7. Économétrie spatiale sur données de panel

BOUAYAD AGHA SALIMA
GAINS (TEPP) et Crest
Le Mans Université
LE GALLO JULIE
CESAER, AgroSup Dijon, INRA,
Université de Bourgogne Franche-Comté, F-21000 Dijon
VÉDRINE LIONEL
CESAER, AgroSup Dijon, INRA,
Université de Bourgogne Franche-Comté, F-21000 Dijon

Résumé

Ce chapitre propose une présentation synthétique des méthodes d'économétrie spatiale appliquées aux données de panel. Nous insistons principalement sur les spécifications et les méthodes implémentées dans le package *splm* disponible sous R. Nous illustrons notre présentation par une analyse de la deuxième "loi" de Verdoorn avant de présenter des extensions récentes des modèles spatiaux sur données de panel.

R La lecture préalable du chapitre 6 : "économétrie spatiale : modèles courants" est recommandée.

Introduction

Les données de panel concernent des observations liées à un ensemble d'individus (firmes, ménages, collectivités locales) observés à plusieurs dates (HSIAO 2014). Relativement aux données en coupe transversale, on considère que le fait de pouvoir disposer d'informations dans les dimensions individuelles et temporelles présente trois avantages principaux. Le gain d'information lié à l'exploitation de la double dimension des données permet de contrôler la présence d'hétérogénéité inobservable. La taille des échantillons généralement plus élevée permet d'améliorer la précision des estimations. Enfin, les données de panel permettent de modéliser des relations dynamiques.

Après une première génération de modèles spatiaux spécifiés pour données en coupe transversale (ELHORST 2014b), de nombreuses applications en économétrie spatiale reposent aujourd'hui sur des données de panel. En effet, si les spécifications a-spatiales sur données de panel permettent effectivement de contrôler une certaine forme d'hétérogénéité inobservée, la dépendance des coupes est prise en compte sans toujours être identifiée ou modélisée et ces modèles ne captent pas le cas particulier des effets de dépendance spatiale. De manière similaire aux modèles en coupe, l'introduction d'effets spatiaux dans les modèles en données de panel permet ainsi de mieux prendre en compte l'interdépendance entre les individus.

Dans ce chapitre, nous présentons les principales spécifications des panels spatiaux, en partant des spécifications standard en données de panel (section 7.1). La section 7.2 est consacrée à la présentation des méthodes d'estimation, et la section 7.3 décrit les principaux tests de spécifications spécifiques aux panels spatiaux. Nous proposons une application empirique en testant la deuxième loi de Verdoorn dans le cadre d'un panel de régions européennes (NUTS3) entre 1991 et 2008 (section 7.4). La section 7.5 présente quelques extensions récentes des panels spatiaux.

7.1 Spécifications

Cette section présente les principales spécifications utilisées pour les modèles statiques sur données de panel avec prise en compte des interactions spatiales. Nous ne considérons que le cas des panels cylindrés : les individus sont observés à toutes les périodes. Les travaux portant sur les méthodes d'estimation sur panels spatiaux non cylindrés sont encore peu développés. Les modèles dynamiques seront brièvement évoqués dans la section 7.5.1. Après un bref rappel de ce qui caractérise les spécifications standard sur données de panel (sans dépendance spatiale) et de ce qui distingue les effets spécifiques fixes des effets aléatoires, nous présentons les différentes façons de prendre en compte l'autocorrélation spatiale dans le contexte de ces modèles.

7.1.1 Modèle standard : modéliser les effets spécifiques individuels

Relativement aux données en coupe transversale, les données de panel, *i.e.* plusieurs observations pour les mêmes individus, permettent de tenir compte de l'influence de certaines caractéristiques non observées invariantes dans le temps de ces individus.
Pour un échantillon comportant des informations sur un ensemble d'individus indicés par $i = 1, ..., N$ que l'on suppose observables pendant toute la période d'étude $t = 1, ..., T$ (*i.e.* il n'y a ni attrition, ni observations manquantes), le modèle standard (*a-spatial*) s'écrit :

