Méthodologie statistique

M 2016/06 – H 2016/02

Econométrie spatiale : une introduction pratique

Ronan Le Saout – Jean-Michel Floch

Document de travail

Institut National de la Statistique et des Études Économiques

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Série des documents de travail « Méthodologie Statistique »

de la Direction de la Méthodologie et de la Coordination Statistique et Internationale et de la Direction de la Diffusion et de l'Action Régionale

M 2016/06 – H2016/02

Économétrie spatiale : une introduction pratique

Ronan Le Saout [*] - Jean-Michel Floch [**]

Nous remercions Salima Bouayad Agha, Marie-Pierre De Bellefon, Pauline Givord, François Hild, Raphaël Lardeux, Alain Le, Julie Le Gallo, Vincent Loonis, Thomas Merly-Alpa, et Olivier Sautory, ainsi que les participants aux JMS 2015 et séminaires de l'INSEE auxquels ce travail a été présenté, pour leurs commentaires et conseils. Cette étude ne reflète pas les opinions de l'INSEE. Nous restons seuls responsables des erreurs ou omissions qui pourraient y demeurer.

[*] DMCSI
[**] DDAR

Direction de la méthodologie et de la coordination statistique et internationale -Département des Méthodes Statistiques - Timbre L101 - 18, bd Adolphe Pinard - 75675 PARIS CEDEX - France -
Tél. : 33 (1) 41 17 66 01 - Fax : 33 (1) 41 17 66 33 - CEDEX - E-mail : L001-dg@insee.fr - Site Web Insee : http://www.insee.fr

Économétrie spatiale : une introduction pratique

Résumé

Ce document de travail décrit la conduite d'une étude d'économétrie spatiale, à travers une modélisation descriptive du taux de chômage par zone d'emploi. 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 document se concentre sur l'étude de la corrélation spatiale, et donc sur ces différentes interactions, et détaille 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 pas à pas de choix de modèle (estimation et tests) est détaillée. Des effets de rétroaction entraînent une interprétation particulière (et plus complexe) des résultats.

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

Spatial econometrics : a practical introduction

This working paper describes the conduct of a spatial econometric study, through a descriptive modeling of the unemployment rate by employment area. Spatial models nevertheless have a broader application, the approach being compatible with any instance in which neighborhood relationships are involved. Indeed, economic theory characterizes many cases of interactions between agents (products, firms, individuals) which are not necessarily geographically based. The document focuses on the study of the spatial correlation, and therefore these various interactions, and details the links with spatial heterogeneity, i.e. spatially differentiated phenomena. Several forms of interaction exist regarding the dependent variable, the explanatory variables or unobserved variables. Many models therefore find themselves in competition from a same prior definition of neighborhood relationships. A stepwise methodology of model selection (estimation and tests) is detailed. Feedback effects lead to a particular (and more complex) interpretation of results.

Keywords : Spatial econometrics, Interaction models, Unemployment rate.

1 Introduction

Les relations entre les valeurs observées sur des territoires proches préoccupent depuis longtemps les géographes. W.Tobler a résumé cette question à l'aide d'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 aisée 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 lors d'études économiques. Une première étape reste bien sûr de caractériser cette proximité à l'aide d'indicateurs descriptifs et de tests (Floch 2012). Une fois l'autocorrélation spatiale des données détectée vient l'étape de la modélisation dans un cadre multivarié. L'objet de ce document de travail est d'aborder la conduite pratique d'une étude d'économétrie spatiale : quel modèle retenir ? comment en interpréter les résultats ? quelles en sont les limites ?

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

Les techniques présentées s'adaptent à des domaines qui dépassent le cadre strictement géographique. Plusieurs types de données 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). Ce dernier type de données est essentiellement physique, par exemple la hauteur du sol, la température, la qualité de l'air... et relève du domaine de la 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

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

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 *et al.* 2005 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 Katossky 2013), les changements de lieux d'habitation (Guymarc 2015) ou la non-réponse dans l'enquête emploi (Loonis 2012).

Les logiciels commerciaux n'incluent pas de procédures standards d'estimation de modèles d'économétrie spatiale. Des outils spécifiques ont été développés. LeSage et Pace mettent à disposition des programmes MatLab. Luc Anselin a initié le projet "GeoDa", qui est un logiciel libre 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. La partie 2 présente les raisons économiques et statistiques de la mise en place de ces modèles. La partie 3 introduit quelques notions de statistique spatiale, notamment les matrices de voisinage. La partie 4 décrit les étapes de l'estimation d'un modèle d'économétrie spatiale. La partie 5 traite de points techniques plus avancés. La partie 6 détaille la mise en oeuvre sous R à travers la modélisation du taux de chômage par zone d'emploi. La partie 7 conclut. Les lecteurs intéressés par approfondir ces méthodes pourront notamment se référer à LeSage et Pace (2009), Elhorst (2013), Arbia (2014) ou Le Gallo (2002, 2004) pour une présentation en langue française.

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

2.1 Les raisons économiques

L'interaction spatiale, organisationnelle ou sociale des agents économiques est classique en économie. Anselin (2002) liste ainsi plusieurs dénominations économiques telles que les effets de voisinage, de pair, les interactions stratégiques, la copie par mimétisme ou par les

normes sociales ("copy-catting"), la concurrence par comparaison ("yardstick competition")... Par exemple, il met en avant deux théories de concurrence entre firmes justifiant le recours à un modèle spatial ou d'interaction.

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

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

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

2.2 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 Pace (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 (PGD), 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. La difficulté tient au lien à effectuer entre les raisons économiques et économétriques, et la capacité à produire à partir de ces modèles des analyses mettant en évidence des causalités de nature économique (Gibbons et Overman 2012).

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

3.1 La nature des effets spatiaux dans les modèles de régression

La célèbre phrase de Waldo Tobler, citée en introduction, résume bien les choses, mais les simplifie sans doute un peu. Anselin (1988), dans son manuel fondateur, distingue 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 document traite principalement des

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

méthodes de prise en compte de cette corrélation spatiale dans les modèles de régression, détaillés en partie 4. 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 sera synthétisée en partie 5. 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 1988 ; Le Gallo 2002 & 2004). Il n'est pas toujours facile de faire le partage entre ces deux composantes, et la mauvaise spécification de l'une peut être la cause de l'autre. Les tests classiques de l'hétéroscédasticité (i.e. une forme particulière d'hétérogénéité sur le terme d'erreur) sont affectés par l'autocorrélation spatiale, et inversement les tests d'autocorrélation spatiale le sont par l'hétéroscédasticité. Il n'y a pas de solution simple pour intégrer simultanément ces deux phénomènes, en dehors du simple ajout d'indicatrices de territoires dans les modèles d'autocorrélation. De plus, la corrélation des valeurs observées fait que l'information apportée par les données est moins riche que celle 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).

3.2 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 :

- Les matrices de contiguïté qui associent à chaque voisin immédiat la valeur 1 (et 0 dans le cas contraire). Bien que cette approche semble simple, elle laisse place à l'interprétation. Un voisin "immédiat" peut en effet être défini de plusieurs manières selon le mode de déplacement entre les zones, à l'instar d'un jeu d'échecs où la reine se déplace librement mais où le fou ne se déplace qu'en diagonale et la tour qu'horizontalement ou verticalement. Il peut exister des éléments naturels (une rivière par exemple) qui explique qu'un voisin "immédiat" ne soit pas directement accessible.
- Des matrices tenant compte de la distance d entre les zones géographiques, les relations devenant plus faibles avec la distance (1 si $d < d_0$, $1/d^\alpha$, $e^{-\alpha \cdot d}$ avec α un paramètre estimé ou défini *a priori*...). Des distances d'arrêt sont utilisées pour limiter le nombre d'éléments non nuls. Le calcul de la distance (entre les centroïdes, les frontières...) ou la définition de cette distance d'arrêt engendrent de nombreux problèmes pratiques. Une distance trop faible peut créer de nombreuses "îles" (des zones sans voisins). Choisir une distance telle que chaque zone ait au moins un voisin peut créer des zones avec un nombre très important de relations de voisinage si les zones sont de taille hétérogène (par exemple les régions européennes définies au niveau NUTS 2). Griffith (1996) précise qu'il est préférable d'utiliser une matrice sous-spécifiée (distance inférieure à la distance optimale) plutôt qu'une matrice sur-spécifiée (distance supérieure). Slade (2005) propose une estimation non paramétrique de tels poids définis sous une forme fonctionnelle $f(d)$.
- Des matrices tenant compte de la force des relations entre les zones, par exemple le pourcentage de frontières communes. Le poids de voisinage entre deux zones i et j peut ainsi être défini par $b_{ij}^\alpha / d_{ij}^\beta$ avec b_{ij} une mesure de la force des relations entre les zones i et j (qui n'est pas forcément symétrique) telles que le pourcentage de frontières communes, la population, la richesse et d_{ij} la distance entre les zones. Le choix de $\alpha = 1$ et $\beta = 2$ correspond à un modèle du type gravitaire.
- Les k plus "proches" voisins notamment lorsque les zones sont de taille hétérogène, une mesure de la proximité sociale entre agents (être dans la même classe par exemple) ou entre produits...

