6. Économétrie spatiale : modèles courants

JEAN-MICHEL FLOCH
Insee

RONAN LE SAOUT
Ensai

Résumé

Ce chapitre décrit la conduite d'une étude d'économétrie spatiale, en s'appuyant sur une modélisation descriptive du taux de chômage par zone d'emploi. Les modèles spatiaux ont néanmoins une application plus large, l'approche étant compatible avec tout problème où des relations de "voisinage" interviennent. La théorie économique caractérise en effet de nombreux cas d'interactions entre agents (produits, entreprises, individus), qui ne sont pas nécessairement de nature géographique. Le chapitre se concentre sur l'étude de la corrélation spatiale, et donc sur ces différentes interactions, et aborde les liens avec l'hétérogénéité spatiale, à savoir les 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 même définition préalable des relations de voisinage. Une méthodologie de choix de modèle (estimation et tests) est détaillée pas à pas. Des effets de rétroaction entraînent une interprétation particulière, et plus complexe, des résultats.

La lecture préalable des chapitre 1 : "Analyse spatiale descriptive", 2 : "Codifier la structure de voisinage" et 3 : "Indices d'autocorrélation spatiale" est recommandée.

Introduction

Les relations entre les valeurs observées sur des territoires proches préoccupent depuis longtemps les géographes. Waldo Tobler a résumé cette problématique par une formule souvent qualifiée de première loi de la géographie : "Tout interagit avec tout, mais deux objets proches ont plus de chance de le faire que deux objets éloignés". La disponibilité de données localisées, associée à des procédures de statistique spatiale désormais pré-programmées dans plusieurs logiciels statistiques, pose la question de la modélisation de cette proximité dans les études économiques. Une première étape reste bien sûr de caractériser cette proximité à l'aide d'indicateurs descriptifs et à l'aide 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 en interpréter les résultats ? Quelles en sont les limites ?

Nous illustrerons notre présentation par 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 œuvre. Nous rappellerons brièvement la définition d'une matrice de voisinage qui décrit les relations de proximité et les tests de corrélation spatiale (décrits plus en détail dans les chapitres 2 : "Codifier la structure de voisinage" et 4 : "Indices d'autocorrélation spatiale"). Nous détaillerons ensuite la spécification, l'estimation et l'interprétation de modèles d'économétrie spatiale.

Les techniques présentées s'appliquent à des domaines qui dépassent le cadre strictement géographique. Plusieurs types de données interconnectées, *i.e.* pouvant interagir entre elles, existent en effet : des points (individus ou entreprises dont on connaît l'adresse), des données par aires géographiques ou administratives (taux de chômage localisés), des réseaux physiques (routes) ou relationnels (élèves d'une même classe) ou des données continues (*i.e.* qui existent en tout point de l'espace). Ces dernières données sont essentiellement issues de la physique, par exemple la hauteur du sol, la température, la qualité de l'air, etc. et relèvent du domaine de la géostatistique (voir chapitre 5 : "Géostatistique"). Elles peuvent néanmoins servir de variables explicatives dans les modèles présentés dans ce document. Un point important à noter est qu'on considère ici des structures de proximité préexistantes, qui n'évoluent pas ou peu. On ne se pose ainsi pas la question de la caractérisation de la formation ou de l'évolution de ces relations de voisinage. On cherche au contraire 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 ? Si la majorité des applications ont une dimension géographique (ABREU

1. BLANC et al. 2008 traitent cette question de manière détaillée à l'aide d'un modèle d'économétrie spatiale pour la France, LOTTMANN 2013 pour l'Allemagne.

et al. 2004 pour la convergence des PIB régionaux, OSLAND 2010 pour les déterminants des prix de l'immobilier pour des exemples classiques), les domaines d'application sont ainsi plus vastes avec par exemple la mesure des effets de pairs dans les réseaux sociaux (FAFCHAMPS 2015 pour une synthèse), de la proximité idéologique en science politique (BECK et al. 2006) ou la prise en compte de la proximité entre produits pour étudier les effets de substitution en économie industrielle (SLADE 2005). Au sein de l'Insee, ces méthodes ont été utilisées pour étudier la relation entre les prix immobiliers et les risques industriels (GRISLAIN-LETRÉMY et al. 2013), les changements de lieux d'habitation ou la non-réponse dans l'enquête emploi (LOONIS 2012).

Des outils spécifiques ont été développés pour estimer les modèles d'économétrie spatiale. LESAGE et al. 2009 mettent à disposition des programmes MatLab. *GeoDa* est un logiciel libre d'analyse spatiale proposé dans le cadre d'un projet initié par Anselin en 2003 d'analyses spatiales. Il existe également des packages complémentaires pour Stata. 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. Les sections 6.1 et 6.2 présentent les raisons économiques et statistiques de la mise en place de ces modèles. La section 6.3 décrit les étapes de l'estimation d'un modèle d'économétrie spatiale. La section 6.4 traite de points techniques plus avancés. La section 6.5 détaille la mise en œuvre sous R à travers la modélisation du taux de chômage par zone d'emploi avant la conclusion. Les lecteurs intéressés par l'approfondissement de ces méthodes pourront notamment se référer à LESAGE et al. 2009, ARBIA 2014 ou LE GALLO 2002, LE GALLO 2004 pour une présentation en langue française.

6.1 Pourquoi tenir compte de la proximité spatiale, organisationnelle ou sociale ?

6.1.1 Les raisons économiques

L'interaction spatiale, organisationnelle ou sociale des agents économiques est classique en économie. ANSELIN 2002a liste ainsi les termes employés pour nommer ces interactions : effets de voisinage, de pair, interactions stratégiques, copie par mimétisme ou par les normes sociales ("*copy-catting*"), concurrence par comparaison ("*yardstick competition*"), etc. Il met notamment en avant deux situations de concurrence entre firmes justifiant le recours à un modèle spatial ou d'interaction.

Dans le premier cas, la décision d'un agent économique (une entreprise par exemple) dépend 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 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 $q_i = R(q_{-i}, x_i)$.

Dans le second cas, la décision d'un agent économique dépend d'une ressource rare. En reprenant l'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 dépend alors des quantités consommées par les autres firmes et donc de leur production q_{-i}. On retrouve la fonction de réaction précédente.

Cet exemple met en évidence que le recours à un modèle d'interaction est microfondé et que la notion de voisinage n'est pas forcément spatiale. Selon les secteurs industriels, les concurrents d'une entreprise seront ceux proches en termes de distance (les services à la personne, les supermarchés) ou de produits vendus (Coca-Cola et Pepsi). ANSELIN 2002a souligne que ces deux situations amènent à implémenter un même modèle spatial ou d'interaction. Ils sont équivalents d'un point de

vue observationnel. Les processus générateurs des données (PGD) sont différents mais fournissent les mêmes observations. De simples données en coupe ne permettent pas d'identifier la source de l'interaction (une concurrence stratégique par les quantités ou une concurrence sur les ressources dans notre exemple), mais seulement de confirmer sa présence et d'évaluer sa force. À l'instar de l'économétrie classique, il reste nécessaire de réfléchir aux effets identifiés par le modèle et les données.

De plus, les externalités ou effets de voisinages sont couramment contrôlés à l'aide de variables spatiales du type distance (par exemple au plus proche concurrent), ou d'indicateurs agrégés par zone géographique (par exemple le nombre de concurrents). Ce type de variable peut s'interpréter comme des variables spatialement décalées (*i.e.* fonction des observations dans les zones voisines), avec une définition *a priori* de relations de voisinage. L'économétrie spatiale justifie et généralise ainsi ces choix empiriques.

6.1.2 Les raisons économétriques

Les raisons économétriques renvoient aux insuffisances de la modélisation linéaire classique (et de l'estimation associée par la méthode des Moindres Carrés Ordinaire -MCO-) lorsque les hypothèses nécessaires à sa mise en œuvre ne sont plus vérifiées. LESAGE et al. 2009 présentent ainsi plusieurs arguments techniques justifiant l'emploi de méthodes spatiales. On observe fréquemment avec des données spatiales 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 (les estimateurs seront sans biais mais moins précis, et les tests n'auront plus les propriétés statistiques usuelles), 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. De plus, la confrontation de plusieurs modèles d'économétrie spatiale permet de discuter l'incertitude du processus générateur des données, qui n'est jamais connu, et de vérifier ainsi la robustesse des résultats.

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. Dans les études appliquées, il est parfois difficile de lier les aspects économétriques et économiques justifiant de la prise en compte de la dépendance spatiale et les causalités de nature économique sont difficiles à établir à partir de modèles économétriques spatiaux (GIBBONS et al. 2012).

6.2 Autocorrélation, hétérogénéité, pondérations : quelques rappels de statistique spatiale

6.2.1 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. ANSELIN et al. 1988, 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 ("*spillover*") 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. Ce chapitre traite principalement des méthodes de prise en compte de cette corrélation spatiale dans les modèles de régression, détaillés en section 6.3. L'hétérogénéité spatiale renvoie quant à elle à des phénomènes d'instabilité structurelle dans l'espace. Cette autre forme de prise en compte de l'espace est détaillée dans le chapitre 9 : "Régression géographiquement pondérée". Elle part de l'idée que 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'immobilier 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 partage "pédagogique" entre autocorrélation et hétérogénéité ne doit pas faire oublier les interactions entre les deux (ANSELIN et al. 1988 ; LE GALLO 2002 ; LE GALLO 2004). Il n'est pas toujours facile de distinguer chacune des 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 fait que l'information apportée par les données est moins riche que dans le cas où les données sont indépendantes. En cas d'autocorrélation, on observe une seule réalisation du processus générateur des données. Tout ceci plaide pour une approche exploratoire préalable des données. Selon la question, la méthodologie traitera en premier lieu l'autocorrélation spatiale des observations (*i.e.* les liens entre les unités proches) ou l'hétérogénéité des comportements (*i.e.* leur variabilité selon la localisation).

6.2.2 La matrice des poids

Pour mesurer la corrélation spatiale entre agents ou zones géographiques, tout commence par la définition *a priori* des 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 carrée de taille $N \times N$, dite matrice de voisinage et notée W, dont les éléments diagonaux sont nuls (on ne peut pas être son propre voisin). La valeur des éléments non diagonaux est le fruit de l'expertise. De nombreuses matrices de voisinage ont été proposées dans la littérature. Leur construction avec le logiciel R est détaillée dans le chapitre 2 : "Codifier la structure de voisinage".

6.2.3 Les méthodes exploratoires

Avant de spécifier 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 décrits dans le chapitre 3 : "Indices d'autocorrélation spatiale".