$$y_{it} = x_{it}\beta + z_i\alpha + \varepsilon_{it} \tag{7.1}$$

Les k variables explicatives du modèle sont regroupées dans k vecteurs x_{it} de dimension $(1,k)$ (qui n'inclut pas de vecteur unitaire) et sont supposées exogènes. Le vecteur β de dimension $(k,1)$

désigne le vecteur des paramètres inconnus à estimer. L'hétérogénéité, ou effet spécifique individuel, est captée par le terme $z_i\alpha$. Le vecteur z_i comprend un terme constant et un ensemble de variables spécifiques aux individus, invariantes dans le temps, qui peuvent être observées (sexe, éducation, etc.) ou non (préférences, compétences, etc.). Les hypothèses formulées sur les termes d'erreur ε_{it} dépendent du type de modèle considéré. En effet, selon la nature des variables prises en compte dans le vecteur z_i, on peut considérer trois classes de modèle : le modèle sur données empilées, le modèle à effets fixes et le modèle à effets aléatoires.

Le premier type de modèle, sur données empilées, correspond au cas pour lequel z_i ne comprend qu'une constante :

$$y_{it} = x_{it}\beta + \alpha + \varepsilon_{it} \tag{7.2}$$

où $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$. L'hétérogénéité individuelle n'est pas modélisée ; la spécification conduit à un simple empilement des données en coupes transversales. Dans ce cas un estimateur convergent et efficace de β et de α est obtenu par la méthode des Moindres Carrés Ordinaires (MCO).

Dans le second modèle, dit à effets fixes, l'hétérogénéité individuelle est modélisée par la prise en compte d'effets spécifiques individuels constants dans le temps. Ce modèle s'écrit :

$$y_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it} \tag{7.3}$$

où l'effet fixe α_i est un paramètre (moyenne conditionnelle) à estimer constant dans le temps et $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$. Dans ce modèle, les différences de comportement inobservables sont ainsi captées par ces paramètres estimables. Ce modèle est alors particulièrement adapté dès lors que l'échantillon est exhaustif au regard de la population qu'il concerne et que le modélisateur souhaite restreindre les résultats obtenus à l'échantillon qui a permis de les obtenir. Les effets individuels α_i peuvent être corrélés avec les variables explicatives x_{it} et l'estimateur *within* (*i.e.* l'estimateur des MCO obtenu à partir d'un modèle où les variables explicatives et expliquée sont centrées sur leur moyenne individuelle respective, voir équation 7.20) reste convergent.

Dans le troisième modèle, à effets aléatoires, l'hétérogénéité individuelle est modélisée par la prise en compte d'effets spécifiques individuels aléatoires (constants au cours du temps). On fait l'hypothèse que cette hétérogénéité individuelle inobservable n'est pas corrélée avec x_{it} :

$$\begin{aligned} y_{it} &= x_{it}\beta + \alpha + u_{it} \\ u_{it} &= \alpha_i + \varepsilon_{it} \end{aligned} \tag{7.4}$$

où $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$.
Contrairement au modèle à effets fixes, les effets individuels ne sont plus des paramètres à estimer, mais les réalisations d'une variable aléatoire. Ce modèle est donc adapté si les spécificités individuelles sont reliées à des causes aléatoires. Il est également préférable au modèle à effets fixes lorsque les individus présents dans l'échantillon sont tirés d'une population plus large et que l'objectif de l'étude empirique est de généraliser les résultats obtenus à la population. Ce modèle présente l'avantage de fournir des estimations plus précises que celles obtenues à partir du modèle à effets fixes. Il s'estime usuellement à l'aide de la méthode des Moindres Carrés Généralisés (MCG).

Dans la suite de ce chapitre, nous adoptons une présentation générale de la spécification de la nature des effets individuels en distinguant les effets individuels fixes des effets aléatoires. Nous présentons également les tests usuels de spécification permettant de choisir la bonne méthode d'estimation et donc la spécification la plus adaptée pour modéliser l'hétérogénéité. Cependant, si ces modèles permettent de prendre en compte l'hétérogénéité individuelle, ils ont en commun

avec le modèle standard en coupe transversale de reposer sur l'hypothèse que les individus sont indépendants les uns des autres. Si les données portent sur des individus pour lesquels on dispose d'informations géolocalisées et que l'on suppose l'existence d'interactions spatiales, cette hypothèse n'est plus acceptable. Il convient donc d'étendre les spécifications présentées précédemment en prenant en compte l'autocorrélation spatiale.