Les matrices de voisinage sont considérées exogènes dans la majorité des applications d'économétrie spatiale (Anselin 2002). Il ne faut donc pas créer des poids de voisinage qui seraient fonctions du phénomène qu'on cherche à expliquer.

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

3.3 Les méthodes exploratoires

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

Le principal indicateur [3] est celui de Moran qui mesure l'association globale, $I = \frac{N}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij}(y_i - \overline{y})(y_j - \overline{y})}{\sum_i (y_i - \overline{y})^2}$ avec w_{ij} le poids de la i-ème ligne et j-ème 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 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 qu'un tel modèle 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, ε i.i.d. $\sim \mathcal{N}(0, \sigma^2)$ 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, il 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

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

fortes ou faibles valeurs de y ou au contraire des valeurs très diverses. Pour chaque zone i, on calcule un indicateur de Moran local $I_i = N \cdot y_i \cdot \frac{\sum_j w_{ij}(y_j - \overline{y})}{\sum_i (y_i - \overline{y})^2}$ qui sous l'hypothèse nulle d'absence d'autocorrélation spatiale locale tend asymptotiquement vers une loi normale. Cet indicateur permet donc d'identifier les régions d'un intérêt particulier, des points "chauds" (regroupement de fortes valeurs de y) et "froids" (regroupement de faibles valeurs de y), ainsi que des localisations atypiques (à l'aide du graphique de Moran présenté dans l'application pratique). Pour plus de détails sur l'approche descriptive et les indicateurs associés, nous renvoyons à Floch (2012).

4 Estimer un modèle d'économétrie spatiale

4.1 La galaxie des modèles d'économétrie spatiale

Elhorst (2010) a établi une classification des principaux modèles d'économétrie spatiale, en s'appuyant sur les 3 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 [4] :

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

4. Par souci de simplification, la constante du modèle est ici incluse dans la matrice des variables explicatives X. Dans le cas d'une matrice de 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.

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

Avec les paramètres β pour les variables explicatives exogènes, ρ pour l'effet d'interaction endogène (de dimension 1) dit autorégressif spatial, θ pour les effets d'interaction exogène (de dimension le nombre de variables exogènes K) et λ pour 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 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 3 interactions spatiales. Il y aurait par exemple des relations de voisinage définies par W_ρ pour le paramètre autorégressif et W_λ pour l'autocorrélation spatiale. Slade (2005) définit ainsi trois 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 est de 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 qu'au contraire tout le monde soit le voisin de tout le monde. On sait de plus 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 Prucha 2010 pour le modèle hétéroscédastique) :

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

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

Les estimateurs β du modèle de Kelejian-Prucha présentent le défaut d'être biaisés et non convergents si le vrai modèle inclut des interactions exogènes WX (LeSage et Pace 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 faut savoir que par ailleurs il existe des versions plus générales de ces modèles, qui autorisent les effets de voisinage à varier selon l'ordre de voisinage ou selon les interactions prises en compte. Ils correspondent à des versions spatiales des modèles temporels ARMA(p,q).

Dans le cadre d'une étude économique, l'ensemble de ces modèles ne seront bien sûr pas présentés. L'analyse se concentrera sur un modèle, en cohérence avec la question économique et les critères statistiques.

4.2 Critères statistiques du choix de modèle

Deux approches principales ont été utilisées pour le choix des modèles (cf. annexe 1 pour une représentation graphique). Ces approches "pratiques" supposent que la matrice de voisinage soit connue et que les variables explicatives soient exogènes. Sous l'hypothèse de normalité des résidus ε i.i.d. $\sim \mathcal{N}(0, \sigma^2)$, elles s'appuient sur une estimation par maximum de vraisemblance des modèles et les tests statistiques associés [5]. La première dite "approche ascendante" ou bottom-up consiste à partir du modèle non spatial (Le Gallo 2002 pour une synthèse). Des tests du multiplicateur de Lagrange (Anselin *et al.* 1996 pour des tests de spécification des modèles SAR et SEM, robustes à la présence d'autres types d'interactions spatiales) permettent ensuite de trancher entre le modèle SAR, SEM ou le modèle non spatial. Cette approche a été celle plébiscitée jusqu'aux années 2000 car les tests développés par Anselin *et al.* (1996) s'appuient sur les résidus du modèle non spatial. Ils sont donc peu coûteux d'un point de vue computationnel. Florax *et al.* (2003) 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. La deuxième dite "approche descendante" ou top-down consiste à partir du modèle spatial de Durbin. Avec des tests du rapport de vraisemblance, on en déduit le modèle le plus adapté aux observations. L'amélioration des performances informatiques a permis de rendre aisée l'estimation de ces modèles plus complexes, dont le modèle spatial de Durbin pris comme référence dans le livre de LeSage et Pace (2009).

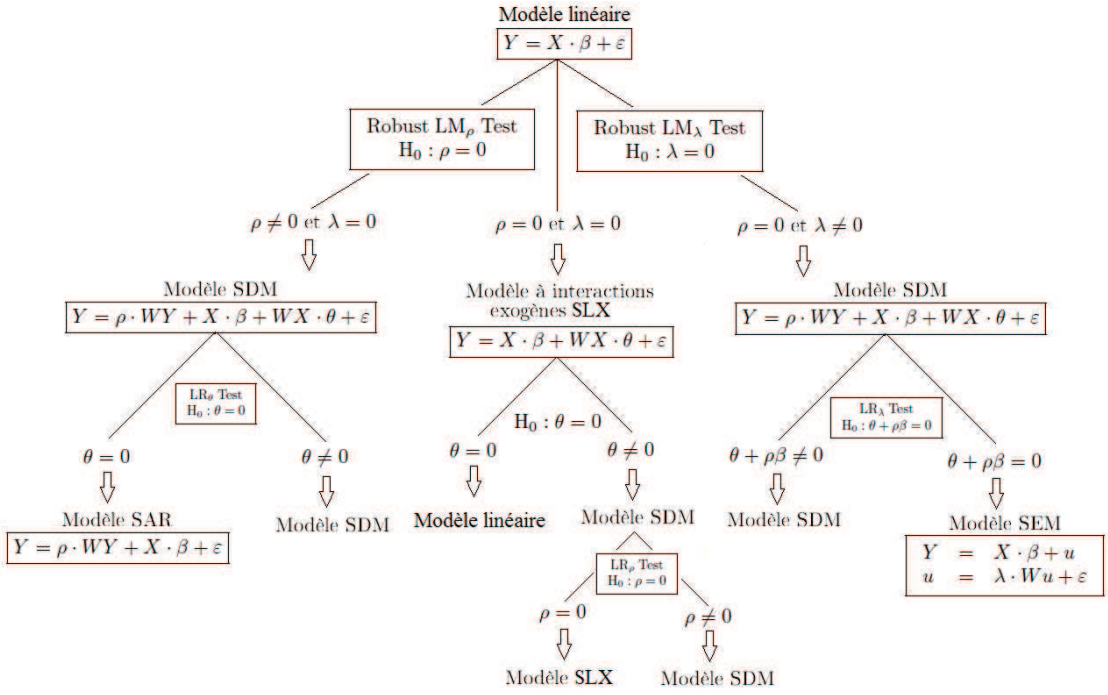
Elhorst (2010) propose une approche "mixte" représentée en figure 1. Elle consiste à commencer par l'approche ascendante mais, en cas d'interactions spatiales ($\rho \neq 0$ ou $\lambda \neq 0$), au lieu de choisir directement un modèle SAR ou SEM, à étudier le modèle spatial de Durbin. Cela permet de 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) [6] concluent à la présence d'un terme autorégressif, i.e. $\rho \neq 0$ et $\lambda = 0$ (branche de gauche de la figure 1). On estime alors le modèle SDM. À l'aide d'un test du rapport de vraisemblance ($\theta = 0$), on peut

5. D'autres méthodes d'estimation existent. Dans le cas de variables explicatives endogènes, Fingleton et Le Gallo (2008 et 2012) proposent une estimation par variables instrumentales et la méthode des moments généralisée. LeSage et Pace (2009) proposent une estimation bayésienne. Enfin, pour relâcher le cadre paramétrique, Lee (2004) propose une estimation par quasi maximum de vraisemblance.