Le principal indicateur[2] est celui de Moran qui mesure l'association globale : $I = \frac{N}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij}(y_i-\bar{y})(y_j-\bar{y})}{\sum_i (y_i-\bar{y})^2}$, avec w_{ij} le poids correspondant au coefficient situé sur la i-ème ligne et la j-ème colonne de la matrice de voisinage W. Les bornes de l'indicateur de Moran I sont comprises entre -1 et 1 et dépendent de la matrice de poids utilisée. La borne supérieure est notamment égale à 1 si la matrice est standardisée en ligne, la borne inférieure reste différente en toute généralité de -1. Une corrélation positive signifie que les zones avec de hautes ou de basses valeurs pour y se regroupent, une corrélation négative que des zones géographiques proches ont des valeurs de y très différentes. Sous l'hypothèse H_0 d'absence d'autocorrélation spatiale ($I = 0$), la

2. Les indicateurs de Geary et de Getis et Ord, ainsi que les autres indicateurs locaux, sont présentés dans FLOCH 2012.

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, le rejet de H_0 ne signifie pas qu'un modèle d'économétrie spatiale soit nécessaire mais que celui-ci doive être envisagé. Il peut en effet ne refléter que la répartition spatiale d'une variable sous-jacente. Par exemple, si le modèle sous-jacent est $Y = X \cdot \beta + \varepsilon$ avec β un paramètre à estimer, $\varepsilon \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma^2\right)$ et X une variable autocorrélée spatialement, un test de Moran conclura à l'autocorrélation spatiale de la variable Y. Pour autant, le modèle linéaire liant Y et X n'est pas un modèle spatial, et peut être estimé classiquement à l'aide des MCO.

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. Leur calcul est détaillé dans le chapitre 3 : "Indices d'autocorrélation spatiale".

6.3 Estimer un modèle d'économétrie spatiale

6.3.1 La galaxie des modèles d'économétrie spatiale

ELHORST 2010 a établi une classification des principaux modèles d'économétrie spatiale, en s'appuyant sur les trois types d'interaction spatiale issus du modèle fondateur de MANSKI 1993 :

— une interaction endogène, lorsque la décision économique d'un agent ou d'une zone géographique va dépendre de la décision de ses voisins ;
— une interaction exogène, lorsque 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 [3] :

$$\begin{aligned} Y &= \rho \cdot WY + X \cdot \beta + WX \cdot \theta + u \\ u &= \lambda \cdot Wu + \varepsilon \end{aligned} \tag{6.1}$$

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 égale au nombre de variables exogènes K) et λ pour l'effet de corrélation spatiale des erreurs dit autocorrélation spatiale. Dans la suite du document, nous emploierons le terme de corrélation spatiale pour désigner un de ces 3 types d'interaction spatiale.

Le modèle de MANSKI 1993 n'est pas identifiable sous cette forme, c'est-à-dire qu'on ne peut pas estimer à la fois β, ρ, θ, et λ. Prenons son exemple des effets de pairs pour en donner l'intuition. Supposons que les mauvais résultats scolaires d'une classe s'expliquent par la composition sociale

3. 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 contiguïté, $W \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ représente le nombre de voisins de chaque observation. Si ce nombre de voisins est le même pour tous les individus, la constante β_0 et le terme $W \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \cdot \theta_0$ ne sont pas identifiables séparément. De plus, le nombre de voisins (ou le nombre moyen si la matrice de voisinage est normée par ligne) 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.

de la classe (interaction exogène) et le fait d'avoir de mauvais professeurs (caractéristique inobservée). On constatera alors une forte corrélation des résultats des élèves au sein de la classe mais cela ne signifie pas que le fait d'être avec des élèves d'un niveau scolaire plus faible (interaction endogène) a un effet.

Une première solution, pour rendre le modèle identifiable, est de supposer que les matrices de voisinage W ne sont pas identiques pour les trois 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. SLADE 2005 définit ainsi deux matrices de voisinage distinctes pour étudier les effets prix en économie industrielle : W_ρ étant fonction de la distance entre entreprises concurrentes et W_X d'un indicateur de proximité entre les produits vendus. Une autre solution consiste à supprimer l'une des 3 formes de corrélation spatiale, représentées par les paramètres 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 $I-\lambda W$, et l'identification des modèles. On peut retenir que les matrices usuelles de contiguïté ou de distance inverse respectent ces contraintes. Ce n'est pas forcément le cas de matrices "atypiques" créées par exemple pour les relations de proximité sociale. Il n'est par exemple pas possible d'avoir uniquement des îles (une zone qui n'a pas de voisin) ou au contraire que tout le monde soit le voisin de tout le monde. On doit de plus supposer que $|\rho|<1$ et $|\lambda|<1$ (critères qu'on peut intuitivement rapprocher des conditions de stationnarité pour les solutions d'un modèle de type ARMA).

Trois principaux types de modèles peuvent être déduits du modèle de MANSKI 1993 selon la contrainte utilisée, $\theta=0$, $\lambda=0$ ou $\rho=0$.
Le cas $\rho=0$ (modèle SDEM, *Spatial Durbin Error Model*) peut être envisagé si on suppose qu'il n'y a pas d'interaction endogène et que l'accent est mis sur les externalités de voisinage. Ce modèle reste néanmoins d'un usage moins courant (LESAGE 2014).

Si on suppose que le modèle est tel que $\theta=0$, on trouve le modèle de Kelejian-Prucha (ou également nommé SAC, *Spatial Autoregressive Confused*, KELEJIAN et al. 2010a pour le modèle hétéroscédastique) :

$$\begin{aligned} Y &= \rho \cdot WY + X \cdot \beta + u \\ u &= \lambda \cdot Wu + \varepsilon \end{aligned} \tag{6.2}$$

Les estimateurs de β du modèle de Kelejian-Prucha présentent le défaut d'être biaisés et non convergents dans le cas où le vrai modèle inclut des interactions exogènes WX (LESAGE et al. 2009). Il y a en effet dans ce cas biais de variables omises. De plus, LE GALLO 2002 souligne que choisir une même matrice de voisinage W pour ce modèle engendre une identification faible des paramètres.

Au contraire, si on suppose que le modèle est tel que $\lambda=0$, $Y=\rho\cdot WY+X\cdot\beta+WX\cdot\theta+\varepsilon$, dit modèle spatial de Durbin (SDM, *Spatial Durbin Model*) , alors les estimateurs seront non biaisés (et les statistiques de test valides) même si, en réalité, nous sommes en présence d'erreurs autocorrélées spatialement (SEM). Ce modèle est ainsi plus robuste à un mauvais choix de spécification.

Ces deux modèles (Kelejian-Prucha et SDM) incluent les cas particuliers du modèle spatial autorégressif (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 $\theta\neq 0$.

Il existe par ailleurs des versions plus générales de ces modèles, qui autorisent une variation des effets de voisinage selon l'ordre de voisinage ou selon les interactions prises en compte. Ils correspondent à des versions spatiales des modèles temporels *ARMA(p,q)*.

Dans le cadre d'une étude économique, on ne présente pas l'ensemble de ces modèles. Les critères statistiques et la cohérence avec la question économique permettent de retenir une spécification plutôt qu'une autre.

6.3.2 Critères statistiques du choix de modèle

Deux approches principales ont été utilisées pour le choix des modèles. Ces approches "pratiques" reposent sur l'hypothèse que la matrice de voisinage soit connue et que les variables explicatives soient exogènes. Sous l'hypothèse de normalité des résidus $\varepsilon \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma^2\right)$, elles reposent sur une estimation par maximum de vraisemblance des modèles et les tests statistiques associés [4]. La première dite "approche ascendante" ou *bottom-up* (figure 6.1) consiste à commencer avec le 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 ont 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.

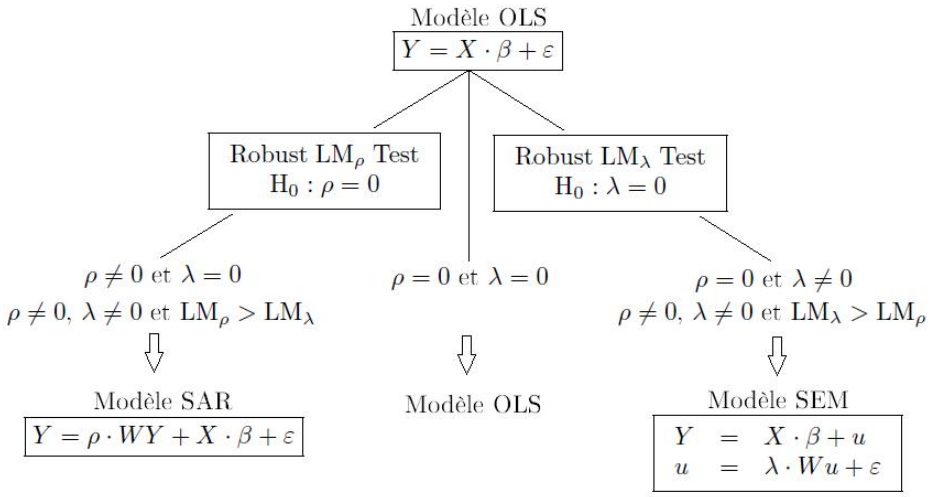

FIGURE 6.1 – Approche *bottom-up*
Source : FLORAX *et al. 2003*

La deuxième approche dite "approche descendante" ou *top-down* (figure 6.2) consiste à commencer avec le modèle spatial de Durbin. À partir des tests du rapport de vraisemblance, on en déduit le modèle le plus adapté aux observations. L'amélioration des performances informatiques

4. D'autres méthodes d'estimation existent. Dans le cas de variables explicatives endogènes, FINGLETON et al. 2008 FINGLETON et al. 2012 proposent une estimation par variables instrumentales et la méthode des moments généralisée. LESAGE et al. 2009 proposent une estimation bayésienne. Enfin, pour relâcher le cadre paramétrique, LEE 2004 propose une estimation par quasi maximum de vraisemblance.

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 al. 2009.

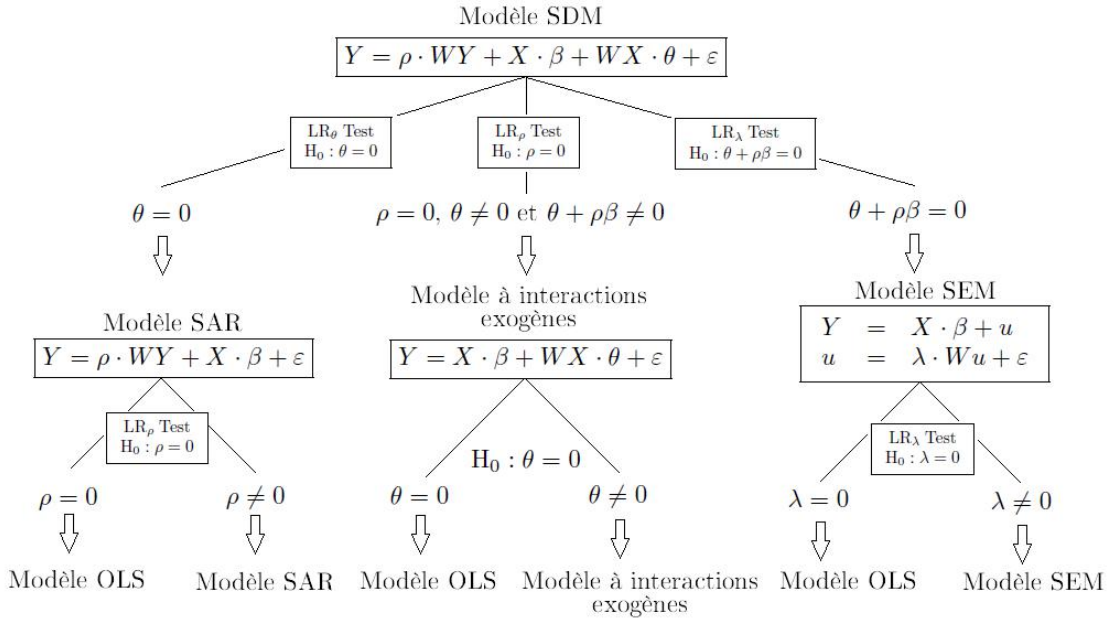

FIGURE 6.2 – Approche *top-down*
Source : LESAGE *et al. 2009*

ELHORST 2010 propose une approche "mixte" représentée en figure 6.3. 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, à étudier 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)[5] concluent à la présence d'un terme autorégressif, *i.e.* $\rho \neq 0$ et $\lambda = 0$ (branche de gauche de la figure 6.1). On estime alors le modèle SDM. À 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 présence d'autocorrélation résiduelle, *i.e.* $\rho = 0$ et $\lambda \neq 0$ (branche de droite de la figure 6.2), on se ramène au modèle SDM ($\rho \neq 0$ et $\theta \neq 0$), puis 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 de corrélation spatiale, *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 et SDM. Enfin, dans le cas où les tests concluent à la présence simultanée de corrélation endogène et résiduelle, *i.e.* $\rho \neq 0$ et $\lambda \neq 0$, le modèle SDM est estimé.

