3.61	13.0	19.5	22.2	24.8	32.2
% Jeunes Actifs15-30 ans	297	21.8	2.02	16.7	20.4	21.8	23.2	27.7
% Emploi Industriel	297	19.7	8.82	3.7	13.3	18.2	24.8	52.0
Variation de l'emploi	297	-1.03	4.04	-13.60	-3.89	-0.86	1.46	10.90

Note de lecture: La zone géographique est la zone d'emploi. Les statistiques ne sont pas pondérées.

6.3.2 Tests d'autocorrélation spatiale et représentations graphiques avancées

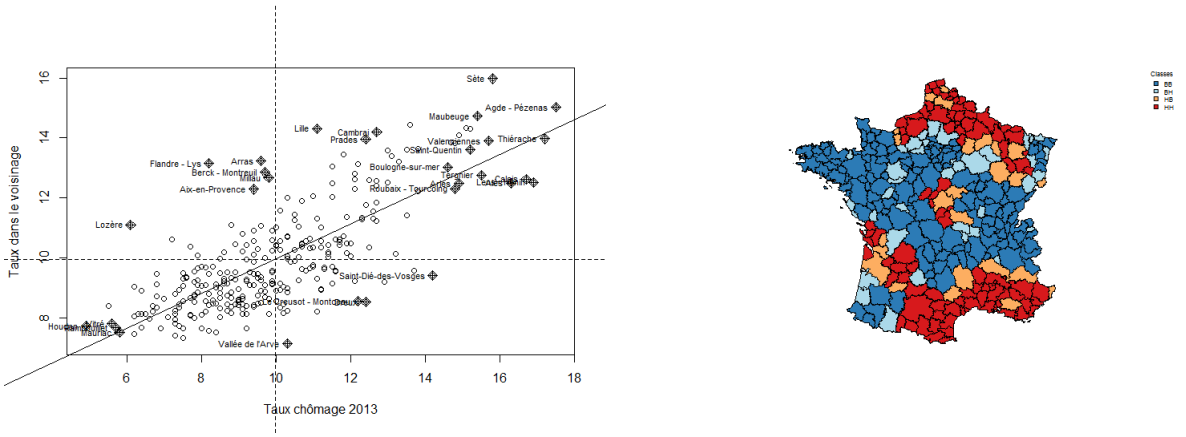
Le package *spdep* fournit des commandes permettant de calculer les indicateurs de Moran (et autres indicateurs spatiaux) pour les valeurs brutes, au niveau global ou local. La commande *moran.test(donnees_ze$txcho_2013, dist.w)* effectue un test de Moran pour le taux de chômage en utilisant la matrice de distance inverse. L'hypothèse nulle est l'absence d'autocorrélation contre une hypothèse alternative d'autocorrélation positive. La p-value quasiment nulle indique que l'hypothèse nulle doit être rejetée. Le résultat est robuste au choix de la matrice de voisinage. Les indicateurs de type LISA peuvent aussi être calculés à l'aide de la commande *localmoran* du package *spdep*.

L'autocorrélation des données brutes peut être illustrée graphiquement à l'aide du graphique de Moran. Il met en relation la valeur observée en un point et celle qui est observée dans le voisinage déterminé par la matrice de poids. Le package *spdep* permet de produire ce graphique à l'aide de la commande *moran.plot*.

```
### Graphique de Moran
> moran.plot(x=donnees_ze$txcho_2013,dist.w,xlab="Taux chômage 2013",
ylab="Taux dans le voisinage",labels=as.character(donnees_ze$libelle_ze))
```

Le graphique 2 est cohérent avec les résultats du test de Moran. Une relation linéaire apparaît entre le taux de chômage d'une zone et celui de son voisinage. Une carte peut être associé permettant de situer les zones d'emploi en fonction de leurs caractéristiques (HH signifie un taux de chômage élevé dans un environnement élevé, HB un taux élevé dans un environnement plus bas...). Elle permet de constater que cette relation n'est pas homogène sur le territoire. Le Nord et le Sud présentent des taux de chômage élevés. Une France du "milieu" présente au contraire des taux de chômage moindres.

Graphique 2 : Graphique de Moran du taux de chômage



6.4 Estimation et choix de modèles

L'analyse descriptive a permis de constater que l'espace n'était pas neutre pour caractériser les taux de chômage locaux. Il n'est néanmoins pas certain qu'un modèle économétrique tenant compte de l'espace soit nécessaire. Le nuage de points des taux de chômage et d'activité montrent une forte relation linéaire des 2 variables. Le taux de chômage et d'activité sont tous les deux corrélés spatialement. Le taux de chômage pourrait donc être relié au taux d'activité, sans autre forme d'autocorrélation spatiale que celle présente dans les deux variables. Première chose, on commence donc par estimer le modèle MCO. Un test de Moran adapté sur les résidus confirme la présence résiduelle d'autocorrélation spatiale (potentiellement associée à de l'hétérogénéité spatiale), quelle que soit la matrice de voisinage.