6. 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).

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 du graphique 2), un test du rapport de vraisemblance de l'hypothèse de facteur commun ($\theta = -\rho\beta$) permet de choisir entre le modèle SEM et le modèle SDM. Dans le cas où les tests soulignent l'absence 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.

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



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 intervenir des déterminants incluant cette matrice, et la matrice de variance-covariance des inverses matricielles. Le coût computationnel peut donc être important avec un nombre élevé d'observations. LeSage et Pace (2009) consacrent ainsi un chapître aux enjeux computationnels (et 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 [7], 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.

4.3 L'interprétation des résultats : attention aux rétroactions

L'économétrie spatiale s'écarte du cadre habituel des modèles linéaires lorsque des variables spatialement décalées $W \cdot Y$ sont présentes dans le modèle. L'interprétation classique des
modèles linéaires reste par contre valide si seule une autocorrélation spatiale des erreurs est prise en compte (modèle SEM). Pour expliquer les différents indicateurs associés à l'interprétation des autres modèles, nous reprenons le cadre de LeSage et Pace (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 :

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

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

La valeur prédite est donc $\hat{y} = (1-\hat{\rho}W)^{-1} X\hat{\beta}$ [8] 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

7. 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).

8. Ce n'est pas la prédiction optimale, Goulard *et al.* (2014) pour la prédiction optimale d'un modèle SAR.

direction à prendre en compte (y_t dépend de y_{t-1} qui n'est expliquée que par des valeurs passées), l'économétrie spatiale est multidirectionnelle. Une modification de mon territoire impacte mes voisins, ce qui m'impacte en retour. Il faut en tenir compte pour l'analyse globale des résultats.

Par ailleurs, l'effet marginal apparait différent pour chaque zone [9]. Les termes diagonaux de la matrice S_r sont les effets directs, pour chaque zone, d'une modification de la variable X_r dans la même zone. Les autres termes représentent des effets indirects, i.e. l'impact de la modification de la variable X_r dans une zone sur une autre zone. Au niveau global, plusieurs indicateurs peuvent donc être calculés pour synthétiser les résultats (LeSage et Pace 2009) :

- L'effet direct moyen correspond à la moyenne des termes diagonaux de la matrice S_r, i.e. $\frac{1}{n}\text{trace}\left(S_r\right)$. 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, $\frac{1}{n}\sum_i \left[\sum_k S_r\left(W\right)_{ik}\right]$. 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. $\sum_k S_r\left(W\right)_{ik}$ (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. $\sum_k S_r\left(W\right)_{ki}$ (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.

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 $S_r\left(W\right)$ s'écrit en effet dans ce cas $\left(1 - \rho W\right)^{-1}\left(I_n\beta_r + W\theta_r\right)$, au lieu de $\left(1 - \rho W\right)^{-1}\beta_r$ dans le cas du modèle SAR.

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

Dans tous les cas, le calcul de la précision de ces estimateurs est complexe. LeSage et

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

Pace (2009) s'appuient ainsi sur des simulations bayésiennes MCMC [10]. Il est clair également que 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 $(1-\rho W)^{-1} = (I_n + \rho W + \rho^2 W^2 + \cdots)$ 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...) est également developpée par LeSage et Pace (2009).

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 ($\frac{1}{n} trace\,(S_r) - \hat{\beta}_r$) permet également d'illustrer la force des effets de rétroaction.

5 Limites et difficultés économétriques

5.1 Que faire des données manquantes ?

En économétrie classique, on observe un échantillon de n individus. Si quelques individus présentent des valeurs manquantes, ils sont 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 oeuvre des méthodes économétriques.

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

10. 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.

avec erreurs [11], ou d'utiliser une estimation adaptée (par exemple algorithme EM espérance-maximisation, Wang et Lee 2013 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 d'enquêtes. On observe en effet dans ce cas uniquement 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. Lardeux et Merly-Alpa (2016) montrent qu'il 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 sinon être estimés.

5.2 Le choix de la matrice de poids

Pour définir une matrice de voisinage, les contraintes sont fortes, puisque l'on recherche une description simple (afin que le modèle soit identifiable), mais adéquate des relations entre territoires. La majorité des auteurs soulignent la sensibilité des résultats au choix de cette matrice (Corrado et Fingleton 2012 ; Harris *et al.* 2011), alors que LeSage et Pace (2012) estiment 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 Overman 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.

11. L'interpolation peut également être utile lorsque les niveaux géographiques servant à mesurer la variable à expliquer et les variables explicatives sont différents, 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 la localisation diffère.

Alstadt et Getis (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 travers les modèles économétriques avec des contraintes fonctionnelles *a priori* faibles (Bhattacharjee et Jensen-Butler 2013). Ces dernières approches sont souvent lourdes en calcul et plus difficiles à implémenter. De plus, on constate facilement qu'une description plus réaliste et plus conforme à la réalité économique risque d'introduire de l'endogénéité. Des travaux faisant intervenir des matrices endogènes ont été récemment proposés (Kelejian et Piras 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.

5.3 Et si le phénomène est hétérogène spatialement ?

Deux formes d'hétérogénéité existent.

La première est l'hétéroscédasticité. Les paramètres du modèle sont les mêmes mais pas sa variabilité individuelle. Une autocorrélation spatiale des erreurs $(I - \lambda W)^{-1}\varepsilon$ (modèle SEM) peut ainsi 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 (Le Sage et Pace 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 variance-covariance (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 Prucha (2007 et 2010) 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 n'ont plus les lois usuelles. Les tests d'autocorrélation spatiale ne sont ainsi 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 dispose d'une bonne connaissance du 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 venues du monde de la géographie ont été développées (Le Gallo 2004). Elles restent en grande partie descriptives et exploratoires (notamment à travers des représentations graphiques), car leur comportement théorique n'est pas complètement connu, notamment la convergence et la prise en compte des ruptures géographiques.

Il existe des méthodes de lissage géographique où la constante (voire chaque variable explicative) est croisée avec des polynômes des coordonnées géographiques. Flachaire (2005) propose un modèle linéaire partiel (et alternatif) $Y_i = X_i\beta + f(u_i, v_i) + \varepsilon_i$ et f 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 régressions locales dont l'application spatiale est la régression géographique pondérée (GWR, Geographically Weighted Regression, Brunsdon *et al.* 1996).

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

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

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

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

Les estimateurs sont obtenus, à l'instar de la méthode des MCO, en minimisant une somme pondérée des carrés des résidus, avec n le nombre d'observations.

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

Comme dans toutes les méthodes utilisant des noyaux, le nombre d'observations dont le poids est non nul (i.e. le voisinage pris en compte dans l'estimation) a un impact beaucoup plus important que la forme fonctionnelle utilisée pour définir les poids $w_k(u_i, v_i)$. Il est donc important de déterminer ce voisinage appelé classiquement "fenêtre" et noté h. Plusieurs méthodes ont été proposées, dont la plus utilisée est celle de la validation croisée. On cherche la valeur de h qui minimise la quantité $\Delta(h) = \sum_{i=1}^{n} (y_i - \widehat{y}_{\neq i}(h))^2$ où $\hat{y}_{\neq i}(h)$ est la valeur prédite de y_i obtenue en utilisant la fenêtre h et en éliminant l'observation i. Cette méthode est sensible à la multicolinéarité (locale) des observations.