La matrice de voisinage W a pour dimension le carré du nombre d'observations. Or le calcul de la vraisemblance de ces modèles spatiaux fait notamment intervenir des déterminants incluant cette matrice. Le coût computationnel peut donc être important lorsque le nombre d'observations devient élevé. LESAGE et al. 2009 consacrent ainsi un chapitre aux enjeux computationnels (et

5. Il existe deux versions de ces tests, l'une robuste à la présence d'autres formes de corrélation spatiale, l'autre non (ANSELIN et al. 1996).

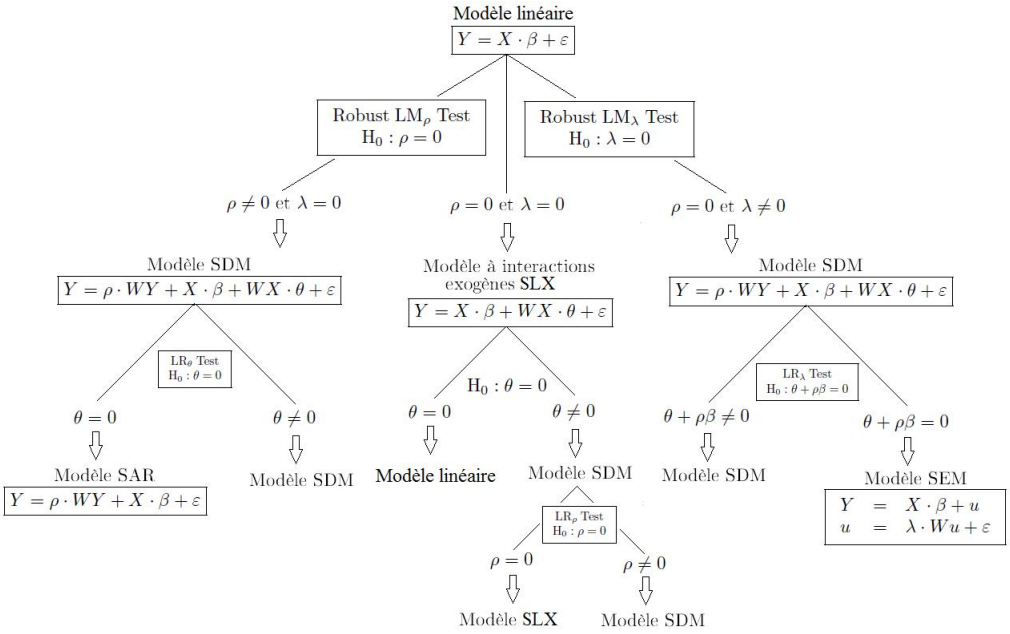

FIGURE 6.3 – Approche d'ELHORST 2010 pour le choix d'un modèle d'économétrie spatiale
Source : ELHORST 2010.

aux méthodes pour les résoudre) associés à l'estimation de ces modèles. En pratique, le nombre d'observations est souvent limité à quelques milliers.

Ces règles ne doivent pas être considérées comme intangibles [6], 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 justifient.

6.3.3 L'interprétation des résultats : attention aux rétroactions

L'économétrie spatiale s'écarte du cadre habituel des modèles linéaires lorsque des variables spatialement décalées WY sont présentes dans le modèle. L'interprétation classique des modèles linéaires reste en revanche valide si seule une autocorrélation spatiale des erreurs est prise en compte (modèle SEM).

En présence d'une variable spatialement décalée WY les paramètres associés aux variables explicatives ne s'interprètent pas comme dans le cadre habituel des modèles linéaires. En effet, du fait des interactions spatiales, la variation d'une variable explicative pour une zone donnée affecte directement son résultat et indirectement les résultats de toutes les autres zones. Les paramètres estimés interviennent alors dans le calcul d'un effet multiplicateur qui est global car il affecte l'ensemble de l'échantillon.

En revanche, l'interprétation des paramètres associés aux variables explicatives reste identique lorsque le modèle ne comporte qu'une autocorrélation des erreurs (modèle SEM). Dans ce cas, il existe un effet de diffusion global lié aux erreurs autocorrélées spatialement : la variation d'une variable explicative pour une zone donnée affecte directement son résultat et indirectement les résultats de toutes les autres zones, mais sans que la valeur de cet effet soit démultipliée.

Lorsqu'on considère des modèles avec variables explicatives spatialement décalées (SLX) les paramètres associés aux variables explicatives permettent de calculer un effet local dans la mesure

6. L'approche séquentielle des tests peut de plus engendrer un biais car la zone de rejet des tests du rapport de vraisemblance (LR) devrait en théorie tenir compte des tests préalables du multiplicateur de Lagrange (LM).

où la variation d'une variable explicative affecte directement son résultat et indirectement le résultat des zones voisines, mais pas celui des zones voisines de ces voisines.

Pour formaliser les différents impacts, nous reprenons le cadre défini par LESAGE et al. 2009.

Le modèle SAR est $Y = \rho \cdot WY + X\beta + \varepsilon$. Il peut se réécrire de plusieurs manières, en notant r l'indice pour une variable explicative et S_r des matrices carrées de la taille du nombre d'observations :

$$\begin{aligned} Y &= (1-\rho W)^{-1} X\beta + (1-\rho W)^{-1} \varepsilon \\ &= \sum_{r=1}^{k} (1-\rho W)^{-1} \beta_r X_r + (1-\rho W)^{-1} \varepsilon \\ &= \sum_{r=1}^{k} S_r(W) X_r + (1-\rho W)^{-1} \varepsilon \end{aligned} \tag{6.3}$$

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

La valeur prédite est donc $\hat{y} = (1-\hat{\rho}W)^{-1} X\hat{\beta}$ [7] et non $X\hat{\beta}$ comme dans un modèle linéaire 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_r 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. À 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 apparaît différent pour chaque zone [8]. 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. Pour l'ensemble des zones (niveau global) on peut donc calculer des effets directs et indirects obtenus en faisant la moyenne de ces effets (LESAGE et al. 2009) :

— L'effet direct moyen correspond à la moyenne des termes diagonaux de la matrice S_r, *i.e.* $\frac{1}{n}\text{trace}(S_r)$. L'interprétation de cet indicateur se rapproche de celle des coefficients β d'un modèle linéaire non spatial calculés par la méthode des MCO.

— L'effet total moyen correspond à une moyenne de l'ensemble des termes de la matrice S_r, $\frac{1}{n}\sum_i [\sum_k S_r(W)_{ik}]$. Il peut s'interpréter de deux manières, soit comme la moyenne des n effets sur une zone i d'une modification d'une unité de la variable X_r dans toutes les zones, *i.e.* $\sum_k S_r(W)_{ik}$ (la somme des termes en ligne de la matrice S_r), soit comme la moyenne des n effets d'une modification d'une unité de la variable X_r dans une zone i sur l'ensemble des zones, *i.e.* $\sum_k S_r(W)_{ki}$ (la somme des termes en colonne de la matrice S_r).

— L'effet indirect moyen est la différence entre l'effet total moyen et l'effet direct moyen.

7. Ce n'est pas la prédiction optimale, voir THOMAS-AGNAN et al. 2014 pour la prédiction optimale d'un modèle SAR.

8. On retrouve cette caractéristique pour l'effet marginal d'un modèle Probit par exemple. Le modèle est $\mathbb{E}(Y|X) = \mathbb{P}(Y=1|X) = \Phi(\beta X)$ avec Φ la fonction de répartition d'une loi normale centrée-réduite. L'effet marginal d'une variable X_r est alors $\beta_r \cdot \varphi(\beta X)$ et diffère donc pour chaque individu. Une solution est alors d'estimer l'effet marginal moyen $\beta_r \cdot \overline{\varphi(\beta X)}$.

Les indicateurs sont identiques pour le modèle de Kelejian-Prucha. De tels indicateurs peuvent être définis pour le modèle SDM $Y = \rho \cdot WY + X\beta + WX \cdot \theta + \varepsilon$, mais leurs calculs doivent tenir compte des interactions exogènes $WX \cdot \theta$. La matrice $S_r(W)$ s'écrit en effet dans ce cas $(1-\rho W)^{-1}(I_n\beta_r + W\theta_r)$, au lieu de $(1-\rho W)^{-1}\beta_r$ dans le cas du modèle SAR.

Lorsqu'une interaction exogène $WX \cdot \theta$ est présente mais qu'il n'y a pas d'interaction endogène (modèles SLX et SDEM), l'effet direct d'une variable X_r est β_r, l'effet indirect est θ_r.

Dans tous les cas, le calcul de la précision de ces estimateurs est complexe. Pour ce calcul, LESAGE et al. 2009 s'appuient ainsi sur des simulations bayésiennes de Monte-Carlo par Chaîne de Markov (MCMC) [9].

Par ailleurs, ces effets dépendent en premier lieu du voisinage proche. Pour le modèle SAR, on peut noter que l'effet direct moyen est supérieur en valeur absolue à l'effet marginal du modèle linéaire non spatial, $|S_r| > |\beta_r|$. Les termes diagonaux de la matrice de voisinage W sont en effet nuls. La décomposition en séries entières $(1-\rho W)^{-1} = \left(I_n + \rho W + \rho^2 W^2 + \cdots\right)$ montre que le premier terme de rétroaction (et qui domine les autres termes d'ordre supérieur) est proportionnel à ρ^2. L'analyse des effets par ordre de voisinage (distinguer l'effet direct, l'effet des voisins, des voisins des voisins, etc.) est également developpée par LESAGE et al. 2009.

En conclusion, pour l'interprétation globale d'un modèle avec interaction endogène, il est utile de calculer pour chaque variable, l'effet direct moyen ($\frac{1}{n}\text{trace}(S_r)$) et l'effet indirect moyen ($\frac{1}{n}\left[\sum_j \sum_k S_r(W)_{kj} - \text{trace}(S_r)\right]$). Calculer l'effet induit par l'espace ($\frac{1}{n}\text{trace}(S_r) - \hat{\beta}_r$) permet également d'illustrer la force des effets de rétroaction.