Pour déterminer la forme de l'autocorrélation spatiale (endogène, exogène ou inobservée), la démarche est pragmatique. L'approche d'Elhorst (2010) conduirait à retenir le modèle SDM. Seuls les modèles MCO et SDM seraient alors estimés. Dans un but pédagogique, l'ensemble des modèles spatiaux sont néanmoins estimés, pour 6 matrices de voisinage: contiguité, plus proches voisins (2, 5 ou 10), distance inverse, et proportionnelle aux trajets domicile-travail (dite matrice endogène). Les régressions s'estiment à l'aide du package *spdep*. Le coût computationnel à estimer ces modèles est par ailleurs faible.

```
### Modèle OLS
> ze.lm <- lm(txcho_2013 ~ tx_act + part_act_peudip + part_act_1530
+ part_emp_ind + var_emp, data=donnees_ze)
> summary(ze.lm)

### Test de Moran adapté sur les résidus
```

```
> lm.morantest(ze.lm,dist.w)

### Test LM-Error et LM-Lag
> lm.LMtests(ze.lm,dist.w,test="LMerr")
> lm.LMtests(ze.lm,dist.w,test="LMlag")
> lm.LMtests(ze.lm,dist.w,test="RLMerr")
> lm.LMtests(ze.lm,dist.w,test="RLMlag")

### Modèle SAR
> ze.sar<-lagsarlm(txcho_2013 ~ tx_act + part_act_peudip + part_act_1530
+ part_emp_ind + var_emp, data=donnees_ze,dist.w)
> summary(ze.sar)

### Modèle SEM
> ze.sem<-errorsarlm(txcho_2013 ~ tx_act + part_act_peudip + part_act_1530
+ part_emp_ind + var_emp, data=donnees_ze,dist.w)
> summary(ze.sem)

### Modèle SDM
> ze.sardm<-lagsarlm(txcho_2013 ~ tx_act + part_act_peudip + part_act_1530
+ part_emp_ind + var_emp, data=donnees_ze,distte.w,type="mixed")
> summary(ze.sardm)
### Test de l'hypothèse de facteur commun
# ze.sardm : Modèle non contraint
# ze.sem : Modèle contraint
> durbin.test<-LR.sarlm(ze.sardm,ze.sem)
> print(durbin.test)
```

On ne présente ici que les résultats associés à la matrice de distance inverse, car c'est celle qui présente le caractère explicatif le plus fort (AIC les plus faibles) et dont l'interprétation économique est la plus intuitive. Les zones d'emploi n'ayant pas la même taille, la contiguité ou les plus proches voisins peut engendrer des effets inattendus. La matrice endogène peut par construction provoquer un biais des estimateurs. Les résultats sur le choix de modèle restent néanmoins cohérents, quelque soit la matrice de voisinage retenue.

Nous nous attendonc ici à une relation négative entre taux de chômage et taux d'activité, mais positive pour le pourcentage d'actifs peu diplômés et de jeunes actifs. Le "halo" du chômage est moins présent dans les zones dynamiques en termes d'emploi. Les personnes les moins diplômées et les jeunes sont réputés plus marqués par le chômage. Les zones de fort emploi industriel sont a priori plus sensibles au chômage (réaction de l'emploi à la consommation et fermeture d'usines). La variation de l'emploi cherche à mesurer les effets de création ou de perte d'emploi sur le chômage. On peut penser que cette relation est négative. Nos suppositions ne sont pas vérifiées pour ces deux dernières variables.

Tableau 2 : Déterminants du taux de chômage par zone d'emploi, à partir d'une matrice inverse de la distance