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

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

Pour chaque point d'estimation (u_i, v_i), plusieurs observations peuvent avoir un poids nul.

Il reste 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. A 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 unique modélisation.

5.4 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 ainsi 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 connue sous l'acronyme MAUP ("Modifiable Areal Unit Problem"). Il a plusieurs conséquences (Floch 2012). Avec des découpages ou des échelles différentes, les résultats des modèles et les interactions entre agents ne sont pas identiques. Il faut également tenir compte de l'étendue spatiale des zones, 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 $\overline{x}_k = \sum_i x_{ik}/n_k$. Le modèle s'écrit $y_{ik} = \alpha + \beta \cdot x_{ik} + \gamma \cdot \overline{x}_k + \varepsilon_{ik}$ où β représente l'effet individuel et γ l'effet de contexte. Si 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 $\overline{y}_k = \alpha + (\beta + \gamma) \cdot \overline{x}_k + \overline{\varepsilon}_k$. Il n'est alors plus possible d'identifier séparément les paramètres β et γ. Le modèle est hétéroscédastique car $\mathbb{V}(\overline{\varepsilon}_k) = \sigma^2/n_k$ dans le cas de perturbations initiales i.i.d. de variance σ^2.

Le problème est encore plus complexe dans le cas de modèles spatiaux. Il n'est en effet pas possible d'agréger une matrice de voisinage W définie au niveau individuel. Avec des données individuelles, un individu i du groupe k peut avoir des voisins parmi le groupe k

mais également parmi un autre groupe k'. Si on considère désormais une matrice de voisinage agrégée au niveau groupe, les relations intra-groupes ne seront plus 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'échelle peuvent engendrer des résultats à première vue paradoxaux.

En pratique, cela signifie que l'interprétation des résultats n'est valable que pour le découpage géographique choisi. Si on étudie des relations économiques à un niveau agrégé avec un modèle spatial, on ne peut rien dire des relations individuelles entre agents. Pour tenir compte de cette imbrication des zones géographiques (régions, départements, cantons, individus) et rendre les analyses cohérentes entre elles, une solution est alors de mener des analyses multi-niveaux (Givord et Guillerm 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 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 (% des peu diplômés et des moins de 30 ans dans la population active), de la structure économique (% des emplois dans le secteur industriel et de l'emploi public) et du marché du travail (taux d'activité). L'objectif de ce modèle est descriptif et illustratif. Il ne sera pas de détailler les résultats d'une étude économique mais d'illustrer les techniques mises en oeuvre : la définition d'une matrice de voisinage qui décrit les relations de proximité entre 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 Hild 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 au millésime 2011 de CLAP (Connaissance locale de l'appareil productif) et du RP (Recensement de la Population). Une interprétation causale reste néanmoins impossible. De nombreuses variables ont en effet été omises de l'analyse, par exemple 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 :

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

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

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

6.1 La gestion des données géographiques

Le logiciel R fournit des commandes permettant de mettre en œuvre les modèles d'économétrie spatiale de manière simple. En pratique, toute étude sur des données géographiques nécessite un travail préalable de mise en forme de ces données. Deux types de bases de données sont en effet utilisés :

- Le fichier de définition des zones géographiques et de ses attributs (points, lignes, polygones...) sous format *shapefile* (principaux SIG), *MIF/MID* (MapInfo) ou *KML* (Google

Earth) ;
- La base de données classique des variables explicatives, avec un identifiant pour les zones géographiques.

Le premier fichier correspond donc à une carte qui permettra de visualiser les résultats obtenus (valeurs prédites par les différents modèles, résidus...) mais également de constituer la matrice de poids à l'aide d'outils disponibles sous R. Il est donc nécessaire de disposer d'un fonds cartographique correspondant au maillage territorial utilisé dans l'étude. Pour pouvoir effectuer les traitements présentés dans ce document, nous utilisons [12] la base de données au format csv (fichier donnees_ze.csv) et une carte des zones d'emploi dans un format réutilisable dans R (fichiers Zempl_2010.shp, Zempl_2010.shx et Zempl_2010.dbf). L'importation de cette carte est effectuée sous R à l'aide de la commande *readOGR* du package *rgdal* (ou de la commande *readShapeSpatial* du package *maptools*). Le package *rgdal* permet la lecture de données vectorielles (sous la forme de points, de lignes ou de polygones). Tous les fichiers de définition de la carte doivent être placés dans le répertoire de travail. L'importation des données statistiques se fait à l'aide des outils classiques d'importation *read.table*, *read.csv*.

```
### Importation des données
> donnees_ze=read.csv("donnees_dt_eco_spatiale.csv",
colClasses=c('character','character',rep('numeric',32)))
### Importation du fonds de carte
# Solution 1 avec readOGR
> carte_ze<-readOGR(dsn="chemin",layer="zempl_2010")
> carte_ze<-carte[carte_ze@data$Reg!="94",]
# Solution 2 avec readShapePoly
> carte_ze <- readShapePoly('chemin/zempl_2010.shp')
### Affichage de la carte de France
> plot(carte_ze,cex=.01)
```

L'une des principales difficultés est que les formats des données géographiques ne sont pas toujours les mêmes selon les logiciels. L'importation et la lecture de données géo-localisées peuvent donc être complexes, ce qui dépasse le cadre de ce document. On pourra se référer à Bivand *et al.* 2013 pour des détails sur l'utilisation de données géographiques sous R. L'annexe 2 détaille les caractéristiques des données géographiques et leur gestion.

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

12. Ces fichiers, ainsi que les programmes, sont mis à disposition des lecteurs sur la page personnelle de Ronan Le Saout.

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

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

6.2 Définir une matrice de voisinage

Pour définir une matrice de contiguïté, l'objet géographique, de type polygone dans notre exemple, doit être transformé en un objet de type .nb (pour neighbours), qui contiendra la liste des relations de voisinages. Cette opération est effectuée à l'aide de la commande *poly2nb*, du package *spdep*, lorsqu'on prend un critère de contiguïté. La relation de contiguïté est du type "Tour" par défaut (Floch 2012 pour la description de ces types).

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

L'objet carteC.nb contient 1 582 liens. Parmi les zones d'emploi, il y a en moyenne un peu plus de cinq liens. 67 zones (sur 297 zones d'emploi) ont 5 voisins. Quatre ont un seul lien. Les objets .nb fournissent un moyen de stockage de l'information. Les liens peuvent être visualisés sous la forme d'un graphe, construit autour des centroïdes des zones.

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

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

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

```
### Récupération des centroides
> coor<-coordinates(carte_ze)
### Définition des 10 plus proches voisins
#   k indique le nombre de voisins
> cartePPV10.knn<-knearneigh(coor,k=10)
> cartePPV10.nb<- knn2nb(carte.knn)
### Visualisation des relations de voisinage
> plot(cartePPV10.nb, coor, col="red",cex=0.1,add=TRUE)
### Matrice de voisinage des 10 plus proches voisins
> PPV10.w<-nb2listw(cartePPV10.nb,style="W")
```

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

```
### Matrice de voisinage fondée sur la distance
> distance<-rdist(coor,coor)
> diag(distance)<-0
> distance[distance>=100000]<-0
> dist<- 1.e12 %/% (distance*distance)
> dist[dist>1.e15]<-0
> dist.w<-mat2listw(dist, row.names = NULL, style="W")
```

Dernier exemple, il est possible de créer une matrice de voisinage spécifique à la question posée. C'est par exemple le cas pour des données non géographiques (proximité sociale, entre produits industriels...). Nous définissons une dernière matrice de voisinage, dont la pondération est proportionnelle aux déplacements domicile-travail entre les zones d'emploi. Il y a en moyenne 9 liens de voisinage avec cette matrice. Le code de création de cette matrice est fourni en annexe 3. Nous la nommerons matrice endogène par la suite car les poids de voisinage sont *a priori* corrélés aux variables explicatives du modèle. Les déplacements domicile-travail sont en rapport avec le niveau de l'emploi et du chômage (et servent même à construire le découpage territorial). Cette matrice a été construite dans un but pédagogique, pour notamment illustrer le caractère exogène ou endogène des relations de voisinage. L'utilisation d'une matrice endogène, avec les méthodes présentées ici, est à proscrire dans le cadre d'une étude économique.