6.4 Limites et difficultés économétriques

6.4.1 Que faire des données manquantes ?

En économétrie classique, on observe un échantillon de n individus. Si quelques individus présentent des valeurs manquantes, ils sont généralement exclus de l'analyse. En l'absence de sélection liée à la non-réponse (le processus de non-réponse est indépendant des variables de notre modèle), cela réduit la taille de l'échantillon mais n'empêche pas la mise en œuvre 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 consiste à interpoler les valeurs manquantes à l'aide de techniques de géostatistique (ANSELIN 2001), mais cela a pour incidence de mesurer les variables avec erreurs [10], ou d'utiliser une estimation adaptée (par exemple algorithme EM espérance-maximisation, WANG et al. 2013b pour le modèle SAR). Ces solutions ne sont néanmoins possibles que pour un faible pourcentage de valeurs manquantes.

Une autre implication est qu'il n'est pas aisé de mettre en place ces techniques sur données individuelles d'enquête. Dans le cas général, l'économétrie spatiale n'est pas adaptée aux données

9. 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 échantillonner.

10. L'interpolation peut également être utile lorsque les niveaux géographiques servant à mesurer la variable à expliquer et les variables explicatives sont différentes, par exemple les prix de logements connus au niveau de l'adresse ou de la commune et des indicateurs de pollution atmosphérique mesurés à l'aide de capteurs dont les localisations diffèrent.

d'enquêtes. En effet, dans ce cas on n'observe que des relations de voisinage partielles, pour les seuls individus enquêtés. Il faut alors faire l'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. Dans le chapitre 11 "Économétrie spatiale sur données d'enquêtes", Lardeux et Merly-Alpa montrent qu'il n'est possible de détecter la corrélation spatiale générée par un modèle SAR uniquement pour un plan de sondage par grappes géographiques. Avec de faibles taux de sondage et des plans de sondages classiques (stratifiés ou systématiques), seuls les effets directs peuvent être estimés. Ce point est développé dans le chapitre 11 : "économétrie spatiale sur données d'enquête".

6.4.2 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. De nombreux auteurs soulignent la sensibilité des résultats au choix de cette matrice (CORRADO et al. 2012 ; HARRIS et al. 2011), alors que LESAGE et al. 2009 considèrent que ces conclusions proviennent d'une mauvaise interprétation des modèles et que cette sensibilité supposée à la matrice de poids est "le plus grand mythe" de l'économétrie spatiale. 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 et al. 2011 : "L'économétrie spatiale souligne l'importance du choix de la matrice W mais nous renseigne peu sur les critères pour effectuer ce choix", difficultés qui ont contribué au scepticisme de plusieurs économistes (GIBBONS et al. 2012). Ces considérations montrent la complexité de la détermination de la matrice W qui reste un sujet de controverses scientifiques.

On a vu que les modèles traitent en général la matrice W comme exogène. D'autres méthodes s'appuient néanmoins sur les données utilisées pour déterminer la matrice des poids. ALDSTADT et al. 2006 définissent ainsi un algorithme de construction de la matrice W à partir des indicateurs locaux d'autocorrélation spatiale des variables d'intérêt. Il est également possible d'estimer les poids à partir de modèles économétriques avec des contraintes fonctionnelles *a priori* faibles (BHATTACHARJEE et al. 2013). Ces dernières approches sont souvent lourdes en calcul et plus difficiles à implémenter. De plus, une description plus réaliste et plus conforme à la réalité économique risque d'introduire de l'endogénéité. Des travaux faisant intervenir des matrices endogènes ont été récemment proposés (KELEJIAN et al. 2014).

Dernier point, la matrice W est considérée fixe, ce qui contraint le cadre de l'analyse économique. Par exemple, dans le cas de matrice de voisinage mesurant la distance entre entreprises ou produits, WAELBROECK 2005 souligne que l'arrivée (ou le départ) d'une entreprise ou d'un produit est un évènement endogène qui devrait amener à modifier les relations de voisinages, ce que ne permet pas la méthodologie usuelle.

6.4.3 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 individuelle. Une autocorrélation spatiale des erreurs $(I-\lambda W)^{-1}\varepsilon$ (modèle SEM) peut s'interpréter comme un effet aléatoire spatial (on suppose que les effets individuels au sein d'un voisinage sont proches, faute de pouvoir estimer des effets fixes) et donc une forme particulière d'hétéroscédasticité et de corrélation spatiale (LESAGE et al. 2009). Une solution alternative à un modèle d'économétrie spatiale serait de définir la forme de l'hétéroscédasticité et de la corrélation spatiale de la matrice de variances-covariances (DUBIN 1998), de définir des clusters spatiaux (BARRIOS et al. 2012) ou d'adopter une correction spatiale du type Newey-West (FLACHAIRE 2005). Enfin, des développements récents de l'économétrie spatiale

relâchent l'hypothèse d'homoscédasticité des résidus ε des modèles présentés dans cette introduction. KELEJIAN et al. 2007 KELEJIAN et al. 2010b ont ainsi proposé une méthode paramétrique de type HAC (*Heteroscedasticity and Autocorrelation Consistent*), issue des séries temporelles, et une méthode non paramétrique.

En présence d'hétéroscédasticité, les estimateurs restent convergents mais les statistiques de tests ne sont plus distribuées selon les lois usuelles. Les tests d'autocorrélation spatiale ne sont donc plus fiables. *A contrario*, en présence d'autocorrélation spatiale, les tests d'hétéroscédasticité usuels (*White*, *Breusch-Pagan*) ne sont également plus valables. LE GALLO 2004 présente des tests joints d'hétéroscédasticité et d'autocorrélation spatiales.

La seconde forme d'hétérogénéité correspond à la variabilité spatiale des paramètres ou de la forme fonctionnelle du modèle. Lorsque l'on connait bien le territoire d'intérêt, elle est souvent traitée dans la littérature empirique en ajoutant des indicatrices de zones géographiques dans le modèle (éventuellement croisées avec chaque variable explicative), en estimant le modèle pour différentes zones ou en conduisant des tests de stabilité géographique des paramètres (dits de Chow). Lorsque le nombre de ces zones géographiques augmente, ce traitement diminue néanmoins le nombre de degrés de liberté et donc la précision des estimateurs. Des méthodes plus complexes couramment utilisées en géographie ont été développées (LE GALLO 2004). Elles restent en grande partie descriptives et exploratoires (notamment à travers des représentations graphiques), car leurs propriétés théoriques sont partiellement connues et tout particulièrement ce qui concerne les propriétés de convergence et la prise en compte des ruptures.

Il existe également des méthodes de lissage géographique où la constante (voire chaque variable explicative) est croisée avec des polynômes fonctions des coordonnées géographiques. FLACHAIRE 2005 propose un modèle linéaire partiel (et alternatif) $Y_i = X_i\beta + f(u_i, v_i) + \varepsilon_i$ où f désigne une forme fonctionnelle dépendant des coordonnées géographiques u_i et v_i (voire d'autres variables explicatives si la proximité n'est pas spatiale mais sociale ou entre produits par exemple). Il montre qu'à l'instar d'un modèle SAR, la fonction f peut s'interpréter comme une somme pondérée des variables endogènes Y. Cette analyse met ainsi en avant que corrélation et hétérogénéité spatiales sont liées.

Il existe également des méthodes de régression locale dont l'extension au contexte spatial est formalisée dans le cadre de la régression géographiquement pondérée (BRUNSDON et al. 1996). Ces méthodes sont détaillées dans le chapitre 9 : "Régression géographiquement pondérée".

Il reste néanmoins délicat de distinguer hétérogénéité et corrélation spatiales. Il n'y a pas à notre connaissance de méthode identifiant de manière distincte ces deux phénomènes. Des approches pragmatiques sont donc retenues. LE GALLO 2004 propose une application sur la criminalité aux états-Unis. À l'aide de tests d'hétéroscédasticité (robustes à la présence d'autocorrélation), elle met en avant la présence de régimes spatiaux distincts entre deux zones géographiques, Est et Ouest. Un modèle SAR est ensuite estimé, pour lequel les variables explicatives X sont croisées avec les deux régimes spatiaux et les variances sont supposées différentes entre ces deux zones. OSLAND 2010 étudie les prix de l'immobilier en Norvège à l'aide de modèles d'économétrie spatiale, de lissage semi-paramétrique et de régressions géographiques pondérées. Les différentes approches donnent des résultats complémentaires mais ne sont pas intégrées dans une modélisation unique.

6.4.4 Le risque d'erreur "écologique"

Les méthodes présentées dans ce document s'appuient sur des zonages géographiques prédéfinis (une zone d'emploi dans notre exemple). De nombreuses variables économiques ne sont disponibles que pour les divisions administratives du territoire (région, département, canton). Or ce découpage administratif ne correspond pas forcément à la réalité économique des relations entre agents. Ce phénomène géographique est connu sous l'acronyme MAUP (*Modifiable Areal Unit Problem*).

Il a plusieurs conséquences (FLOCH 2012). Avec des échelles ou des découpages différents, les résultats des modèles et les interactions entre agents ne sont pas identiques. Il faut également tenir compte de l'étendue spatiale des zones : 1 000 agents économiques n'interagissent pas de la même manière dans 1 km^2 ou dans 10 000 km^2. Lorsque des données individuelles sont disponibles (par exemple les caractéristiques d'emploi issues du recensement de la population plutôt que les taux de chômage par zone d'emploi), il est possible de faire abstraction de ce découpage administratif ou de construire le niveau géographique *a priori* le plus pertinent. Mais en général, il n'y a pas de solution pour résoudre le problème du MAUP.

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, connue sous le nom de *régression écologique*, entraîne des problèmes d'identification et d'hétéroscédasticité. Anselin (2002) donne l'exemple d'un modèle où les décisions d'un individu i, y_{ik}, s'expliquent par ses caractéristiques x_{ik} mais également par les caractéristiques du groupe k auquel il appartient $\bar{x}_k = \sum_i x_{ik}/n_k$. Le modèle s'écrit $y_{ik} = \alpha + \beta \cdot x_{ik} + \gamma \cdot \bar{x}_k + \varepsilon_{ik}$ où β représente l'effet individuel et γ l'effet de contexte. Si on ne dispose que de données par groupe (par exemple les résultats moyens d'une classe à un examen et non les résultats individuels), le modèle estimé devient $\bar{y}_k = \alpha + (\beta + \gamma) \cdot \bar{x}_k + \bar{\varepsilon}_k$. Il n'est alors plus possible d'identifier séparément les paramètres β et γ. Le modèle est hétéroscédastique car $\mathbb{V}(\bar{\varepsilon}_k) = \sigma^2/n_k$ 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-groupe ne seront plus prises 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 contiguïté agrégée au niveau groupe, la force des relations individuelles ne sera plus prise 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'échelles 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 (GIVORD et al. 2016). Dans le cas d'études macro-économiques telles que la croissance régionale, ce problème est moins présent. Le niveau agrégé est en effet le niveau pertinent.