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	MCO	SEM	SAR	SDM	SAC	SLX	SDEM	Manski
Taux d'activité	-0.559***	-0.458***	-0.375***	-0.433***	-0.458***	-0.428***	-0.438***	-0.435***
	(0.035)	(0.039)	(0.036)	(0.041)	(0.040)	(0.049)	(0.039)	(0.041)
% Actifs Peu Diplômés	0.197***	0.184***	0.142***	0.179***	0.182***	0.173***	0.177***	0.179***
	(0.027)	(0.028)	(0.023)	(0.028)	(0.027)	(0.034)	(0.028)	(0.029)
% Jeunes Actifs 15-30 ans	0.163***	0.222***	0.116***	0.237***	0.216***	0.234***	0.228***	0.239***
	(0.044)	(0.046)	(0.037)	(0.047)	(0.046)	(0.056)	(0.046)	(0.047)
% Emploi Industriel	-0.038***	-0.002	-0.017*	-0.002	-0.003	-0.010	-0.008	-0.001
	(0.011)	(0.010)	(0.009)	(0.010)	(0.010)	(0.012)	(0.010)	(0.010)
Variation de l'emploi	0.022	-0.013	0.000	-0.018	-0.014	-0.016	-0.014	-0.018
	(0.024)	(0.020)	(0.020)	(0.020)	(0.021)	(0.024)	(0.021)	(0.020)
$\hat{\rho}$			0.522***	0.640***	0.136			0.680***
			(0.050)	(0.062)	(0.121)			(0.129)
$\hat{\lambda}$		0.756***			0.684***		0.665***	-0.089
		(0.049)			(0.086)		(0.061)	(0.266)
$\hat{\theta}$, Taux d'activité				0.201***		-0.201***	-0.204**	0.229**
				(0.070)		(0.069)	(0.093)	(0.099)
$\hat{\theta}$, % Actifs Peu Diplômés				-0.136***		-0.017	-0.008	-0.144***
				(0.048)		(0.055)	(0.070)	(0.050)
$\hat{\theta}$, % Jeunes Actifs 15-30 ans				-0.123 *		0.012	0.043	-0.134*
				(0.071)		(0.082)	(0.116)	(0.075)
$\hat{\theta}$, % Emploi Industriel				-0.040**		-0.095***	-0.056**	-0.038*
				(0.018)		(0.021)	(0.028)	(0.020)
$\hat{\theta}$, Variation de l'emploi				0.037		0.078*	0.049	0.035
				(0.039)		(0.046)	(0.061)	(0.038)
Constante	43.569***	34.449***	26.721***	17.962***	33.252***	49.141***	48.397***	15.960***
	(2.840)	(3.262)	(2.891)	(4.266)	(3.595)	(3.641)	(7.263)	(6.788)
Observations	297	297	297	297	297	297	297	297
AIC	1083	973	995	968	975	1040	972	970
R^2 Ajusté	0.609					0.667		
Test Moran	0.000					0.000		
Test LM-Error	0.000					0.000		
Test LM-Lag	0.000					0.000		
Test Robuste LM-Error	0.000					0.847		
Test Robuste LM-Lag	0.000					0.001		
Test Facteur Commun				0.008				
Test LM residual auto.			0.000	0.715				

Note de lecture: L'ensemble des modèles est estimé avec une matrice inverse de la distance (avec un seuil à 100 kms). Les écarts-types sont indiqués entre parenthèses. Pour les tests, la p-value est indiquée. Significativité: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Concernant le choix du modèle, on peut retenir les points suivants du tableau 2:

- Les critères statistiques conduiraient à retenir un modèle SDM. L'ensemble des tests d'autocorrélation spatiale menés à partir des résidus du modèle MCO sont rejetés. De même, l'hypothèse de facteur commun du modèle SDM est rejeté. Le modèle SDM présente l'AIC le plus faible. Enfin, pour le modèle à interactions exogènes (SLX), on ne rejette pas l'hypothèse d'absence d'autocorrélation résiduelle sous l'hypothèse d'une autocorrélation endogène (Test Robuste LM-Error, p-value=0.847).

- Pour des raisons de parcimonie, le choix d'un modèle SEM, voire SDEM, pourrait être envisagé. Les critères AIC sont en effet proches du modèle SDM. L'interprétation de ces modèles est plus aisée mais se limite aux effets directs.

- Le choix d'un modèle SAR serait ici mauvais. Un test montre qu'une autocorrélation spatiale résiduelle reste présente. Les conséquences sont importantes sur l'interprétation des résultats. C'est le seul modèle pour lequel la variable "Pourcentage d'emploi industriel" reste significative à 10% (5% avec les autres matrices de voisinage), alors que le signe négatif est contre-intuitif. Les autres modèles corrigent au contraire ce biais spatial du modèle MCO. Le modèle est plus complexe, sans gain en termes de biais dans notre exemple.

- Le modèle des Manski fournit des résultats divergents selon la matrice de voisinage, certainement par manque d'identifiabilité de ce modèle. Le modèle SAC (autocorrélation endogène et résiduelle) estime une autocorrélation endogène faible et non significative en comparaison de l'autocorrélation résiduelle. Ce résultat est difficile à interpréter. Il peut provenir d'un biais lié à la non prise en compte des interactions exogènes (cf. partie 4.1). Mais il peut égalementrenforcer le choix d'un modèle SEM pour des raisons de parcimonie.

Les divergences de résultats (pour différentes matrices de voisinage) sont analysées pour les modèles SEM et SDM. Le modèle SEM peut s'interpréter comme le modèle MCO. L'effet marginal correspond bien aux paramètres du modèle. Cette comparaison est cohérente avec un biais du modèle MCO. Pour le taux d'activité, l'effet est surévalué de 0.08 à 0.1 point par rapport au modèle SEM. Pour le pourcentage d'emploi indutriel, le modèle MCO conclut à tort à un effet négatif significatif alors qu'il est jugé nul avec le modèle SEM. Les résultats pour différentes matrices de voisinage sont cohérents, mais peuvent présenter des écarts. L'effet du taux d'activité pourrait être sur-évalué avec un matrice de contiguité ou un nombre faible de plus proches voisins. L'effet du pourcentage de jeunes actifs semble sous-évalué avec une matrice endogène. Pour le modèle SDM (tableau en annexe 3), une interprétation directe n'est pas possible car les effets doivent tenir compte des effets d'interaction endogène. On constate des effets d'interactions exogènes variables selon la matrice de voisinage.