On trouvera ci-dessous une représentation graphique des zones de voisinage pour Lyon. Pour les matrices endogène et celle liée à la distance, les pondérations plus faibles sont représentées en jaune. Ces cartes permettent de constater que le choix de la matrice de voisinage n'est pas neutre. Avec une matrice endogène, la région parisienne devient ainsi très proche de la zone d'emploi de Lyon.

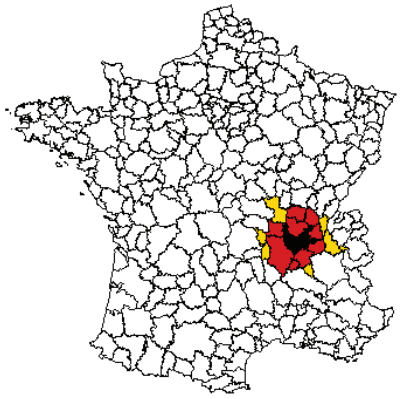
Cartes 1 : Matrices de voisinage autour de la zone d'emploi de Lyon

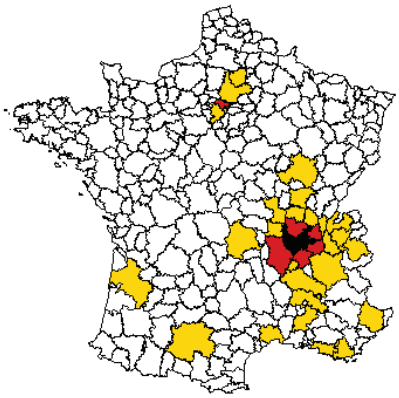
contiguïté

Distance

5 plus proches voisins

Endogène





6.3 Cartographie et tests

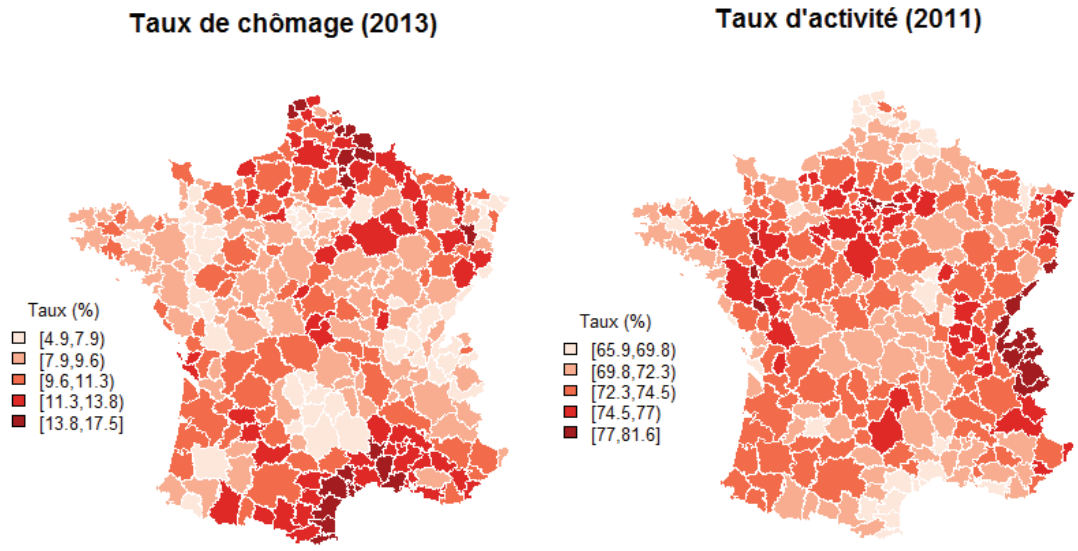
Avant de mettre en place un modèle d'économétrie spatiale, il convient de vérifier qu'il y a bien un phénomène spatial à prendre en compte. Cela commence par une caractérisation de l'autocorrélation spatiale à l'aide de représentations graphiques (carte) et de tests statistiques (Moran). Il est utile,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

6.3.1 Cartographie

L'objectif n'est pas ici de présenter les nombreuses techniques cartographiques disponibles dans R, mais de donner quelques exemples utiles. Les codes des cartes présentées sont fournis en annexe 2.

La carte 2 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.

Carte 2 : Distribution du taux de chômage et d'activité, par zone d'emploi



Le tableau 1 décrit la distribution des variables. Le taux de chômage moyen est de 10%, pour un taux d'activité de 73%. Il y a 22% d'actifs peu diplômés et de jeunes actifs de moins de 30 ans. 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.

Tableau 1 : Descriptif de l'échantillon

	N	Moyenne	Ecart-Type	Min	Q25	Médiane	Q75	Max
Taux de chômage	297	10.0	2.4	4.9	8.3	9.6	11.4	17.5
Taux d'activité	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

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

6.3.2 Tests d'autocorrélation spatiale et représentations graphiques avancées

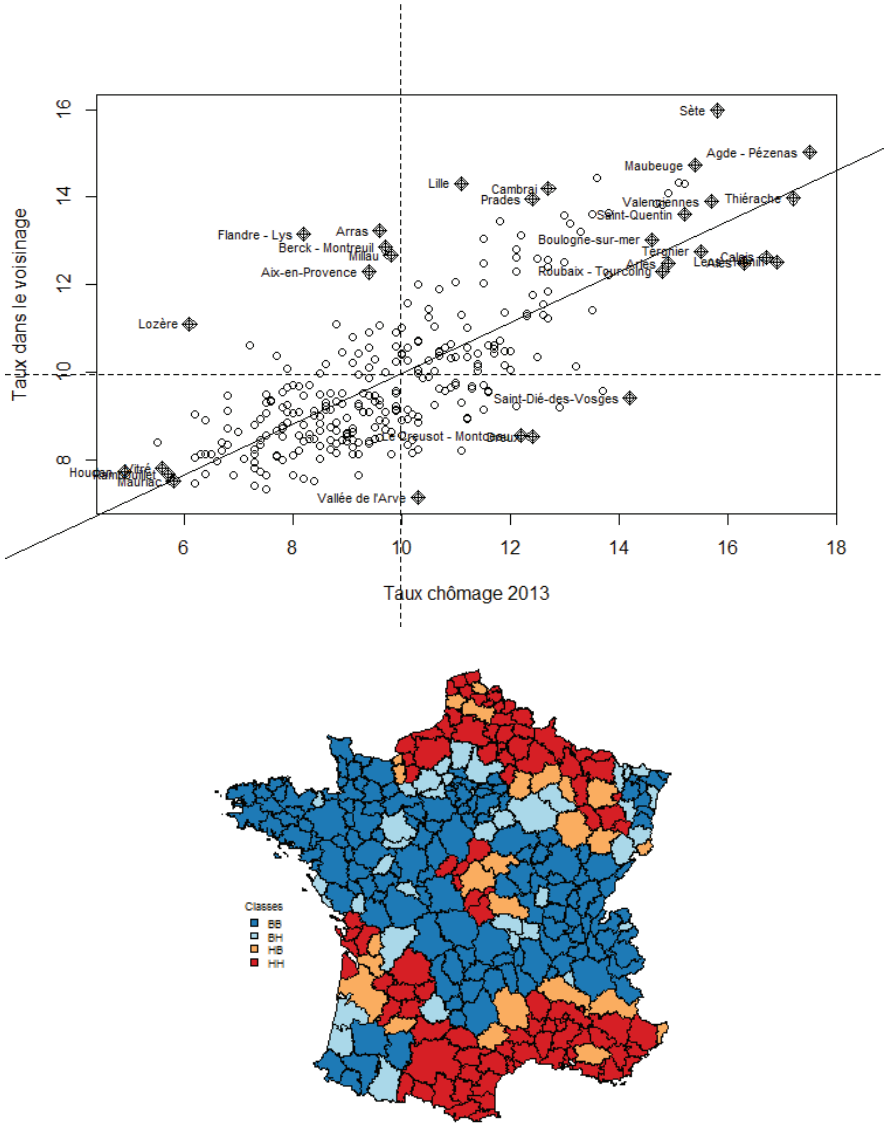
Le package *spdep* fournit des commandes permettant de calculer les indicateurs de Moran (et autres indicateurs spatiaux) pour les valeurs brutes, au niveau global ou local. La commande *moran.test(donnees_ze$txcho_2013, dist.w)* effectue un test de Moran pour le taux de chômage en utilisant la matrice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

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

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

La figure 2 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 2 : Graphique de Moran du taux de chômage et carte associée



6.4 Estimation et choix de modèles

L'analyse descriptive a permis de constater que l'espace n'était pas neutre pour caractériser les taux de chômage locaux. Il n'est néanmoins pas certain qu'un modèle économétrique tenant compte de l'espace soit nécessaire. Le nuage de points des taux de chômage et d'activité montrent une forte relation linéaire des 2 variables. Le taux de chômage et d'activité sont tous les deux corrélés spatialement. Le taux de chômage pourrait donc être relié au taux d'activité, sans autre forme 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 d'Elhorst (2010) conduirait à retenir le modèle SDM. Seuls les modèles MCO et SDM seraient alors estimés. Dans un but pédagogique, l'ensemble des modèles spatiaux 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 distance inverse, car c'est celle qui présente le caractère explicatif le plus fort (AIC les plus faibles) et dont l'interprétation économique est la plus intuitive. Les zones d'emploi n'ayant pas la même taille, la 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 terme 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.