7.1.2 Les effets spatiaux dans les modèles en données de panel

Comme pour les modèles en coupe transversale, la prise en compte de l'autocorrélation spatiale peut se faire de plusieurs manières : par des variables spatiales décalées, endogènes ou exogènes, ou par une autocorrélation spatiale des erreurs.

Les effets spatiaux dans les modèles sur données empilées

Dans un premier temps, nous reprenons le modèle empilé en incorporant ces trois termes spatiaux potentiels :

$$\begin{aligned} y_{it} &= \rho \sum_{i \neq j} w_{ij} y_{jt} + x_{it}\beta + \sum_{i \neq j} w_{ij} x_{jt}\theta + \alpha + u_{it} \\ u_{it} &= \lambda \sum_{i \neq j} w_{ij} u_{jt} + \varepsilon_{it} \end{aligned} \tag{7.5}$$

w_{ij} est un élément d'une matrice de pondération spatiale W_N de dimension (N,N) dans laquelle sont définies les relations de voisinage entre les individus de l'échantillon. Par convention, les éléments diagonaux w_{ii} sont tous fixés à zéro. La matrice de poids est généralement normalisée en ligne. La plupart des travaux académiques considèrent une matrice de pondération spatiale fixe dans le temps. La variable $\sum_{i \neq j} w_{ij} y_{jt}$ désigne la variable endogène spatialement décalée ; elle est égale à la valeur moyenne de la variable dépendante prise par les voisins (au sens de la matrice de poids) de l'observation i. Le paramètre ρ capte l'effet d'interaction endogène. L'interaction spatiale est également prise en compte par la spécification d'un processus autorégressif spatial dans les erreurs $\sum_{i \neq j} u_{jt}$ selon lequel les chocs inobservables affectant l'individu i interagissent avec les chocs affectant son voisinage. Le paramètre λ capte un effet corrélé des inobservables. Enfin, un effet contextuel (ou d'interaction exogène) est capté par le vecteur θ de dimension $(k,1)$. Comme précédemment, on suppose que $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$.

En empilant les données pour chaque période t, le modèle précédent s'écrit de la façon suivante :

$$\begin{aligned} y_t &= \rho W_N y_t + x_t\beta + W_N x_t\theta + \alpha + u_t \\ u_t &= \lambda W_N u_t + \varepsilon_t \end{aligned} \tag{7.6}$$

où y_t est le vecteur de dimension $(N,1)$ des observations de la variable expliquée pour la période t, x_t est la matrice (N,k) des observations sur les variables explicatives pour la période t. Enfin, en empilant les données pour tous les individus, le modèle s'écrit sous forme matricielle de la façon suivante :

$$\begin{aligned} y &= \rho (I_T \otimes W_N) y + x\beta + (I_T \otimes W_N) x\theta + \alpha + u \\ u &= \lambda (I_T \otimes W_N) u + \varepsilon \end{aligned} \tag{7.7}$$

où $\otimes$ désigne le produit Kronecker et $(I_T \otimes W_N)$ est une matrice de dimension (NT,NT) de la forme suivante :

$$\begin{pmatrix} W_N & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & W_N \end{pmatrix}$$

Comme vu dans le chapitre précédent : "économétrie spatiale : modèles courants", les paramètres de ce modèle ne sont généralement pas identifiables (MANSKI 1993). Il convient de faire des choix sur la nature des termes spatiaux à privilégier dans le modèle. Ces choix peuvent s'appuyer sur une modélisation théorique et/ou reposer sur une stratégie de spécification allant du spécifique au général à partir des résultats des tests du multiplicateur de Lagrange utilisés pour les modèles en coupe transversale.