Tableau 3 : Modèle SEM, pour différentes matrices de voisinage

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	MCO	SEM	SEM	SEM	SEM	SEM	SEM
		Contiguité	2 Voisins	5 Voisins	10 Voisins	Distance	Endogène
Taux d'activité	-0.559***	-0.482***	-0.487***	-0.490***	-0.464***	-0.458***	-0.469***
	(0.035)	(0.039)	(0.039)	(0.039)	(0.038)	(0.039)	(0.040)
% Actifs Peu Diplômés	0.197***	0.193***	0.213***	0.188***	0.183***	0.184***	0.188***
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.028)	(0.027)
% Jeunes Actifs 15-30 ans	0.163***	0.203***	0.211***	0.226***	0.228***	0.222***	0.163***
	(0.044)	(0.046)	(0.045)	(0.046)	(0.047)	(0.046)	(0.046)
% Emploi Industriel	-0.038***	-0.007	-0.010	-0.006	-0.007	-0.002	-0.008
	(0.011)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.011)
Variation de l'emploi	0.022	-0.004	0.011	-0.011	-0.013	-0.013	-0.010
	(0.024)	(0.020)	(0.020)	(0.020)	(0.021)	(0.020)	(0.021)
$\hat{\lambda}$		0.699***	0.519***	0.692***	0.771***	0.756***	0.701***
		(0.049)	(0.046)	(0.049)	(0.052)	(0.049)	(0.044)
Constante	43.569***	36.593***	36.310***	36.668***	34.940***	34.449***	36.256***
	(2.840)	(3.244)	(3.191)	(3.314)	(3.174)	(3.262)	(3.309)
Observations	297	297	297	297	297	297	297
AIC	1083	983	1001	980	981	973	1004

Note de lecture: Le modèle SEM est estimé avec 6 matrices de voisinage différentes. Les écarts-types sont indiqués entre parenthèses. Significativité: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

6.5 Interprétation des résultats

Pour le modèle SDM, pour permettre une interprétation au regard du modèle MCO et SEM, on calcule les effets directs et indirects tels que décrits dans la partie 4.4. Les intervalles de confiance empiriques sont obtenus à l'aide de 1000 simulations MCMC (Monte-Carlo Markov Chain). Pour les effets directs, on retrouve l'interprétation du modèle SEM. Pour les effets indirects, seul le pourcentage d'emploi industriel a un effet négatif significatif. Ces effets indirects ont en effet une variabilité plus grande, qui ne permet pas de conclure sur les effets éventuels. Il met en avant le rôle particulier du poucentage d'emploi industriel, qui n'aurait qu'un effet indirect. Le modèle SDM peut amener à interpréter de manière fallacieuse l'autocorrélation endogène, qui n'a pas ici une interprétation économique claire. Au vu de ces résultats, le modèle SEM pourrait ainsi être privilégié par principe de parcimonie.

Tableau 4 : Impacts directs du modèle SDM, pour différentes matrices de voisinage

	(1) MCO	(2) SDM Contiguité	(3) SDM 2 Voisins	(4) SDM 5 Voisins	(5) SDM 10 Voisins	(6) SDM Distance	(7) SDM Endogène
Taux d'activité	-0.559	-0.462	-0.468	-0.479	-0.461	-0.444	-0.453
	[-0.629,-0.490]	[-0.544,-0.389]	[-0.539,-0.391]	[-0.555,-0.405]	[-0.540,-0.385]	[-0.517,-0.372]	[-0.524,-0.373]
% Actifs Peu Diplômés	0.197	0.175	0.209	0.183	0.179	0.176	0.180
	[0.144,0.250]	[0.121,0.231]	[0.155,0.266]	[0.132,0.239]	[0.129,0.233]	[0.120,0.232]	[0.130,0.234]
% Jeunes Actifs 15-30 ans	0.163	0.234	0.245	0.251	0.248	0.241	0.218
	[0.077,0.249]	[0.137,0.321]	[0.156,0.335]	[0.161,0.342]	[0.157,0.341]	[0.150,0.330]	[0.126,0.314]
% Emploi Industriel	-0.038	-0.013	-0.017	-0.011	-0.010	-0.007	-0.014
	[-0.060,-0.016]	[-0.034,0.006]	[-0.039,0.003]	[-0.031,0.011]	[-0.030,0.010]	[-0.027,0.014]	[-0.036,0.007]
Variation de l'emploi	0.022	-0.012	0.009	-0.010	-0.010	-0.014	-0.017
	[-0.025,0.068]	[-0.053,0.027]	[-0.029,0.051]	[-0.051,0.030]	[-0.050,0.031]	[-0.052,0.027]	[-0.058,0.023]

Note de lecture: Le modèle SDM est estimé avec 6 matrices de voisinage différentes. Les intervalles de confiance empirique (quantiles à 2.5% et 97.5% de 1000 simulations MCMC) sont indiqués entre crochets.