6.5 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'emploi, hors Corse) à l'aide de caractéristiques structurelles relatives aux caractéristiques de la population active (proportion des personnes peu

diplômées et des personnes de moins de 30 ans dans la population active), de la structure économique (proportion des emplois dans le secteur industriel et de l'emploi public) et du marché du travail (taux d'activité). L'objectif de cette partie n'est pas de détailler les résultats d'une étude économique mais d'illustrer les techniques mises en œuvre : la définition d'une matrice de voisinage qui décrit les relations de proximité entre territoires, 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 (BLANC et al. 2008 ; LOTTMANN 2013). Les variables économiques sont supposées structurelles et peu variables à court terme. Pour limiter les problèmes d'endogénéité, le taux de chômage est calculé sur l'année 2013 et les variables explicatives correspondent aux millésimes 2011 de la source 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, comme par exemple 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 exemples et les codes sont présentés à l'aide de R, logiciel le plus complet pour l'estimation de modèles d'économétrie spatiale. Nous listons ci-dessous quelques packages utiles dans R :

— *sp* et *rgdal* pour l'importation et 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* (spatial dependancies) pour l'ensemble des modèles classiques, et *spgwr* pour la régression géographique pondérée.

6.5.1 Cartographie et tests

Après avoir importé les données et défini une matrice de voisinage grâce aux méthodes présentées en section 6.2, on peut cartographier les données et effectuer une première analyse de l'autocorrélation spatiale.

La figure 6.4 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 contiguës 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.

La table 6.1 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. Hormis pour le pourcentage d'emploi industriel et d'emploi public, les écarts interquartiles sont faibles, inférieurs à 5%. Le pourcentage d'emploi industriel apparaît comme la variable la plus polarisée.

Tests d'autocorrélation spatiale et représentations graphiques avancées

La p-value quasiment nulle du test de Moran indique que l'hypothèse nulle d'absence d'autocorrélation spatiale doit être rejetée (voir chapitre 3 : "Indices d'autocorrélation spatiale"). Le résultat est robuste au choix de la matrice de voisinage.

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.

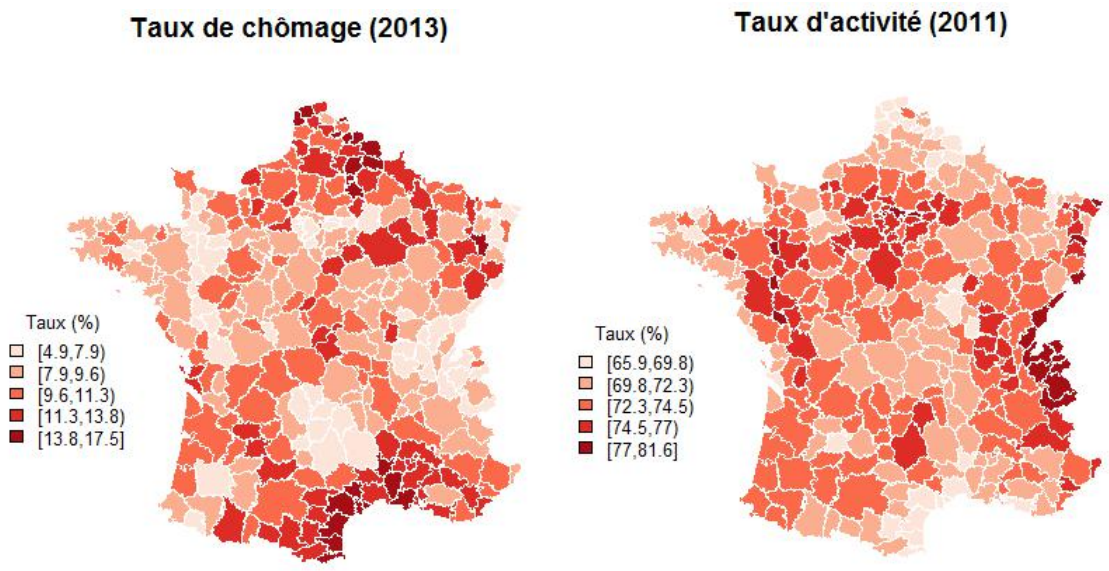

FIGURE 6.4 – Distribution du taux de chômage et d'activité, par zone d'emploi

	N	Moyenne	Écart-Type	Min	Q25	Médiane	Q75	Max
Taux de chômage (en %)	297	10.0	2.4	4.9	8.3	9.6	11.4	17.5
Taux d'activité (en %)	297	72.8	2.6	65.9	71.3	72.8	74.2	81.6
% Actifs Peu Diplômés	297	22.1	3.6	13.0	19.5	22.2	24.8	32.2
% Jeunes Actifs 15-30 ans	297	21.8	2.0	16.7	20.4	21.8	23.2	27.7
% Emploi Industriel	297	19.7	8.8	3.7	13.3	18.2	24.8	52.0
% Emploi Public	297	33.5	6.2	15.0	29.5	33.2	36.9	51.0

TABLE 6.1 – Descriptif de l'échantillon
Note : La zone géographique est la zone d'emploi. Les statistiques ne sont pas pondérées.

La figure 6.5 est cohérente 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 associée 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.

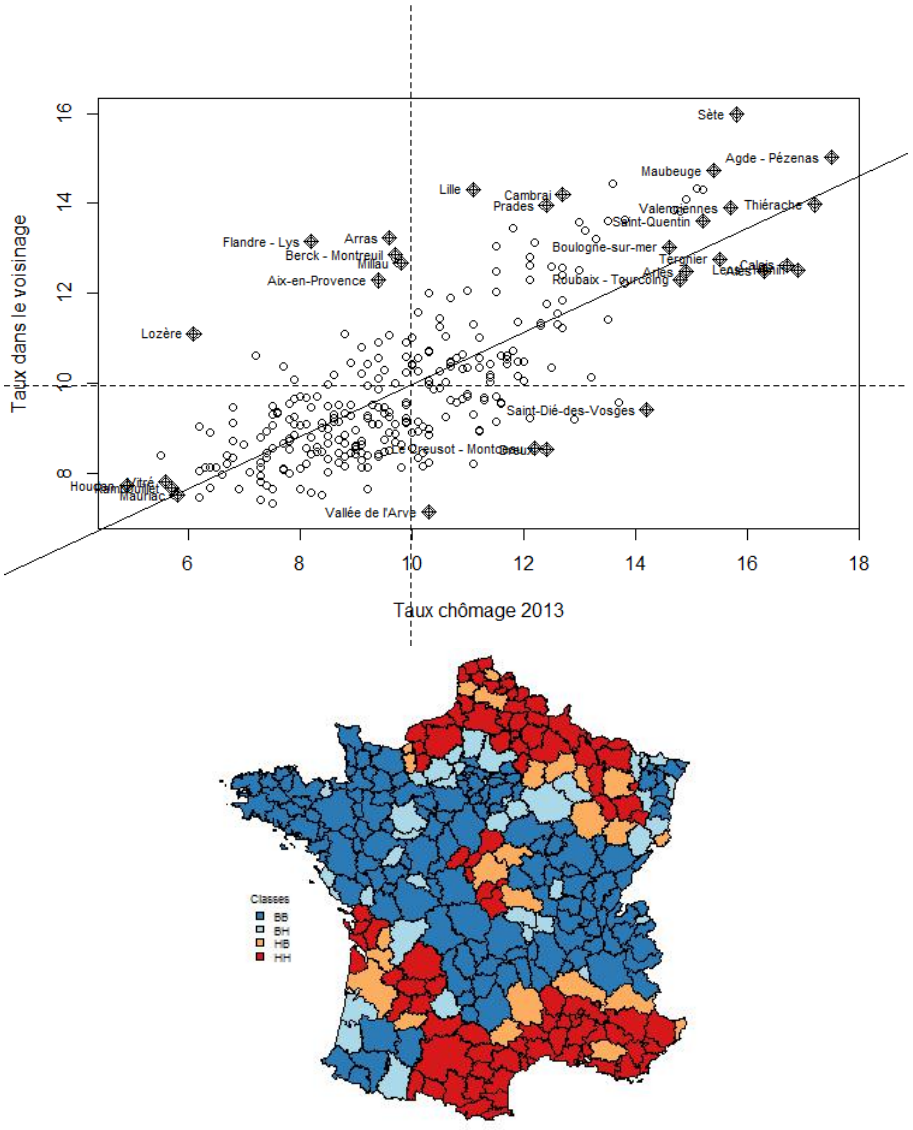

FIGURE 6.5 – Graphique de Moran du taux de chômage et carte associée

6.5.2 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 deux 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 de

corrélation spatiale que celle présente dans les deux variables. Première chose, on commence donc par estimer un modèle linéaire non spatial à l'aide des 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 corrélation spatiale (endogène, exogène ou inobservée), la démarche est pragmatique. L'approche de 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 est néanmoins estimé, pour 6 matrices de voisinage : contiguïté, 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 estimé
modele <- txcho_2013 ~ tx_act+part_act_peudip+part_act_1530+part_emp_ind+
    part_emp_pub
### Matrice de voisinage
matrice <- dist.w

### Modèle MCO
ze.lm <- lm(modele, data=donnees_ze)
summary(ze.lm)

### Test de Moran adapté sur les résidus
lm.morantest(ze.lm,matrice)

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

### Modèle SEM
ze.sem<-errorsarlm(modele, data=donnees_ze, matrice)
summary(ze.sem)
### Test d'Hausman
Hausman.test(ze.sem)