Cependant, l'intérêt du modèle sur données empilées reste limité, puisque celui-ci ne permet pas de considérer la présence d'hétérogénéité individuelle alors que les individus sont susceptibles de différer du fait de caractéristiques inobservables ou difficilement mesurables. Selon la manière dont est modélisée l'hétérogénéité inobservable (fixe par opposition à aléatoire) l'omission de ces caractéristiques peut compromettre la convergence des estimateurs pour les paramètres β, θ et α. En conséquence, les modèles à effets spécifiques, fixes ou aléatoires, sont à privilégier. Nous présentons, dans ce cadre, les spécifications faisant intervenir simultanément un ou deux des termes spatiaux présentés plus haut, pour lesquels nous disposons d'estimateurs documentés dans la littérature.

Les effets spatiaux dans les modèles à effets fixes

Plusieurs spécifications spatiales peuvent être considérées pour tenir compte de l'autocorrélation spatiale dans le modèle à effets fixes. La première spécification est le modèle autorégressif spatial (SAR) , qui s'écrit :

$$y_{it} = \rho \sum_{i \neq j} w_{ij} y_{jt} + x_{it}\beta + \alpha_i + u_{it} \tag{7.8}$$

où $u_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. L'interaction spatiale est ici modélisée à travers l'introduction de la variable dépendante spatialement décalée ($\sum_{i \neq j} w_{ij} y_{jt}$). Comme dans les modèles en coupe transversale, l'introduction de cette variable implique des effets de débordement globaux : en moyenne, la valeur de y au temps t pour une observation i n'est pas seulement expliquée par les valeurs des variables explicatives pour cette observation, mais aussi par celles associées à toutes les observations (voisines de i ou non). C'est l'effet de multiplicateur spatial. Un effet global de diffusion spatiale est également à l'œuvre : un choc aléatoire dans une observation i au temps t affecte non seulement la valeur de y de cette observation à la même période mais a également un effet sur les valeurs de y des autres observations.

Le deuxième modèle est connu sous le nom de modèle à erreur spatiale (SEM) :

$$\begin{aligned} y_{it} &= x_{it}\beta + \alpha_i + u_{it} \\ u_{it} &= \lambda \sum_{i \neq j} w_{ij} u_{jt} + \varepsilon_{it} \end{aligned} \tag{7.9}$$

où $u_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. L'interaction spatiale est captée à travers une spécification autorégressive spatiale du terme d'erreur ($\lambda \sum_{i \neq j} w_{ij} u_{jt}$). Seul l'effet de diffusion spatiale est présent dans un modèle SEM, il reste cependant global.

Un troisième modèle, préconisé par LESAGE et al. 2009, est le modèle spatial de Durbin (SDM) qui contient une variable dépendante spatialement décalée ($\sum_{i \neq j} w_{ij} y_{jt}$) et des variables explicatives spatialement décalées ($\sum_{i \neq j} w_{ij} x_{jt}$) :

$$y_{it} = \rho \sum_{i \neq j} w_{ij} y_{jt} + x_{it}\beta + \sum_{i \neq j} w_{ij} x_{jt}\theta + \alpha_i + u_{it} \tag{7.10}$$

où $u_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$.

Une alternative à ce modèle est le modèle de Durbin spatial dans les erreurs (SDEM) , qui est composé d'un terme d'erreur spatialement autocorrélé ($\sum_{i\neq j} w_{ij}u_{jt}$) et des variables explicatives spatialement décalées ($\sum_{i\neq j} w_{ij}x_{jt}$) :

$$\begin{aligned} y_{it} &= x_{it}\beta + \sum_{i\neq j} w_{ij}x_{jt}\theta + \alpha_i + u_{it} \\ u_{it} &= \lambda \sum_{i\neq j} w_{ij}u_{jt} + \varepsilon_{it} \end{aligned} \tag{7.11}$$

où $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$. À travers l'autocorrélation spatiale des erreurs, il existe bien un effet de diffusion globale mais il n'y a pas d'effet multiplicateur. En effet, introduire des variables spatiales explicatives décalées induit des effets de débordements locaux et non globaux (voir chapitre 6 : "économétrie spatiale : modèle courants").

Enfin, certains auteurs utilisent une modélisation faisant intervenir simultanément un processus autorégressif spatial de la variable dépendante et du terme d'erreur (SARAR), les pondérations spatiales (w_{ij} et m_{ij}) étant distinctes pour chacun des processus (LEE et al. 2010b ; ERTUR et al. 2015) :

$$\begin{aligned} y_{it} &= \rho \sum_{i\neq j} w_{ij}y_{jt} + x_{it}\beta + \alpha_i + u_{it} \\ u_{it} &= \lambda \sum_{i\neq j} m_{ij}u_{jt} + \varepsilon_{it} \end{aligned} \tag{7.12}$$

où $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$.