Tableau 5 : Impacts indirects du modèle SDM, pour différentes matrices de voisinage

	(1) SDM Contiguité	(2) SDM 2 Voisins	(3) SDM 5 Voisins	(4) SDM 10 Voisins	(5) SDM Distance	(6) SDM Endogène
Taux d'activité	-0.192	-0.104	-0.118	-0.185	-0.202	-0.326
	[-0.411,0.019]	[-0.217,0.015]	[-0.328,0.076]	[-0.441,0.073]	[-0.445,0.037]	[-0.523,-0.153]
% Actifs Peu Diplômés	-0.032	-0.055	-0.032	-0.062	-0.057	-0.041
	[-0.218,0.138]	[-0.147,0.035]	[-0.193,0.130]	[-0.319,0.169]	[-0.267,0.165]	[-0.201,0.120]
% Jeunes Actifs15-30 ans	0.082	-0.036	-0.008	0.090	0.076	0.025
	[-0.186,0.371]	[-0.193,0.104]	[-0.272,0.233]	[-0.245,0.469]	[-0.238,0.390]	[-0.226,0.279]
% Emploi Industriel	-0.110	-0.046	-0.070	-0.112	-0.110	-0.094
	[-0.193,-0.034]	[-0.085,-0.010]	[-0.139,-0.002]	[-0.231,-0.018]	[-0.205,-0.021]	[-0.178,-0.020]
Variation de l'emploi	0.046	0.017	0.080	0.057	0.068	0.096
	[-0.138,0.230]	[-0.063,0.097]	[-0.068,0.240]	[-0.188,0.275]	[-0.130,0.279]	[-0.093,0.277]

Note de lecture: Le modèle SDM est estimé avec 6 matrices de voisinage différentes. Les intervalles de confiance empirique (quantiles à 2.5% et 97.5% de 1000 simulations MCMC) sont indiqués entre crochets.

L'analyse descriptive a mis en avant une hétérogénéité spatiale possible du modèle. Il serait possible de tester la présence de ce phénomène en incluant des indicatrices de zones géographiques dans le modèle, en autorisant le modèle à être hétéroscédastique (via le package *sphet*) ou en conduisant une analyse géographique pondérée (via le package *spgwr*).

7 Conclusion

Les modèles d'économétrie spatiale définissent un cadre cohérent (et paramétrique) pour modéliser tout type d'interactions entre agents économiques: zones géographiques mais également produits, entreprises ou individus. Ils reposent sur une définition *a priori* de relations de voisinage. Les principales critiques qui leur sont adressées sont leur manque de robustesse quant au choix de la matrice de voisinage et leur manque d'identification du processus générateur des données. Ces critiques nous semblent néanmoins exagérées. Comme pour tout travail empirique, des choix toujours discutables de spécification sont nécessaires. La force de ces modèles est de mettre en avant si un problème "spatial" se pose et sous quelle forme.

A contrario, estimer un modèle d'économétrie spatiale dès qu'on dispose de données "spatiales" n'est pas toujours nécessaire. Le raffinement méthodologique doit être mis en regard

de la complexité des nouveaux modèles, en termes d'interprétation. Dans notre exemple, prendre en compte l'autocorrélation spatiale pour modéliser le taux de chômage localisé est nécessaire d'après les tests statistiques. Cela corrige certaines interprétations erronées issues du modèle linéaire classique (MCO). Mais il faut justifier le choix du modèle spatial. Introduire une seule interaction endogène (modèle SAR) serait ainsi à la fois contraire à l'approche statistique, inutile et complexe. Il convient ici de privilégier un modèle spatial de Durbin (SDM) ou aux erreurs spatialement autocorrélées (modèle SEM). De plus, estimer ces modèles d'économétrie spatiale suppose de disposer de données exhaustives. Dans le cas général, ils ne sont donc pas adaptés aux données d'enquêtes.

La prise en compte de l'hétérogénéité spatiale a été brièvement abordée, notamment au travers des régressions localisées. Les enjeux théoriques de ces méthodes restent encore mal maîtrisés. Elles permettent néanmoins des approches descriptives, de définir de grands ensembles régionaux homogènes et des analyses complémentaires à des tests de rupture régionale.

Bibliographie indicative

Abreu Maria, Henri L.F. De Groot, et Raymond J.G.M. Florax. (2005) "Space and Growth: A Survey of Empirical Evidence and Methods." *Région et Développement*, 21.

Aldstadt, Jared, et Arthur Getis. (2006) "Using AMOEBA to Create a Spatial Weights Matrix and Identify Spatial Clusters." *Geographical Analysis,* 38, 327-343.

Anselin, Luc. (1988) "Spatial Econometrics: Methods and Models." Dordrecht: Kluwer Academic Publishers.

Anselin, Luc, et Daniel A. Griffith. (1988) "Do Spatial Effects Really Matter in Regression Analysis?" *Papers in Regional Science*, 65(1), 11-34.