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

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	MCO	SEM	SAR	SDM	SAC	SLX	SDEM	Manski
Taux d'activité	-0.622***	-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				

Note de lecture : 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 : * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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

- L'approche séquentielle d'Elhorst (partie 4.2) conduirait à retenir un modèle SDM (col.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 (col.1). De même, l'hypothèse de facteur commun du modèle SDM est rejetée (p-value : 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, col.6), on ne rejette pas l'hypothèse d'absence d'autocorrélation résiduelle sous l'hypothèse de corrélation endogène (Test Robuste LM-Error, p-value=0.787).
- Le choix d'un modèle SAR (col.3) serait ici déconseillé. Un test montre qu'une autocorrélation spatiale résiduelle reste présente (p-value (test LM residual auto.) : 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 (col.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, col.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 (tableau 2, col.2) voire SDEM (col.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 (tableau 3, col.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 amèner 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 (tableau 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 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 (tableau en annexe 4), 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.

Tableau 3 : Modèle SEM, pour différentes matrices de voisinage

	(1) MCO	(2) SEM Contiguité	(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

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

6.5 Interprétation des résultats

Pour le modèle SDM, afin de permettre une interprétation au regard du modèle MCO et SEM, on calcule les effets directs et indirects tels que décrits dans la partie 4.4. Les intervalles de confiance empiriques sont obtenus à l'aide de 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)
```

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

	(1) MCO	(2) SDM Contiguité	(3) SDM 2 Voisins	(4) SDM 5 Voisins	(5) SDM 10 Voisins	(6) SDM Distance	(7) SDM Endogène
Taux d'activité	-0.622	-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]

Note de lecture : 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.

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

	(1) SDM Contiguité	(2) SDM 2 Voisins	(3) SDM 5 Voisins	(4) SDM 10 Voisins	(5) SDM Distance	(6) SDM Endogène
Taux d'activité	-0.323	-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]

Note de lecture : 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.

6.6 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 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 (via le package *spgwr*).

Nous allons détailler la mise en oeuvre d'une analyse géographique pondérée, à partir du même modèle que précédemment. La première étape, réalisée à l'aide de la commande *gwr.sel*, consiste à déterminer pour une fonction de pondération donnée, gaussienne ou biweight, la valeur optimale de la fenêtre (i.e. le nombre de voisins ou la distance pris en compte). On doit introduire dans la commande les coordonnées, qui peuvent être récupérées à partir de la carte. Pour la méthode utilisée, on a le choix entre *cv* (validation croisée) et *aic* (critère

d'information d'Akaike). La fonction de pondération est selon le cas *gwr.bisquare* (biweight) ou *gwr.gauss* (normale). Par défaut, cette commande utilise la fonction gaussienne et la validation croisée. La fonction crée un objet qui contient les informations calculées et qui sera réutilisé dans la commande *gwr*. La taille de la fenêtre diffère, selon la fonction de pondération et le critère choisis, du simple au quintuple. On a choisi pour l'exemple la fenêtre obtenue avec le noyau gaussien, et le critère de validation croisée.