Les effets spatiaux dans les modèles à effets aléatoires

Dans les modèles à effets aléatoires, les effets individuels non observés sont supposés non corrélés avec les autres variables explicatives du modèle et peuvent donc être traités comme des composants du terme d'erreur. Dans ce contexte, le modèle SAR s'écrit de manière similaire à ce qui a été proposé dans le cadre du modèle à effets fixes, à l'exception du terme d'effet individuel :

$$\begin{aligned} y_{it} &= \rho \sum_{i\neq j} w_{ij}y_{jt} + x_{it}\beta + \alpha + u_{it} \\ u_{it} &= \alpha_i + \varepsilon_{it} \end{aligned} \tag{7.13}$$

où $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$.

L'effet aléatoire étant une partie du terme d'erreur, deux spécifications SEM sont proposées dans la littérature. Dans la première (SEM-RE), l'effet de diffusion spatiale n'est considéré que pour le terme d'erreur idiosyncratique [1] et non pour l'effet individuel aléatoire (BALTAGI et al. 2003). On peut donc écrire :

$$\begin{aligned} y_{it} &= x_{it}\beta + u_{it} \\ u_{it} &= \alpha_i + \lambda \sum_{i\neq j} w_{ij}u_{jt} + v_{it} \end{aligned} \tag{7.14}$$

où $v_{it} \overset{i.i.d.}{\sim} \mathcal{N}iid(0,\sigma^2)$.

1. ie. le terme d'erreur individuel temporel.

Dans une seconde spécification (RE-SEM), suggérée par KAPOOR et al. 2007 (on désigne souvent cette spécification par KKP), on considère que la structure de corrélation spatiale s'applique à la fois aux effets individuels et à la composante restante du terme d'erreur :

$$\begin{aligned} y_{it} &= x_{it}\beta + \alpha + u_{it} \\ u_{it} &= \lambda \sum_{i \neq j} w_{ij} u_{jt} + v_{it} \\ v_{it} &= \alpha_i + \varepsilon_{it} \end{aligned} \tag{7.15}$$

où $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.

Ces deux spécifications impliquent des effets de reports spatiaux assez différents régis par des matrices de variances-covariances de structure différente, ce qui a des implications en matière d'estimation. D'autre part, comme le soulignent BALTAGI et al. 2013, ces deux modèles ont des implications différentes : dans le premier, seule la composante qui varie dans le temps se diffuse spatialement, tandis que dans le second cela caractérise également la composante permanente.

Enfin, on peut finalement envisager une spécification plus générale comme celle suggérée par BALTAGI et al. 2007 [2] :

$$\begin{aligned} y_{it} &= x_{it}\beta + u_{it} \\ u_{it} &= \alpha_i + \lambda \sum_{i \neq j} w_{ij} u_{jt} + v_{it} \\ \alpha_i &= \eta \sum_{i \neq j} w_{ij} \alpha_j + e_i \end{aligned} \tag{7.16}$$

où $e_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.

Le processus autorégressif spatial sur l'effet individuel s'interprète comme un effet de diffusion spatiale permanent sur la période.

7.1.3 Interprétation des coefficients en présence d'un terme autorégressif spatial

Comme dans les modèles de régression en coupe transversale, on peut, à partir des spécifications précédentes, donner l'expression des effets marginaux des variables explicatives ainsi que celles des impacts directs, indirects et totaux qui facilitent l'interprétation des coefficients des modèle estimés. En effet, à la différence des modèles a-spatiaux, l'effet marginal d'une variation d'une variable explicative peut être différent d'un individu à l'autre. En effet, du fait des interactions spatiales, la variation d'une variable explicative pour un individu affecte directement son résultat et indirectement les résultats de tous les autres individus. La fonction `impacts.splm`, du package *splm* de R, étend les méthodes de calcul d'impact développées pour les modèles en coupe en tenant compte de la spécificité de la dimension (NT, NT) de la matrice de pondération spatiale qui intervient dans les spécifications sur données de panel [3].