Anselin, Luc, Anil K. Bera, Raymond Florax, et Mann J. Yoon. (1996) "Simple Diagnostic Tests for Spatial Dependance." *Regional Science and Urban Economics*, 26(1), 77-104.

Anselin, Luc. (2002) "Under the Hood Issues in the Specification and Interpretation of Spatial Regression Models." *Agricultural Economics*, 27, 247-267.

Blanc, Michel, et François Hild. (2008) "Analyse des Marchés Locaux du Travail: du Chômage à l'Emploi." *Economie et Statistique*, 415-416, 45-60.

Bhattacharjee, Arnabet Chris Jensen-Butler. (2013) "Estimation of the Spatial Weights Matrix Under Structural Constraints." *Regional Science and Urban Economics*, 43(4), 617-634.

Bivand, Roger S., Edzer Pebesma, et Virgilio Gómez-Rubio. (2013) "Applied Spatial Data Analysis with R" Second Edition, Springer.

Brunsdon, Chris, A. Stewart Fotheringham, et Martin E. Charlton. (1996) "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity." *Geographical Analysis*, 28(4), 281-298.

Corrado, Luisa, et Bernard Fingleton. (2012) "Where is The Economics in Spatial Econometrics?" *Journal of Regional Science,* 52(2), 210-239.

Dubin, Robin A. (1998) "Spatial Autocorrelation: A Primer" *Journal of Housing Economics*, 7, 304-327.

Elhorst J. Paul. (2010) "Applied Spatial Econometrics: Raising the Bar." *Spatial Economic Analysis,* 5(1), 9-28.

Elhorst J. Paul. (2013) "Spatial Econometrics From Cross-Sectional Data to Spatial Panels." Springer.

Fingleton, Bernard, et Julie Le Gallo. (2007) "Finite Sample Properties of Estimators of Spatial Models with Autoregressive, or Moving Average, Disturbances and System Feedback." *Annales d'Économie et de Statistique*, 87/88, 39-62.

Fingleton, Bernard, et Julie Le Gallo. (2008) "Estimating Spatial Models with Endogenous Variables, a Spatial Lag and Spatially Dependent Disturbances: Finite Sample Properties." *Papers in Regional Science*, 87(3), 319-339.

Fingleton, Bernard, et Julie Le Gallo. (2012) "Endogeneité et Autocorrélation Spatiale : Quelle Utilité pour le Modèle de Durbin ?" *Revue d'Économie Régionale & Urbaine*, 1.

Floch, Jean-Michel. (2012) "Détection des Disparités Socio-économiques - l'Apport de la Statistique Spatiale." Document de travail INSEE H2012/04.

Florax Raymond J.G.M., Hendrik Folmer, et Sergio J. Rey. (2003) "Specification Searches in Spatial Sconometrics: the Relevance of Hendry's Methodology." *Regional Science and Urban Economics*, 33(5), 557-579.

Gibbons, Stephen, et Henry G. Overman. (2012) "Mostly Pointless Spatial Econometrics?" *Journal of Regional Science*, 52(2), 172-191.

Griffith, Daniel A. (1996) "Some Guidelines for Specifying the Geographic Weights Matrix Contained in Spatial Statistical Models." in Arlinghaus S.L (ed). Practical Handbook of Spatial Statistics, CRC, Boca Raton.

Grislain-Letrémy, Céline, et Arthur Katossky. (2013) "Les Risques Industriels et le Prix des Logements." *Economie et Statistique*, 460-461, 79-106.

Guymarc, Gaël. (2015) "Analyse Econométrique des Migrations Résidentielles." Séminaire de Méthodologie Statistique du département des méthodes statistiques, Statistique et géographie.

Harris, Richard, John Moffat, et Victoria Kravtsova. (2011) "In Search of "W"." *Spatial Economic Analysis*, 6(3), 249-270.

Kelejian, Harry H., et Gianfranco Piras. (2014) "Estimation of Spatial Models with Endogenous Weighting Matrices, and an Application to a Demand Model for Cigarettes." *Regional Science and Urban Economics,* 46, 140-149.

Lee, Lung-Fei. (2004) "Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models." *Econometrica*, 72, 1899-1925.

LeSage, James P., et Kelleyils R. Pace. (2009) "Introduction to Spatial Econometrics." CRC Press Taylor & Francis Group.

LeSage, James P., et Kelley R. Pace. (2012) "The Biggest Myth in Spatial Econometrics." Mimeo.

Le Gallo, Julie. (2002) "Économétrie Spatiale : l'Autocorrélation Spatiale dans les Modèles de Régression Linéaire." *Economie & Prévision*, 155(4), 139-157.

Le Gallo, Julie. (2004) "Hétérogénéité Spatiale, Principes et Méthodes." *Economie & Prévision*, 162(1), 151-172.

Loonis, Vincent. (2012) "Non Réponse à l'Enquête Emploi et Modèles Probit Spatiaux." Septième colloque francophone sur les sondages, Rennes.