### Modèle SAR
ze.sar<-lagsarlm(modele, data=donnees_ze, matrice)
summary(ze.sar)

### Modèle SDM
ze.sardm<-lagsarlm(modele, data=donnees_ze, matrice, type="mixed")
summary(ze.sardm)
### Test de l'hypothèse de facteur commun
# ze.sardm : Modèle non contraint
# ze.sem : Modèle contraint
FC.test<-LR.sarlm(ze.sardm,ze.sem)
print(FC.test)
```

On ne présente ici que les résultats associés à la matrice de distances 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 contiguïté ou les plus proches voisins peuvent 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, quelle que soit la matrice de voisinage retenue.

Nous nous attendons 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 touchés par le chômage. Les zones de fort emploi industriel sont *a priori* plus affectées par le chômage (réaction de l'emploi à la conjoncture et fermeture d'usines). Au contraire, les emplois publics étant plus stables, le pourcentage d'emploi public devrait être négativement corrélé avec le taux de chômage. Rappelons ici que ce modèle se veut illustratif des techniques d'économétrie spatiale, aucune conclusion économique ne peut en être tirée.

Concernant le choix du modèle, on peut retenir les points suivants de la table 6.2.

— L'approche séquentielle d'Elhorst (présentée en 6.3.2) conduirait à retenir un modèle SDM (colonne 4). Il présente l'AIC le plus faible (960). L'ensemble des tests d'autocorrélation spatiale menés à partir des résidus du modèle MCO sont rejetés (colonne 1). De même, l'hypothèse de facteur commun du modèle SDM est rejetée (p-value de 0.004). Plusieurs effets d'interaction exogène sont significativement non nuls (le pourcentage d'actifs non diplômés au seuil de 1 %). Enfin, pour le modèle à interactions exogènes (SLX, colonne 6), on ne rejette pas l'hypothèse d'absence d'autocorrélation résiduelle sous l'hypothèse de corrélation endogène (test robuste LM-Error, p-value de 0.787).

— Le choix d'un modèle SAR (colonne 3) serait ici déconseillé. Un test montre qu'une autocorrélation spatiale résiduelle reste présente (p-value (test LM residual auto.) de 0.003). Les conséquences sont importantes sur l'interprétation des résultats. La variable "pourcentage d'emploi industriel" reste significative à 1 % (quelle que soit la matrice de voisinage), alors que le signe négatif peut paraître contre-intuitif.

— Le modèle de Manski (colonne 8) fournit des résultats divergents selon la matrice de voisinage (non présentés ici), certainement par manque d'identifiabilité de ce modèle. De même, le modèle SAC (corrélation endogène et résiduelle, colonne 5) estime une corrélation endogène faible et non significative en comparaison de l'autocorrélation résiduelle. Ce résultat est difficile à interpréter et peut provenir d'une mauvaise spécification du modèle (Le Gallo 2002).

Enfin, pour des raisons de parcimonie, le choix d'un modèle SEM (table 6.2, colonne 2) voire SDEM (colonne 7) pourrait être envisagé, après avoir vérifié la cohérence des résultats avec ceux du modèle SDM. L'interprétation de ce modèle SEM est en effet plus aisée mais se limite aux effets directs. Le critère AIC (967) est proche du modèle SDM, et pour des matrices de poids des 5 ou 10 plus proches voisins (table 6.3, colonnes 4 et 5), l'hypothèse de facteur commun n'est pas rejetée à 1 %. La divergence de résultats entre les modèles MCO et SEM pourrait amener à conclure que la spécification du modèle SEM n'est pas juste, *i.e.* qu'elle souffre d'un biais de variable omise. Un test d'Hausman (LeSage et Pace 2009 p.61-63) entre les modèles MCO et SEM repose sur l'hypothèse nulle de validité des deux modèles, le modèle SEM étant plus efficace. On constate alors que cette hypothèse n'est pas rejetée au seuil de 1 %, hormis pour la matrice de poids des 2 plus proches voisins (table 6.3).

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 0.09 à 0.12 point par rapport au

	(1) MCO	(2) SEM	(3) SAR	(4) SDM	(5) SAC	(6) SLX	(7) SDEM	(8) Manski
Taux d'activité	-0.622***	-0.498***	-0.437***	-0.472***	-0.499***	-0.470***	-0.486***	-0.473***
	(0.039)	(0.041)	(0.038)	(0.042)	(0.041)	(0.050)	(0.041)	(0.042)
% Actifs Peu Diplômés	0.186***	0.184***	0.138***	0.182***	0.179***	0.179***	0.181***	0.183***
	(0.026)	(0.027)	(0.022)	(0.027)	(0.026)	(0.033)	(0.027)	(0.028)
% Jeunes Actifs 15-30 ans	0.138***	0.196***	0.087**	0.209***	0.180***	0.205***	0.197***	0.211***
	(0.043)	(0.045)	(0.037)	(0.046)	(0.045)	(0.055)	(0.045)	(0.047)
% Emploi Industriel	-0.062***	-0.018	-0.036***	-0.015	-0.021*	-0.022	-0.024**	-0.014
	(0.012)	(0.012)	(0.010)	(0.012)	(0.012)	(0.014)	(0.012)	(0.012)
% Emploi Public	-0.068***	-0.044***	-0.063***	-0.042**	-0.048***	-0.044**	-0.049***	-0.041**
	(0.019)	(0.016)	(0.016)	(0.016)	(0.017)	(0.019)	(0.017)	(0.016)
$\hat{\rho}$			0.519***	0.629***	0.205*			0.689***
			(0.049)	(0.064)	(0.109)			(0.120)
$\hat{\lambda}$		0.747***			0.616***		0.651***	-0.137
		(0.051)			(0.096)		(0.063)	(0.257)
$\hat{\theta}$, Taux d'activité				0.157*		-0.300***	-0.277***	0.205*
				(0.083)		(0.082)	(0.105)	(0.111)
$\hat{\theta}$, % Actifs Peu Diplômés				-0.135***		-0.027	-0.021	-0.145***
				(0.045)		(0.052)	(0.066)	(0.046)
$\hat{\theta}$, % Jeunes Actifs 15-30 ans				-0.140 *		-0.041	-0.003	-0.153**
				(0.072)		(0.085)	(0.115)	(0.072)
$\hat{\theta}$, % Emploi Industriel				-0.044**		-0.118***	-0.073**	-0.038*
				(0.020)		(0.023)	(0.029)	(0.023)
$\hat{\theta}$, % Emploi Public				-0.024		-0.084*	-0.070	-0.018
				(0.037)		(0.043)	(0.052)	(0.037)
Constante	51.653***	39.729***	34.470***	27.456***	38.427***	66.077***	63.650***	23.530***
	(3.635)	(3.685)	(3.407)	(6.766)	(3.901)	(6.514)	(10.213)	(9.065)
Observations	297	297	297	297	297	297	297	297
AIC	1072	967	980	960	967	1029	964	962
R^2 Ajusté	0.624					0.679		
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					0.787		
Test Robuste LM-Lag	0.000					0.001		
Test Facteur Commun				0.004				
Test LM residual auto.			0.003	0.572				

TABLE 6.2 – Déterminants du taux de chômage par zone d'emploi, à partir d'une matrice inverse de la distance

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

modèle SEM. Pour le pourcentage d'emploi industriel, le modèle MCO conclut à un effet négatif significatif alors qu'il est jugé nul avec le modèle SEM dans le cas d'une matrice inverse de la distance ou plus faible avec les autres matrices. L'effet du taux d'activité pourrait être surévalué avec une matrice de contiguïté 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 (table 6.7 en annexe de ce chapitre), 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.

Les résultats pour le modèle SEM ne sont pas toujours robustes au choix de la matrice de voisinage, le "pourcentage d'emploi industriel" pouvant se révéler ou non significatif. Il n'y a pas de choix évident de matrice de voisinage, qui amènerait à privilégier les résultats obtenus avec une matrice inverse de la distance par exemple. Le choix ne doit bien sûr en aucun cas être dicté par un argument de significativité des résultats, mais reposer sur une analyse associée à la question économique.

	(1) MCO	(2) SEM Contiguïté	(3) SEM 2 Voisins	(4) SEM 5 Voisins	(5) SEM 10 Voisins	(6) SEM Distance	(7) SEM Endogène
Taux d'activité	-0.622***	-0.518***	-0.517***	-0.530***	-0.507***	-0.498***	-0.515***
	(0.039)	(0.040)	(0.040)	(0.040)	(0.040)	(0.041)	(0.041)
% Actifs Peu Diplômés	0.186***	0.188***	0.204***	0.185***	0.181***	0.184***	0.184***
	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)	(0.027)	(0.026)
% Jeunes Actifs 15-30 ans	0.138***	0.179***	0.195***	0.201***	0.198***	0.196***	0.139***
	(0.043)	(0.045)	(0.044)	(0.045)	(0.046)	(0.045)	(0.044)
% Emploi Industriel	-0.062***	-0.023*	-0.027**	-0.023*	-0.024**	-0.018	-0.026**
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
% Emploi Public	-0.068***	-0.042**	-0.039**	-0.047***	-0.048***	-0.044***	-0.050***
	(0.019)	(0.017)	(0.017)	(0.017)	(0.017)	(0.016)	(0.016)
$\hat{\lambda}$		0.687***	0.506***	0.681***	0.763***	0.747***	0.700***
		(0.050)	(0.047)	(0.051)	(0.053)	(0.051)	(0.044)
Constante	51.653***	41.535***	40.672***	42.166***	40.685***	39.729***	42.414***
	(3.635)	(3.681)	(3.643)	(3.639)	(3.644)	(3.685)	(3.745)
Observations	297	297	297	297	297	297	297
AIC	1072	977	996	972	973	967	995
Test Hausman		0.030	0.000	0.042	0.114	0.029	0.115
Test Facteur Commun		0.002	0.001	0.040	0.035	0.004	0.000

TABLE 6.3 – Modèle SEM, pour différentes matrices de voisinage
Note : Le modèle SEM est estimé avec 6 matrices de voisinage différentes. Les écarts-types sont indiqués entre parenthèses. Significativité : * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

6.5.3 Interprétation des résultats

Pour le modèle SDM, afin de permettre une interprétation au regard du modèle MCO et SEM, on calcule les effets directs et indirects tels que décrits en section 6.4 (tables 6.4 et 6.5). Les intervalles de confiance empiriques sont obtenus à l'aide de 1 000 simulations à partir de la distribution empirique. 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. Le modèle SDM met en avant le rôle particulier du pourcentage d'emploi industriel, qui seul aurait un effet indirect (négatif) associé à un effet direct (négatif) faible ou nul selon la matrice de voisinage retenue. La compréhension économique d'un tel résultat demeure délicate. Le modèle SDM peut amener à interpréter de manière fallacieuse la corrélation 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.