Quelle que soit la nature des données prises en compte, du fait des interactions spatiales, toute variation d'une variable explicative x_k pour un individu i entraîne une variation de la variable dépendante pour ce même individu (effet direct) mais également pour les autres (effet indirect). Pour une même variation unitaire ces effets peuvent être différents d'un individu à l'autre. Les mesures d'impacts proposées par LESAGE et al. 2009 sont donc des effets moyens dont l'expression va dépendre de la spécification spatiale retenue.

2. Ce modèle admet la spécification de KAPOOR et al. 2007 comme cas particulier pour $\eta = \lambda$ et BALTAGI et al. 2003 pour $\eta = 0$.

3. Le lecteur peut se référer à PIRAS 2014 pour plus de détails dans le calcul des effets directs, indirects et totaux sous R.

Dans le modèle de régression en coupe, si l'on part de la forme réduite du modèle autorégressif spatial (SAR), les mesures d'impacts de la variable explicative k se déduisent de l'équation suivante :

$$S_k(W_N) = (I_N - \lambda W_N)^{-1} I_N \beta_k. \tag{7.17}$$

Par analogie, dans un panel spatial statique, pour calculer les effets directs et indirects il suffit de remplacer W_N, invariante dans le temps, par la matrice diagonale par blocs $W_N = I_N \otimes W_N$. C'est cette matrice qui figure sur la diagonale de W_N dans l'équation précédente (PIRAS 2014), soit :

$$S_k(I_N \otimes W_N) = (I_{NT} - \lambda (I_N \otimes W))^{-1} I_{NT} \beta_k. \tag{7.18}$$

Plus généralement, si l'on considère un modèle spatial Durbin (SDM ; équation 7.10), la matrice des dérivées partielles de la variable dépendante, pour chacune des unités, relativement à la variable explicative k à un instant t donné s'écrit :

$$\Gamma = \left(\frac{\partial y}{\partial x_{1k}} \dots \frac{\partial y}{\partial x_{Nk}} \right)_t = (I - \rho W_N)^{-1} \begin{pmatrix} \beta_k & w_{12}\theta_k & \dots & w_{1N}\theta_k \\ w_{21}\theta_k & \beta_k & \dots & w_{2N}\theta_k \\ \vdots & \vdots & \vdots & \vdots \\ w_{N1}\theta_k & w_{N2}\theta_k & \dots & \beta_k \end{pmatrix}. \tag{7.19}$$

LESAGE et al. 2009 définissent l'effet direct comme la moyenne des éléments diagonaux de la matrice figurant dans le terme de droite de l'équation 7.19 et l'effet indirect comme la moyenne de la somme des éléments en lignes (ou en colonnes) en dehors de ceux situés sur la diagonale principale.

Dans le cas du modèle SEM, la matrice du terme de droite de l'équation 7.19 est une matrice diagonale dont les éléments sont égaux à β_k. De ce fait, l'effet direct d'une variation de la variable explicative k est égal à β_k et l'effet indirect est nul comme dans les modèles a-spatiaux et dans les modèles spatiaux en coupe.

Dans le cas du modèle SAR, bien que les éléments en dehors de la diagonale principale de la seconde matrice du terme de droite de l'équation 7.19 soient nuls, du fait de la dimension de W, le calcul des effets directs et indirects exige de mettre en œuvre des calculs matriciels et de calculer la trace de la matrice Γ qui fait intervenir des puissances de W. D'autre part, les statistiques permettant de tester la significativité de ces mesures d'impact sont obtenues par simulation de Monte Carlo (pour plus de détail voir PIRAS 2014).