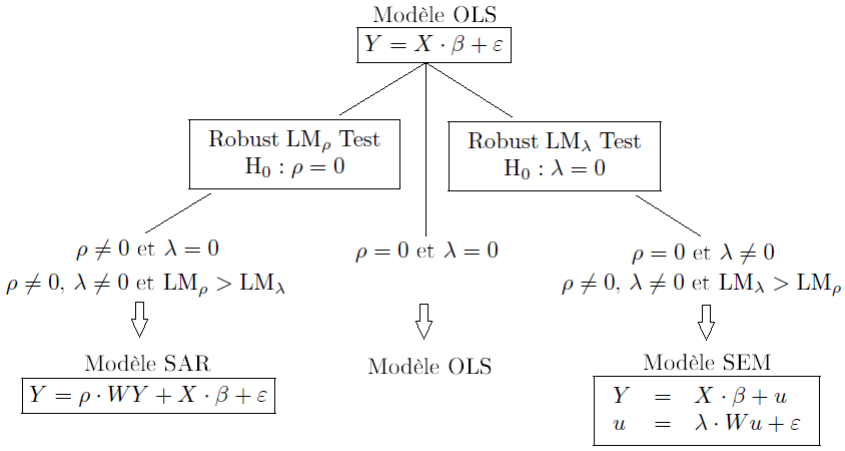
Lottmann Franziska. (2013) "Spatial Dependence in German Labor Markets." Thèse de l'Université de Humboldt.

Manski, Charles F. (1993) "Identification of Endogenous Social Effects: the Reflection Problem." *Review of Economic Studies*, 60, 531-542.

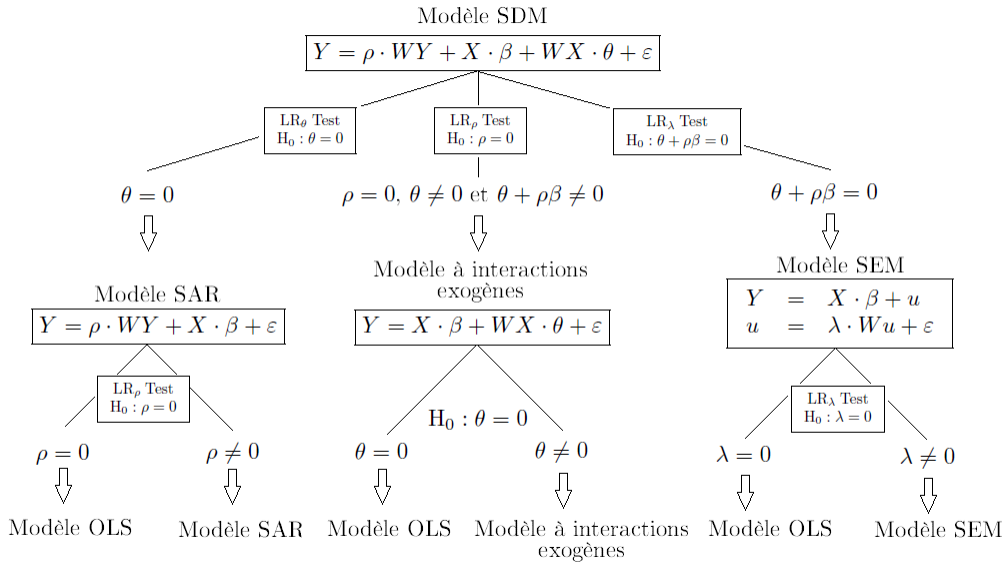
Annexes

Annexe 1 : Réprésentation graphique des procédures ascendante (bottom-up) et descendante (top-down)

Graphique A1.1 : Approche bottom-up, à partir de Florax et al. (2003)



Graphique A1.2 : Approche top-down, à partir de LeSage et Pace (2009)



Annexe 2: Codes R

Création d'une matrice de voisinage endogène, basée sur les déplacements domicile-travail

```
## Lecture du fichier SAS, des flux domicile-travail
library ("sas7bdat")
flux<-read.sas7bdat("flux.sas7bdat")
## Numérotation des zones
zeo<-unique(flux[,1])
zed<-unique(flux[,1])
lig<-c(rep(1:297))
col<-c(rep(1:297))
dzeo<-data.frame(zeo,lig)
dzed<-data.frame(zed,col)
flux$zeo<-flux$ZEMPL2010_RESID
flux$zed<-flux$ZEMPL2010_TRAV
flux<-merge(flux,dzeo,by="zeo")
flux<-merge(flux,dzed,by="zed")
## Construction de la matrice des poids
lien<-matrix(0,nrow=297,ncol=297)
for (i in 1:297)
{   for (j in 1:297)
                {ze<-flux$IPONDI[flux$lig==i & flux$col==j]
                                if(length(ze)>0)
                                lien[i,j]<-ze
                }
}
mig1.w<-mat2listw(lien,style="W")
```