```
### Estimation géographique pondérée du modèle linéaire
## Calcul de la fenêtre d'estimation optimale
> h.bw <- gwr.sel(modele, data=donnees_ze, coords=coor, method="cv", gweight=gwr.Gauss)
## Estimation du modèle
> modele.gwr <- gwr(modele, data=donnees_ze, coords=coor, bandwidth=h.bw, hatmatrix=TRUE)
```

Le tableau 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.

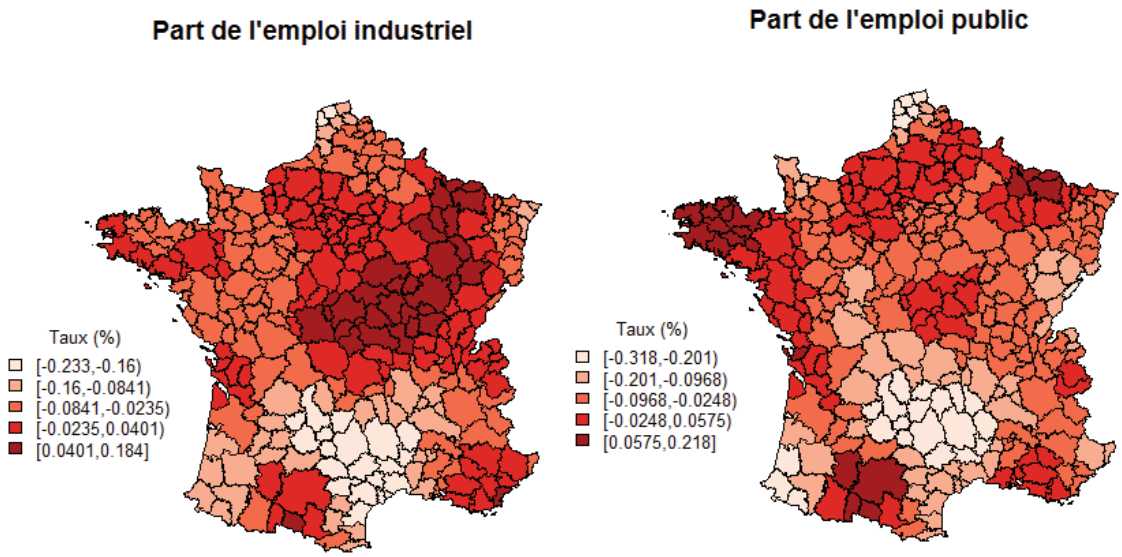
Tableau 6 : Résultats de la régression géographique pondérée

	(1)	(2)	(3)	(4)	(5)	(6)
	MCO	Minimum	P1	Médiane	P3	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

On récupère en sortie de la commande *gwr* 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 (carte 3) 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...) 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.

Carte 3 : Distribution des paramètres locaux



Le package *spgwr* propose des tests (avec les commandes *gwr.test*) permettant de tester 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, permettant notamment de visualiser des phénomènes de non-stationnarité. Mais elle a fait aussi l'objet d'un certain nombre de critiques. Wheeler et Paez (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.

7 Conclusion

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

Bibliographie

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

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

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

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

Anselin, Luc. (2001) "Spatial Effects in Econometric Practice in Environmental and Resource Economics." *American Journal of Agricultural Economics*, 83(3), 705-710.

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

Arbia, Giuseppe. (2014) "A Primer for Spatial Econometrics - With Applications in R." Palgrave Macmillan Publisher, Palgrave Texts in Econometrics Series.

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, 27-44.

Barrios, Thomas, Rebecca Diamond, Guido W. Imbens, et Michal Kolesar. (2012) "Clustering, Spatial Correlations, and Randomization Inference." *Journal of the American Statistical Association*, 107(498), 578-591.

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

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

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

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

Brunsdon, Chris, A. Stewart Fotheringham, et Martin E. Charlton. (2002) "Geographically Weighted Regression : The Analysis of Spatially Varying Relationships." Wiley.

Brunsdon, Chris, Martin E. Charlton, Isabella Gollini, Paul Harris, et Binbin Lu. (2015) "GWmodel : an R Package for Exploring Spatial Heterogeneity using Geographically Weighted Models." *Journal of statistical software*, 63(17).

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

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

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

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

Fafchamps, Marcel. (2015) "Causal Effects in Social Networks." *Revue Economique*, 66(4), 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), 319-339.

Fingleton, Bernard, et Julie Le Gallo. (2012) "Endogénéité et Autocorrélation Spatiale : Quelle Utilité pour le Modèle de Durbin ?" *Revue d'Économie Régionale & Urbaine*, 1, 3-17.

Flachaire, Emmanuel. (2005) "The Role of Economic Space in Decision Making : Comment." *Annals of Economics and Statistics*, 77, 21-28.

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

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

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

Givord, Pauline, et Marine Guillerm. (2016) "Les modèles multiniveaux : principes et pratiques" Document de travail INSEE, M2016/05.

Goulard, Michel, Laurent Thibault, et Christine Thomas-Agnan. (2013) "About Predictions in Spatial Autoregressive Models : Optimal and Almost Optimal Strategies." TSE Working Papers 13-452, Toulouse School of Economics.

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

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

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

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

Kelejian, Harry H., et Ingmar R. Prucha. (2007) "HAC Estimation in a Spatial Framework." *Journal of Econometrics*, 140(1), 131-154.

Kelejian, Harry H., et Ingmar R. Prucha. (2010) "Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances." *Journal of Econometrics*, 157(1), 53-67.

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

Lardeux, Raphaël, et Thomas Merly-Alpa. (2016) "Spatial Econometrics on Survey Data." Mimeo.

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

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

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

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, 9-32.

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

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

LeSage, James P. (2014) "What Regional Scientists Need to Know about Spatial Econometrics." *The Review of Regional Studies*, 44(1), 13-32.

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

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

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

Osland, Liv. (2010) "An Application of Spatial Econometrics in Relation to Hedonic House Price Modeling." *Journal of Real Estate Research*, 32(3), 289-320.

Piras, Gianfranco. (2010) "Spatial Models with Heteroskedastisc Innovations in R." *Journal of Statistical Software,* 35(1).

Slade, Margaret E. (2005) "The Role of Economic Space in Decision Making." *Annals of Economics and Statistics*, 77, 1-20.

Waelbroeck, Patrick. (2005) "The Role of Economic Space in Decision Making : Comment." *Annals of Economics and Statistics*, 77, 29-31.

Wang, Wei, et Lung-Fei Lee. (2013) "Estimation of Spatial Autoregressive Models with Randomly Missing Data in the Dependent Variable." *The Econometrics Journal*, 16(1), 73-102.

Wheeler David C., et Antonio Paez. (2009) "Geographically Weighted Regression" in Handbook of Spatial Analysis, Fischer & Getis Eds., pages 461-486.

Annexes

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

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

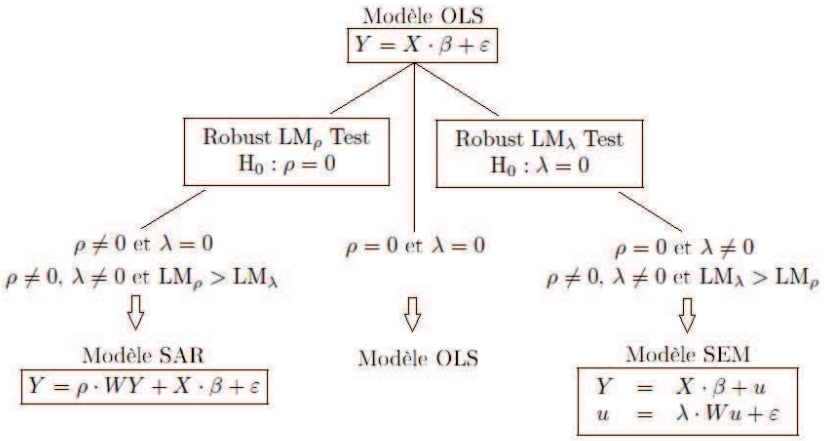
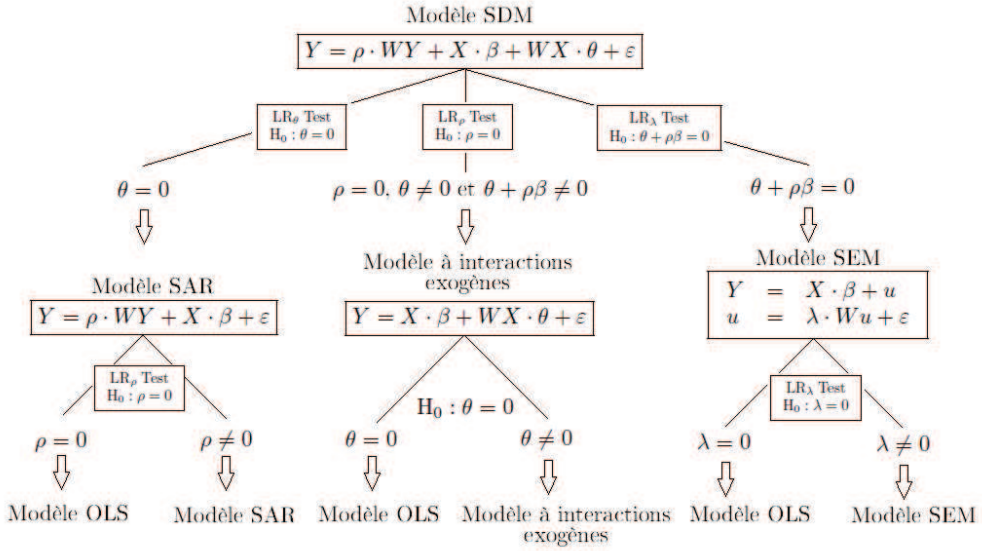


Figure A1.2 : Approche top-down, à partir de LeSage et Pace (2009)



Annexe 2 : Fonds de carte et cartographie sous R

Les données géographiques peuvent être représentées sous forme vectorielle (point, ligne ou polygone) et matricielle (objet nommé raster, par exemple des grilles de pixels). Des exemples sont définis ci-dessous pour chacun de ces objets géographiques. Les zones d'emploi, exemple de ce présent document, sont des polygones. Un exemple de point correspond aux coordonnées de bâtiment, par exemple les gares routières. Le réseau hydrographique correspond lui à des lignes. Une décomposition de la presqu'île de Crozon (Bretagne) en une grille de pixels est au contraire une approche matricielle, moins usitée en économétrie spatiale.

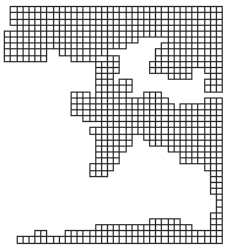
Figure A2 : Représentation des objets spatiaux

Des polygones : les zones d'emploi

Des lignes : les fleuves

Des points : les gares

Un raster : la presqu'île de Crozon pixellisée



Les exemples présentés dans ce document nécessitent donc le fond de carte des zones d'emploi. Lorsqu'on n'en dispose pas, des fonds libres de droit peuvent être récupérés. Les

fonds de carte communaux permettent de reconstituer par agrégation les zonages supra-communaux conçus par l'Insee (unités urbaines, zonage en aires urbaines 2010, zones d'emploi 2010), ainsi que les regroupements officiels de commune (Communauté d'Agglomération ou de Communes...). Ils peuvent être exportés depuis Openstreetmap ou l'IGN. Le site de l'Insee consacré aux zonages d'étude diffuse des tables de passage permettant de construire, dans des logiciels cartographiques ou directement dans R de tels agrégats supra-communaux. Des fonds de carte européens sont disponibles sur le site d'Eurostat, des fonds de cartes mondiaux à partir du GADM database of Global Administrative Areas.

On peut alors créer de nouveaux fonds de carte par agrégation en utilisant par exemple la commande *unionSpatialPolygons* du package *maptools* ou la commande *gUnaryUnion* du package *rgeos*. Par exemple, pour constituer une carte de France à partir des zones d'emploi, on ajoute à la carte des zones d'emploi un identifiant permettant l'agrégation *Carte_ze@data$fr <- "FR"*. On crée ainsi une variable supplémentaire associée à la carte. On fusionne ensuite les zones d'emploi à l'aide de la commande *carte_fr <- unionSpatialPolygons(carte_ze, ID=carte@data$fr)*.

Les cartes présentées dans ce document nécessitent les packages *rcolorBrewer* (pour la gestion des couleurs) et *classInt* (pour le partage en classes de la variable cartographiée). Pour une introduction à la cartographie, on peut utilement se référer à "R pour les géographes" du groupe ElementR. Un exemple de code pour la génération d'une carte est fourni ci-dessous.