7.2 Méthodes d'estimations

Deux grandes catégories de méthodes d'estimation des modèles spatiaux sur données de panel sont principalement utilisées : les méthodes fondées sur le principe du maximum de vraisemblance et les méthodes fondées sur la méthode des moments généralisée (incluant les variables instrumentales). Comme précédemment, nous nous restreignons ici au cas standard d'un panel cylindré et d'une matrice de pondération spatiale fixe dans le temps. Généralement, les estimateurs par le maximum de vraisemblance (MV) sont plus efficaces, mais reposent sur des conditions plus fortes sur la distribution du terme d'erreur. La méthode des moments généralisée (MMG) est souvent privilégiée car moins coûteuse en temps de calcul et plus facile à mettre en œuvre. D'autre part, dans la majorité des cas, comme ces estimateurs ne reposent pas sur l'hypothèse de normalité, les estimateurs que cette méthode permet d'obtenir sont plus robustes à l'hétéroscédasticité. Enfin, la flexibilité que permet la définition des conditions sur les moments permet également d'estimer les modèles spatiaux en présence d'une variable explicatives endogène. Ces deux méthodes sont implémentables sous R. Cette section présente les estimateurs des modèles à effets fixes (section 7.3.1), puis des modèles à effets aléatoires (section 7.3.2).

7.2.1 Modèle à effets fixes

Encadré 7.2.1 — Estimation d'un modèle à effets fixes par maximum de vraisemblance. Lorsque l'effet individuel spécifique est considéré comme fixe, la procédure la plus souvent employée (approche directe) consiste à transformer les variables du modèle de telle sorte à éliminer l'effet fixe, puis à estimer directement le modèle sur ces variables transformées. La transformation la plus courante est la déviation intra-individuelle (*within*). Elle consiste à différencier chaque variable par rapport à sa moyenne intra-individuelle :

$$y_{it}^* = y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it} \quad et \quad x_{it}^* = x_{it} - \frac{1}{T}\sum_{t=1}^{T} x_{it} \tag{7.20}$$

Dans un deuxième temps, l'estimation se fait à partir des variables transformées. Dans un modèle sans autocorrélation spatiale, la fonction de vraisemblance s'écrit :

$$LogL = -\frac{NT}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}^* - x_{it}^*\beta)^2 \tag{7.21}$$

Si le modèle intègre une variable endogène décalée ($\sum_{i\neq j} w_{ij}y_{jt}$), alors la fonction de vraisemblance doit être dérivée en prenant en compte l'endogénéité de $\sum_{i\neq j} w_{ij}y_{jt}$ *via* un terme jacobien (ANSELIN et al. 2006) :

$$LogL = -\frac{NT}{2}log(2\pi\sigma^2) + Tlog|I_n - \rho W| - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}^* - \rho\sum_{j\neq i} w_{ij}y_{jt}^* - x_{it}^*\beta)^2 \tag{7.22}$$

Cette fonction est très proche de celle dérivée pour le modèle SAR en coupe transversale. Son estimation suit donc une procédure semblable. Les estimateurs de β et σ^2 étant fonction de ρ, ELHORST 2003 propose d'utiliser une fonction de log vraisemblance concentrée que l'on peut maximiser à partir des résidus (u_0^* et u_1^*) de deux régressions respectivement de y_{it}^* et de $\sum_{i\neq j} w_{ij}y_{jt}^*$ sur x_{it}^* :

$$LogL_C = C + Tlog|I_n - \rho W| - \frac{NT}{2}((u_0^* - \rho u_1^*)'(u_0^* - \rho u_1^*)) \tag{7.23}$$

Il faut utiliser une procédure par itération nécessitant de fixer initialement ρ pour calculer $\hat{\beta}$ et $\hat{\sigma}^2$. Dans un deuxième temps, il faut estimer $\hat{\rho}$ de telle sorte à maximiser la fonction de log vraisemblance concentrée et recommencer à calculer $\hat{\beta}$ et $\hat{\sigma}^2$ en fixant $\hat{\rho}$ jusqu'à obtenir des résultats qui convergent numériquement.

La modélisation de l'autocorrélation spatiale à travers un terme d'erreur spatialement autocorrélé modifie seulement l'estimation de σ^2 (l'estimation de β n'est pas affectée). La méthode des moindres carrés généralisée pourrait permettre d'obtenir un estimateur de σ^2 si λ était connu. En toute généralité, ce n'est pas le cas et il est nécessaire encore une fois d'estimer de manière itérative β, λ puis σ^2. La fonction de vraisemblance concentrée peut être maximisée à l'aide des résidus (ε_{it}^*) de la régression de y_{it}^* sur x_{it}^* :

$$LogL_C = Tlog(I_N - \lambda W) - \frac{NT}{2}log(\varepsilon_{it}^*(I_N - \lambda W)'\varepsilon_{it}^*(I_N - \lambda W)) \tag{7.24}$$

LEE et al. 2010b ont remis en cause cette approche en montrant qu'elle ne permettait pas nécessairement d'obtenir des estimateurs convergents des coefficients et des écart-types.