```
### Estimation des effets directs et indirects du modèle SDM
```

```
impactssdm<-impacts(ze.sardm, listw=matrice, R=1000)
summary(impactssdm)
```

	(1) MCO	(2) SDM Contiguïté	(3) SDM 2 Voisins	(4) SDM 5 Voisins	(5) SDM 10 Voisins	(6) SDM Distance	(7) SDM Endogène
Taux d'activité	-0.622	-0.509	-0.510	-0.529	-0.505	-0.490	-0.508
	[-0.700,-0.545]	[-0.588,-0.435]	[-0.589,-0.434]	[-0.611,-0.451]	[-0.583,-0.422]	[-0.574,-0.409]	[-0.588,-0.429]
% Actifs Peu Diplômés	0.186	0.178	0.208	0.183	0.177	0.180	0.178
	[0.136,0.237]	[0.122,0.232]	[0.154,0.261]	[0.132,0.235]	[0.125,0.230]	[0.122,0.230]	[0.129,0.232]
% Jeunes Actifs 15-30 ans	0.138	0.194	0.223	0.213	0.212	0.207	0.184
	[0.054,0.223]	[0.102,0.288]	[0.135,0.312]	[0.123,0.309]	[0.119,0.306]	[0.119,0.299]	[0.092,0.279]
% Emploi Industriel	-0.062	-0.026	-0.032	-0.027	-0.027	-0.022	-0.033
	[-0.087,-0.038]	[-0.048,-0.003]	[-0.053,-0.008]	[-0.051,-0.005]	[-0.050,-0.005]	[-0.045,0.001]	[-0.055,-0.011]
% Emploi Public	-0.068	-0.045	-0.048	-0.052	-0.051	-0.049	-0.052
	[-0.106,-0.030]	[-0.078,-0.010]	[-0.081,-0.011]	[-0.084,-0.017]	[-0.083,-0.018]	[-0.081,-0.014]	[-0.084,-0.019]

TABLE 6.4 – Impacts directs du modèle SDM, pour différentes matrices de voisinage
Note : Le modèle SDM est estimé avec 6 matrices de voisinage différentes. Les intervalles de confiance empiriques (quantiles à 2,5 % et 97,5 % de 1000 simulations MCMC) sont indiqués entre crochets.

	(1) SDM Contiguïté	(2) SDM 2 Voisins	(3) SDM 5 Voisins	(4) SDM 10 Voisins	(5) SDM Distance	(6) SDM Endogène
Taux d'activité	-0.323	-0.200	-0.241	-0.306	-0.357	-0.351
	[-0.587,-0.091]	[-0.337,-0.068]	[-0.488,0.007]	[-0.700,0.030]	[-0.658,-0.073]	[-0.638,-0.107]
% Actifs Peu Diplômés	-0.015	-0.059	-0.032	-0.050	-0.053	-0.079
	[-0.161,0.142]	[-0.146,0.032]	[-0.205,0.124]	[-0.291,0.158]	[-0.254,0.137]	[-0.251,0.085]
% Jeunes Actifs 15-30 ans	-0.016	-0.079	-0.082	0.016	-0.023	0.047
	[-0.321,0.249]	[-0.214,0.058]	[-0.334,0.174]	[-0.321,0.390]	[-0.352,0.301]	[-0.230,0.332]
% Emploi Industriel	-0.130	-0.064	-0.100	-0.135	-0.136	-0.111
	[-0.208,-0.055]	[-0.105,-0.022]	[-0.170,-0.030]	[-0.244,-0.041]	[-0.229,-0.059]	[-0.187,-0.043]
% Emploi Public	-0.120	-0.078	-0.113	-0.098	-0.130	-0.037
	[-0.274,0.017]	[-0.140,-0.011]	[-0.257,0.031]	[-0.345,0.132]	[-0.335,0.046]	[-0.186,0.106]

TABLE 6.5 – Impacts indirects du modèle SDM, pour différentes matrices de voisinage
Note : Le modèle SDM est estimé avec 6 matrices de voisinage différentes. Les intervalles de confiance empiriques (quantiles à 2,5 % et 97,5 % de 1000 simulations MCMC) sont indiqués entre crochets.

6.5.4 Autres modélisations spatiales

L'analyse descriptive a mis en avant une hétérogénéité spatiale possible du modèle. Il serait possible d'intégrer et de tester la présence de ce phénomène, soit en autorisant le modèle à être hétéroscédastique (*via* le package *sphet*, PIRAS et al. 2010), soit en modélisant une variabilité spatiale des paramètres ou de la forme fonctionnelle du modèle. Cette seconde forme d'hétérogénéité est obtenue en incluant des indicatrices de zones géographiques dans le modèle, à l'aide d'un modèle de lissage géographique (*via* le package *McSpatial*, qui inclut des modèles spatiaux semi-paramétriques ou par splines) ou en conduisant une analyse géographique pondérée.

La mise en œuvre pratique d'une régression géographiquement pondérée est détaillée dans le chapitre 9 : "Régression géographiquement pondérée". Nous présentons ici les résultats de l'estimation géographiquement pondérée du modèle linéaire reliant le taux de chômage et les caractéristiques structurelles présenté plus haut.

La table 6.6 fournit les valeurs minimales, maximales et les quartiles des coefficients obtenus. On peut ainsi apprécier la variabilité des coefficients, et comparer ces résultats avec ceux des MCO. L'utilisation de la régression géographique conduit à des coefficients qui ne sont pas toujours de même signe. Cela peut conduire à s'interroger sur le bien-fondé de la spécification. Les coefficients peuvent varier de façon sensible, notamment pour les actifs de 15 à 30 ans, le coefficient médian s'écartant très sensiblement de celui des MCO.

	(1) MCO	(2) Minimum	(3) P1	(4) Médiane	(5) P3	(6) Maximum
Taux d'activité	-0.622	-1.492	-0.653	-0.508	-0.379	-0.133
% Actifs Peu Diplômés	0.186	-0.116	0.081	0.188	0.250	0.607
% Jeunes Actifs 15-30 ans	0.138	-0.753	-0.040	0.183	0.340	0.875
% Emploi Industriel	-0.062	-0.233	-0.066	-0.029	0.006	0.184
% Emploi Public	-0.068	-0.318	-0.098	-0.048	-0.002	0.218
Constante	51.650	-7.485	29.940	40.440	52.310	130.500

TABLE 6.6 – Résultats de la régression géographique pondérée

On récupère une table contenant pour chacun des points d'estimation (ici les centroïdes des zones d'emploi) la valeur des coefficients, la valeur prédite par le modèle, les résidus et la valeur locale du R^2. Cela permet notamment de cartographier les variations locales des paramètres. Cette dimension cartographique est importante pour apprécier les tendances spatiales. On peut également vérifier si les résidus restent autocorrélés spatialement, à l'aide de cartes et de tests de Moran adaptés. Il n'y a pas une structure spatiale marquée des résidus dans le cas présent. La distribution des paramètres spatiaux pour la part de l'emploi industriel et de l'emploi public (figure 6.6) met en avant des particularités régionales, qui peuvent permettre de comprendre des résultats surprenants, par exemple la relation nulle (ou négative) entre emploi industriel et taux de chômage. Cette relation négative est présente principalement dans la partie Sud de la France (ainsi que quelques zones du Nord), alors que des zones du Centre et de l'Est, régions ayant subi de fortes restructurations industrielles, présentent une corrélation positive entre taux de chômage et part de l'emploi industriel. Concernant l'emploi public, on constate une relation négative avec le taux de chômage pour une partie du Sud de la France et du Nord, alors que la relation est positive en Bretagne par exemple. Notre modèle inclut un nombre limité de variables, l'effet de certaines particularités régionales (restructurations industrielles, caractéristiques de l'offre d'emploi, etc.) pourrait ainsi être capté à tort par nos variables explicatives, biais classique d'endogénéité. Il est également possible que les comportements soient hétérogènes entre zones d'emploi. Dans tous les cas, cette analyse devrait nous amener à modifier notre modèle, par l'inclusion d'autres variables ou de paramètres de corrélation spatiale par zones géographiques. Nous limitons ici notre analyse, en rappelant que les résultats présentés ne visent qu'à illustrer la démarche du choix et de l'estimation d'un modèle spatial. Prendre en compte à la fois l'hétérogénéité et la corrélation spatiales demeure délicat.

Nous avons effectué les tests permettant de vérifier la non-stationnarité, et donc d'apprécier si la régression géographique pondérée est préférable au modèle linaire estimé par les MCO (BRUNSDON et al. 2002 ; LEUNG et al. 2000). La stationnarité est rejetée ici quel que soit le test, au niveau global et pour chaque variable explicative (résultats non présentés ici).

La régression géographique pondérée est considérée comme une bonne méthode exploratoire,

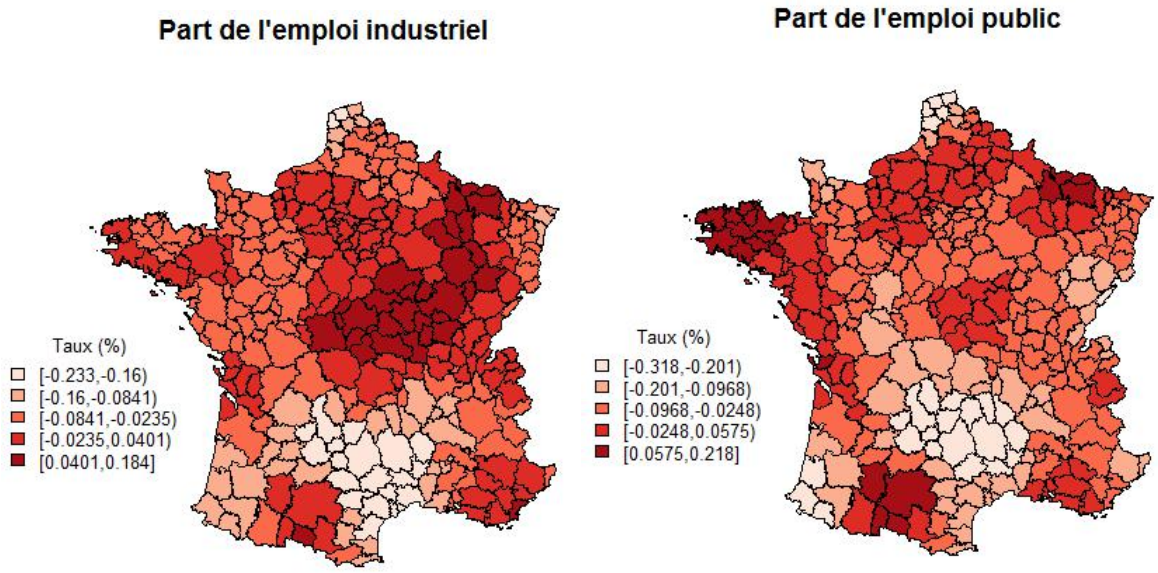

FIGURE 6.6 – Distribution des paramètres locaux

permettant notamment de visualiser des phénomènes de non-stationnarité. Mais elle a fait aussi l'objet d'un certain nombre de critiques. WHEELER et al. 2009 soulignent que les résultats ne sont pas robustes à une forte corrélation entre variables explicatives ou à la présence conjointe d'autocorrélation spatiale. De plus, comme dans toutes les méthodes statistiques non paramétriques, la distance introduite (*i.e.* le choix de la fenêtre) n'est pas neutre. Une grande distance, introduisant de nombreux points, va conduire à des coefficients variant peu localement. À l'inverse une faible distance introduira beaucoup de variabilité. Le choix opéré peut avoir des conséquences sur les tests appréciant le choix de la régression géographique pondérée par rapport aux MCO. Le package *GWmodel* (BRUNSDON et al. 2015) tente de répondre à ces critiques.

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 question économique et de la complexité de ces nouveaux modèles, en particulier en termes d'interprétation.

Le choix de modéliser la corrélation ou l'hétérogénéité spatiale, voire les deux simultanément, est délicat. Dans notre exemple, prendre en compte la corrélation spatiale pour modéliser le taux de chômage localisé apparaît nécessaire d'après les tests statistiques. Cela corrige certaines interprétations erronées issues du modèle linéaire classique. Il conviendrait ici de privilégier un modèle spatial de Durbin (SDM), voire aux erreurs spatialement autocorrélées (modèle SEM). Mais l'analyse de l'hétérogénéité spatiale à partir de régressions géographiques pondérées souligne également que la spécification devrait être améliorée, certains résultats surprenants pouvant provenir

d'un biais de variables omises et d'une mauvaise prise en compte de l'hétérogénéité spatiale des marchés du travail. Cette incertitude sur le choix du modèle doit amener à rester prudent quant à l'interprétation des effets directs et indirects du modèle SDM. De plus, ce n'est pas parce que le modèle est plus compliqué qu'il règle le problème de l'endogénéité des variables explicatives ou du sens de la causalité entre les variables du modèle. Aucune interprétation causale n'est ici possible.

Les enjeux théoriques de ces méthodes, et en particulier les liens entre corrélation et hétérogénéité spatiales, ne sont pas complètement maîtrisés. Les modèles d'économétrie spatiale permettent une prise en compte de l'espace ou des relations entre agents, préférable bien souvent à ne rien faire. La régression géographique pondérée et le lissage géographique permettent en complément des approches descriptives, de définir de grands ensembles régionaux homogènes et des analyses complémentaires à des tests de rupture régionale. Néanmoins, estimer ces modèles suppose de disposer de données exhaustives. Dans le cas général, ils ne sont donc pas adaptés aux données d'enquêtes.

Annexes

Annexe 1 : Codes R complémentaires

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
                }
}
mig.w<-mat2listw(lien,style="W")
```