```
### Cartographie des zones d'emploi
> vPal5 <- brewer.pal(n = 5, name = "Reds")
> carte_ze@data <- data.frame(carte_ze@data, donnees_ze[match(carte_ze@data[,"Codgeo"],
donnees_ze[,"ze2010"]),])
> lJenks2013 <- classIntervals(var = carte_ze@data$txcho_2013, n = 5, style = "jenks")
> vJenks2013 <- lJenks2013$brks
> carte_ze@data$cho <- as.character(cut(carte_ze@data$txcho_2013, breaks = vJenks2013,
labels = vPal5, include.lowest = TRUE, right = FALSE))
> vLegendBoxJ5 <- as.character(levels(cut(carte_ze@data$txcho_2013, breaks = vJenks2013,
include.lowest = TRUE, right = FALSE)))
> plot(carte_ze, col = carte_ze@data$cho, border = "white")
> legend("bottomleft", legend = vLegendBoxJ5, bty = "n",
fill = vPal5, cex = 0.8, title = "Taux (%)")
> title(main="Taux de chômage (2013)")
```

Annexe 3 : 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 4 : Modèle SDM, pour différentes matrices de voisinage

	(1) SDM Contiguité	(2) SDM 2 Voisins	(3) SDM 5 Voisins	(4) SDM 10 Voisins	(5) SDM Distance	(6) SDM Endogène
Taux d'activité	-0.486***	-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

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

Série des Documents de Travail « Méthodologie Statistique »

9601 : Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population.
G. DECAUDIN, J.-C. LABAT

9602 : Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises.
N. CARON, P. RAVALET, O. SAUTORY

9603 : La procédure FREQ de SAS - Tests d'indépendance et mesures d'association dans un tableau de contingence.
J. CONFAIS, Y. GRELET, M. LE GUEN

9604 : Les principales techniques de correction de la non-réponse et les modèles associés.
N. CARON

9605 : L'estimation du taux d'évolution des dépenses d'équipement dans l'enquête de conjoncture : analyse et voies d'amélioration.
P. RAVALET

9606 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT).
S. LOLLIVIER, M. MARPSAT, D. VERGER

9607 : Enquêtes régionales sur les déplacements des ménages : l'expérience de Rhône-Alpes.
N. CARON, D. LE BLANC

9701 : Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?
J.-C. DEVILLE

9702 : Modèles univariés et modèles de durée sur données individuelles.
S. LOLLIVIER

9703 : Comparaison de deux estimateurs par le ratio stratifiés et application aux enquêtes auprès des entreprises.
N. CARON, J.-C. DEVILLE

9704 : La faisabilité d'une enquête auprès des ménages.
1. au mois d'août.
2. à un rythme hebdomadaire

C. LAGARENNE, C. THIESSET

9705 : Méthodologie de l'enquête sur les déplacements dans l'agglomération toulousaine.
P. GIRARD.

9801 : Les logiciels de désaisonnalisation TRAMO & SEATS : philosophie, principes et mise en œuvre sous SAS.
K. ATTAL-TOUBERT, D. LADIRAY

9802 : Estimation de variance pour des statistiques complexes : technique des résidus et de linéarisation.
J.-C. DEVILLE

9803 : Pour essayer d'en finir avec l'individu Kish.
J.-C. DEVILLE

9804 : Une nouvelle (encore une !) méthode de tirage à probabilités inégales.
J.-C. DEVILLE

9805 : Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish.
J.-C. DEVILLE

9806 : Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE.
N. CARON, J.-C. DEVILLE, O. SAUTORY

9807 : Estimation de données régionales à l'aide de techniques d'analyse multidimentionnelle.
K. ATTAL-TOUBERT, O. SAUTORY

9808 : Matrices de mobilité et calcul de la précision associée.
N. CARON, C. CHAMBAZ

9809 : Échantillonnage et stratification : une étude empirique des gains de précision.
J. LE GUENNEC

9810 : Le Kish : les problèmes de réalisation du tirage et de son extrapolation.
C. BERTHIER, N. CARON, B. NEROS

9901 : Perte de précision liée au tirage d'un ou plusieurs individus Kish.
N. CARON

9902 : Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen.
N. CARON

0001 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT) (version actualisée).
S. LOLLIVIER, M. MARPSAT, D. VERGER

0002 : Modèles structurels et variables explicatives endogènes.
J.-M. ROBIN

0003 : L'enquête 1997-1998 sur le devenir des personnes sorties du RMI - Une présentation de son déroulement.
D. ENEAU, D. GUILLEMOT

0004 : Plus d'amis, plus proches ? Essai de comparaison de deux enquêtes peu comparables.
O. GODECHOT

0005 : Estimation dans les enquêtes répétées : application à l'Enquête Emploi en Continu.
N. CARON, P. RAVALET

0006 : Non-parametric approach to the cost-of-living index.
F. MAGNIEN, J. POUGNARD

0101 : Diverses macros SAS : Analyse exploratoire des données, Analyse des séries temporelles.
D. LADIRAY

0102 : Économétrie linéaire des panels : une introduction.
T. MAGNAC

0201 : Application des méthodes de calages à l'enquête EAE-Commerce.
N. CARON

C 0201 : Comportement face au risque et à l'avenir et accumulation patrimoniale - Bilan d'une expérimentation.
L. ARRONDEL, A. MASSON, D. VERGER

C 0202 : Enquête Méthodologique Information et Vie Quotidienne - Tome 1 : bilan du test 1, novembre 2002.
J.-A. VALLET, G. BONNET, J.-C. EMIN, J. LEVASSEUR, T. ROCHER, P. VRIGNAUD, X. D'HAULTFOEUILLE, F. MURAT, D. VERGER, P. ZAMORA

0203 : General principles for data editing in business surveys and how to optimise it.
P. RIVIERE

0301 : Les modèles logit polytomiques non ordonnés : théories et applications.
C. AFSA ESSAFI

0401 : Enquête sur le patrimoine des ménages - Synthèse des entretiens monographiques.
V. COHEN, C. DEMMER

0402 : La macro SAS CUBE d'échantillonnage équilibré
S. ROUSSEAU, F. TARDIEU

0501 : Correction de la non-réponse et calage de l'enquêtes Santé 2002
N. CARON, S. ROUSSEAU

0502 : Correction de la non-réponse par répondération et par imputation
N. CARON

0503 : Introduction à la pratique des indices statistiques - notes de cours
J-P BERTHIER

0601 : La difficile mesure des pratiques dans le domaine du sport et de la culture - bilan d'une opération méthodologique
C. LANDRE, D. VERGER

0801 : Rapport du groupe de réflexion sur la qualité des enquêtes auprès des ménages
D. VERGER

M2013/01 : La régression quantile en pratique
P. GIVORD, X. D'HAULTFOEUILLE

M2014/01 : La microsimulation dynamique : principes généraux et exemples en langage R
D. BLANCHET

M2015/01 : la collecte multimode et le paradigme de l'erreur d'enquête totale
T. RAZAFINDROVONA

M2015/02 : Les méthodes de Pseudo-Panel
M. GUILLERM

M2015/03 : Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse
E. GROS – K.MOUSSALAM

M2016/01 : Le modèle Logit Théorie et application.
C. AFSA

M2016/02 : Les méthodes d'estimation de la précision de l'Enquête Emploi en Continu
E. GROS – K.MOUSSALAM

M2016/03 : Exploitation de l'enquête expérimentale Vols, violence et sécurité.
T. RAZAFINDROVONA

M2016/04 : Savoir compter, savoir coder. Bonnes pratiques du statisticien en programmation.
E. L'HOUR – R. LE SAOUT B. ROUPPERT

M2016/05 : Les modèles multiniveaux
P. GIVORD – M. GUILLERM