Cartographie des zones d'emploi

```
### Cartographie des zones d'emploi
> vPal5 <- brewer.pal(n = 5, name = "Reds")
> carte@data <- data.frame(carte@data, donnees_ze[match(carte@data[,"Codgeo"],
donnees_ze[,"ze2010"]),])
> lJenks2011 <- classIntervals(var = carte@data$txcho_2011, n = 5, style = "jenks")
> vJenks2011 <- lJenks2011$brks
> carte@data$cho <- as.character(cut(carte@data$txcho_2011, breaks = vJenks2011,
labels = vPal5, include.lowest = TRUE, right = FALSE))
> vLegendBoxJ5 <- as.character(levels(cut(carte@data$txcho_2011, breaks = vJenks2011,
include.lowest = TRUE, right = FALSE)))
> plot(carte, col = carte@data$cho, border = "white")
> legend("bottomleft", legend = vLegendBoxJ5, bty = "n",
fill = vPal5, cex = 0.8, title = "Taux (%)")
> title(main="Taux de chômage (2011)")
```

Modèles linéaires spatiaux: estimations complémentaires

```
### Modèle SAC
> ze.sac<-sacsarlm(txcho_2013 ~ tx_act + part_act_peudip + part_act_1530
+ part_emp_ind + var_emp, data=donnees_ze,dist.w)
> summary(ze.sac)

### Modèle SLX
> ze.slx<-lmSLX(txcho_2013 ~ tx_act + part_act_peudip + part_act_1530
+ part_emp_ind + var_emp, data=donnees_ze,dist.w)
> summary(ze.slx)

### Modèle SDEM
> ze.sdem<-errorsarlm(txcho_2013 ~ tx_act + part_act_peudip + part_act_1530
+ part_emp_ind + var_emp, data=donnees_ze,dist.w,etype="emixed")
> summary(ze.sdem)

### Modèle Manski
> ze.manski<-sacsarlm(txcho_2013 ~ tx_act + part_act_peudip + part_act_1530
+ part_emp_ind + var_emp, data=donnees_ze,dist.w,type="sacmixed")
> summary(ze.manski)
```

Annexe 3: Modèle SDM, pour différentes matrices de voisinag

	(1) SDM Contiguité	(2) SDM 2 Voisins	(3) SDM 5 Voisins	(4) SDM 10 Voisins	(5) SDM Distance	(6) SDM Endogène
Taux d'activité	-0.448***	-0.455***	-0.471***	-0.454***	-0.433***	-0.432***
	(0.041)	(0.041)	(0.040)	(0.039)	(0.041)	(0.041)
% Actifs Peu Diplômés	0.177***	0.216***	0.185***	0.181***	0.179***	0.182***
	(0.028)	(0.029)	(0.029)	(0.027)	(0.028)	(0.028)
% Jeunes Actifs 15-30 ans	0.228***	0.249***	0.251***	0.244***	0.237***	0.217***
	(0.048)	(0.047)	(0.048)	(0.049)	(0.047)	(0.048)
% Emploi Industriel	-0.005	-0.011	-0.006	-0.006	-0.002	-0.008
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.011)
Variation de l'emploi	-0.016	0.007	-0.016	-0.012	-0.018	-0.024
	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.021)
$\hat{\rho}$	0.616***	0.463***	0.615***	0.659***	0.640***	0.600***
	(0.056)	(0.049)	(0.056)	(0.067)	(0.062)	(0.052)
$\hat{\theta}$, Taux d'activité	0.197***	0.148***	0.241***	0.234***	0.201***	0.120*
	(0.067)	(0.053)	(0.063)	(0.070)	(0.070)	(0.061)
$\hat{\theta}$, % Actifs Peu Diplômés	-0.122***	-0.133***	-0.127***	-0.141***	-0.136***	-0.127**
	(0.043)	(0.035)	(0.043)	(0.050)	(0.048)	(0.044)
$\hat{\theta}$, % Jeunes Actifs 15-30 ans	-0.107	-0.137**	-0.158**	-0.129*	-0.123*	-0.119*
	(0.068)	(0.054)	(0.065)	(0.076)	(0.071)	(0.066)
$\hat{\theta}$, % Emploi Industriel	-0.042**	-0.023*	-0.025	-0.035*	-0.040**	-0.035*
	(0.018)	(0.013)	(0.016)	(0.020)	(0.018)	(0.018)
$\hat{\theta}$, Variation de l'emploi	0.029	0.007	0.043	0.028	0.037	0.055
	(0.037)	(0.025)	(0.035)	(0.045)	(0.039)	(0.040)
Constante	19.217***	24.124***	17.893***	19.906***	17.962***	24.046***
	(4.055)	(3.368)	(3.869)	(4.507)	(4.266)	(4.032)
Observations	297	297	297	297	297	297
AIC	974	995	979	980	968	987

Note de lecture: Le modèle SDM est estimé avec 6 matrices de voisinage différentes. Les écarts-types sont indiqués entre parenthèses. Significativité: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.