Modèles linéaires spatiaux : estimations complémentaires

```
### Modèle SAC
```

```
ze.sac<-sacsarlm(modele, data=donnees_ze, matrice)
summary(ze.sac)

### Modèle SLX
ze.slx<-lmSLX(modele, data=donnees_ze, matrice)
summary(ze.slx)

### Modèle SDEM
ze.sdem<-errorsarlm(modele, data=donnees_ze, matrice, etype="emixed")
summary(ze.sdem)

### Modèle Manski
ze.manski<-sacsarlm(modele, data=donnees_ze, matrice, type="sacmixed")
summary(ze.manski)
```

Annexe 2 : Modèle SDM, pour différentes matrices de voisinage

	(1) SDM Contiguïté	(2) SDM 2 Voisins	(3) SDM 5 Voisins	(4) SDM 10 Voisins	(5) SDM Distance	(6) SDM Endogène
Taux d'activité	-0.486***	-0.485***	-0.513***	-0.494***	-0.472***	-0.485***
	(0.042)	(0.042)	(0.041)	(0.041)	(0.042)	(0.042)
% Actifs Peu Diplômés	0.180***	0.215***	0.186***	0.179***	0.182***	0.184***
	(0.027)	(0.028)	(0.028)	(0.027)	(0.027)	(0.028)
% Jeunes Actifs 15-30 ans	0.196***	0.232***	0.219***	0.211***	0.209***	0.181***
	(0.047)	(0.046)	(0.047)	(0.048)	(0.046)	(0.047)
% Emploi Industriel	-0.016	-0.024**	-0.020*	-0.022*	-0.015	-0.026**
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
% Emploi Public	-0.037**	-0.038**	-0.044***	-0.048***	-0.042**	-0.050*
	(0.017)	(0.017)	(0.017)	(0.017)	(0.016)	(0.016)
$\hat{\rho}$	0.601***	0.448***	0.606***	0.647***	0.629***	0.609***
	(0.057)	(0.050)	(0.057)	(0.068)	(0.064)	(0.051)
$\hat{\theta}$, Taux d'activité	0.153**	0.094	0.209***	0.207**	0.157*	0.149**
	(0.075)	(0.057)	(0.072)	(0.087)	(0.083)	(0.075)
$\hat{\theta}$, % Actifs Peu Diplômés	-0.114***	-0.133***	-0.126***	-0.134***	-0.135***	-0.145***
	(0.040)	(0.034)	(0.041)	(0.047)	(0.045)	(0.040)
$\hat{\theta}$, % Jeunes Actifs 15-30 ans	-0.124*	-0.153***	-0.167***	-0.131*	-0.140*	-0.090
	(0.069)	(0.053)	(0.065)	(0.078)	(0.072)	(0.068)
$\hat{\theta}$, % Emploi Industriel	-0.046**	-0.029**	-0.030	-0.035	-0.044**	-0.031*
	(0.021)	(0.015)	(0.019)	(0.022)	(0.020)	(0.018)
$\hat{\theta}$, % Emploi Public	-0.029	- 0.031	-0.020	- 0.005	-0.024	0.015
	(0.033)	(0.022)	(0.031)	(0.043)	(0.037)	(0.031)
Constante	28.582***	33.848***	26.710***	24.504***	27.456***	27.662***
	(6.184)	(4.814)	(5.844)	(7.372)	(6.766)	(6.312)
Observations	297	297	297	297	297	297
AIC	968	985	970	971	960	987
Test Facteur Commun	0.002	0.001	0.040	0.035	0.004	0.000
Test LM residual auto.	0.054	0.263	0.071	0.715	0.572	0.135

TABLE 6.7 – Modèle SDM, pour différentes matrices de voisinage
Note : Le modèle SDM est estimé avec 6 matrices de voisinage différentes. Les écarts-types sont indiqués entre parenthèses. Significativité : * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Références - Chapitre 6

ABREU, Maria, Henri DE GROOT et Raymond FLORAX (2004). « Space and growth : a survey of empirical evidence and methods ».

ALDSTADT, Jared et Arthur GETIS (2006). « Using AMOEBA to create a spatial weights matrix and identify spatial clusters ». *Geographical Analysis* 38.4, p. 327–343.

ANSELIN, Luc (2001). « Spatial econometrics ». *A companion to theoretical econometrics* 310330.

— (2002a). « Under the hood : Issues in the specification and interpretation of spatial regression models ». *Agricultural economics* 27.3, p. 247–267.

ANSELIN, Luc et Daniel A GRIFFITH (1988). « Do spatial effects really matter in regression analysis ? » *Papers in Regional Science* 65.1, p. 11–34.

ANSELIN, Luc et al. (1996). « Simple diagnostic tests for spatial dependence ». *Regional science and urban economics* 26.1, p. 77–104.

ARBIA, Giuseppe (2014). *A primer for spatial econometrics : with applications in R*. Springer.

BARRIOS, Thomas et al. (2012). « Clustering, spatial correlations, and randomization inference ». *Journal of the American Statistical Association* 107.498, p. 578–591.

BECK, Nathaniel, Kristian Skrede GLEDITSCH et Kyle BEARDSLEY (2006). « Space is more than geography : Using spatial econometrics in the study of political economy ». *International studies quarterly* 50.1, p. 27–44.

BHATTACHARJEE, Arnab et Chris JENSEN-BUTLER (2013). « Estimation of the spatial weights matrix under structural constraints ». *Regional Science and Urban Economics* 43.4, p. 617–634.

BLANC, Michel et François HILD (2008). « Analyse des marchés locaux du travail : du chômage à l'emploi ». fre. *Economie et Statistique* 415.1, p. 45–60. ISSN : 0336-1454. DOI : 10.3406/estat.2008.7019. URL : https://www.persee.fr/doc/estat_0336-1454_2008_num_415_1_7019.

BRUNSDON, C. et al. (2002). « Geographically weighted summary statistics : a framework for localised exploratory data analysis ». *Computers, Environment and Urban Systems*.

BRUNSDON, Chris et Lex COMBER (2015). *An Introduction to R for Spatial Analysis Et Mapping*. Sage London.

BRUNSDON, Chris, A Stewart FOTHERINGHAM et Martin E CHARLTON (1996). « Geographically weighted regression : a method for exploring spatial nonstationarity ». *Geographical analysis* 28.4, p. 281–298.

CORRADO, Luisa et Bernard FINGLETON (2012). « Where is the economics in spatial econometrics ? » *Journal of Regional Science* 52.2, p. 210–239.

DUBIN, Robin A (1998). « Spatial autocorrelation : a primer ». *Journal of housing economics* 7.4, p. 304–327.

ELHORST, J Paul (2010). « Applied spatial econometrics : raising the bar ». *Spatial Economic Analysis* 5.1, p. 9–28.

FAFCHAMPS, Marcel (2015). « Causal Effects in Social Networks ». *Revue économique* 66.4, p. 657–686.

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, p. 319–339.

— (2012). « Endogeneité et autocorrélation spatiale : quelle utilité pour le modèle de Durbin ? » *Revue d'Économie Régionale & Urbaine* 1, p. 3–17.

FLACHAIRE, Emmanuel (2005). « Bootstrapping heteroskedastic regression models : wild bootstrap vs. pairs bootstrap ». *Computational Statistics & Data Analysis* 49.2, p. 361–376.

FLOCH, J.M. (2012). « Détection des disparités socio-économiques - L'apport de la statistique spatiale ». *Documents de Travail Insee*.

FLORAX, Raymond JGM, Hendrik FOLMER et Sergio J REY (2003). « Specification searches in spatial econometrics : the relevance of Hendry's methodology ». *Regional Science and Urban Economics* 33.5, p. 557–579.

GIBBONS, Stephen et Henry G OVERMAN (2012). « Mostly pointless spatial econometrics ? » *Journal of Regional Science* 52.2, p. 172–191.

GIVORD, Pauline et al. (2016). « Quels outils pour mesurer la ségrégation dans le système éducatif ? Une application à la composition sociale des collèges français ». *Education et formation*.

GRISLAIN-LETRÉMY, Céline et Arthur KATOSSKY (2013). « Les risques industriels et le prix des logements ». *Economie et statistique* 460.1, p. 79–106.

HARRIS, Richard, John MOFFAT et Victoria KRAVTSOVA (2011). « In search of 'W' ». *Spatial Economic Analysis* 6.3, p. 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, p. 140–149.

KELEJIAN, Harry H et Ingmar R PRUCHA (2007). « HAC estimation in a spatial framework ». *Journal of Econometrics* 140.1, p. 131–154.

— (2010a). « Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances ». *Journal of Econometrics* 157.1, p. 53–67.

KELEJIAN, H.H. et I.R. PRUSHA (2010b). « Spatial models with spatially lagged dependent variables and incomplete data ». *Journal of geographical systems*.

LE GALLO, Julie (2002). « Econométrie spatiale : l'autocorrélation spatiale dans les modèles de régression linéaire ». *Economie & prévision* 4, p. 139–157.

— (2004). « Hétérogénéité spatiale ». *Économie & prévision* 1, p. 151–172.

LEE, Lung-Fei (2004). « Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models ». *Econometrica* 72.6, p. 1899–1925.

LESAGE, James (2014). « What regional scientists need to know about spatial econometrics ».

LESAGE, James et Robert K PACE (2009). *Introduction to spatial econometrics*. Chapman et Hall/CRC.

LEUNG, Yee, Chang-Lin MEI et Wen-Xiu ZHANG (2000). « Statistical tests for spatial nonstationarity based on the geographically weighted regression model ». *Environment and Planning A* 32.1, p. 9–32.

LOONIS, Vincent (2012). « Non-réponse à l'Enquête Emploi et modèles probit spatiaux ».

LOTTMANN, Franziska (2013). « Spatial dependence in German labor markets ».

MANSKI, Charles F (1993). « Identification of Endogenous Social Effects : The Reflection Problem ». *Review of Economic Studies* 60.3, p. 531–542.

OSLAND, Liv (2010). « An application of spatial econometrics in relation to hedonic house price modeling ». *Journal of Real Estate Research* 32.3, p. 289–320.

PIRAS, Gianfranco et al. (2010). « sphet : Spatial models with heteroskedastic innovations in R ». *Journal of Statistical Software* 35.1, p. 1–21.

SLADE, Margaret E (2005). « The role of economic space in decision making ». *Annales d'Economie et de Statistique*, p. 1–20.

THOMAS-AGNAN, Christine, Thibault LAURENT et Michel GOULARD (2014). « About predictions in spatial autoregressive models ».

WAELBROECK, Patrick (2005). « The Role of Economic Space in Decision Making : Comment ». *Annales d'Economie et de Statistique*, p. 29–31.

WANG, Wei et Lung-Fei LEE (2013b). « Estimation of spatial autoregressive models with randomly missing data in the dependent variable ». *The Econometrics Journal* 16.1, p. 73–102.

WHEELER, D et A PÁEZ (2009). *Geographically weighted regression. 1er MM, Getis A (eds) Handbook of applied spatial analysis*.