L'ampleur des biais et les paramètres affectés diffèrent en fonction des cas. Par exemple, lorsque le modèle contient un effet fixe individuel, σ^2 est biaisé pour N grand et T fixe. Si le modèle intègre à la fois des effets temporels et individuels, les β et σ^2 seront biaisés pour N et T grand. À partir de ces résultats, LEE et al. 2010b suggèrent des corrections, spécifiques à chaque cas, permettant d'obtenir des estimateurs convergents à partir de l'approche directe. Ces corrections sont disponibles dans les principaux logiciels d'économétrie. Nous renvoyons à LEE et al. 2010b et ELHORST 2014b pour plus de précisions sur cette approche.

Encadré 7.2.2 — Estimation d'un modèle à effets fixes par la méthode des moments généralisée. Une stratégie d'estimation alternative repose sur la méthode des moments généralisée. Dans le cadre des modèles spatiaux, la stratégie proposée par KELEJIAN et al. 1999 pour les données en coupe est étendue aux données de panel par KAPOOR et al. 2007 et MUTL et al. 2011.

Pour un modèle SAR, la stratégie d'estimation mise en œuvre repose sur la méthode des variables instrumentales proposée par KELEJIAN et al. 1998 sur le modèle en déviation intra-individuelle (*within*). Les instruments utilisés sont les variables exogènes du modèle ainsi que leur décalage spatial.

Dans le cas d'un modèle SEM, la stratégie d'estimation du paramètre d'autocorrélation spatiale sur les erreurs repose sur les trois conditions sur les moments proposées par KELEJIAN et al. 1999 pour les données en coupe, celles-ci étant étendues aux résidus du modèle en déviation intra-individuelle. Les autres paramètres du modèle peuvent alors être estimés par les moindres carrés ordinaires à partir d'un modèle auquel il a été appliqué une transformation de type Cochrane-Orcutt.

7.2.2 Modèle à effets aléatoires

Encadré 7.2.3 — Estimation d'un modèle à effets aléatoires par maximum de vraisemblance. Lorsqu'on considère un modèle à effets aléatoires, on fait l'hypothèse que les effets individuels non observés ne sont pas corrélés avec les variables explicatives du modèle. Comme dans le cas du modèle à effets fixes, on peut mettre en œuvre une méthode en deux étapes en utilisant des variables pour lesquelles la transformation dépend de ϕ tel que $\phi^2 = \sigma^2/(T\sigma_\alpha^2 + \sigma^2)$, soit :

$$y_{it}^{o} = y_{it} - (1-\phi)\frac{1}{T}\sum_{t=1}^{T} y_{it} \quad et \quad x_{it}^{o} = x_{it} - (1-\phi)\frac{1}{T}\sum_{t=1}^{T} x_{it} \tag{7.25}$$

On remarque que si $\phi = 0$, on se ramène à la transformation *within* et le modèle à effets aléatoires se ramène à un modèle à effets fixes.
Dans un modèle sans autocorrélation spatiale, la fonction de vraisemblance s'écrit :

$$LogL = -\frac{NT}{2}log(2\pi\sigma^2) + \frac{N}{2}log(\phi^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}^{o} - x_{it}^{o}\beta)^2 \tag{7.26}$$

Si le modèle intègre une variable endogène décalée, alors la fonction de vraisemblance

s'écrit :

$$LogL = -\frac{NT}{2}log(2\pi\sigma^2) + Tlog|I_n - \rho W| + \frac{N}{2}log(\phi^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}^{o} - \rho\sum_{j\neq i} w_{ij}y_{jt}^{o} - x_{it}^{o}\beta)^2 \tag{7.27}$$

Pour ϕ donné, cette fonction est très proche de celle dérivée pour le modèle SAR à effets fixes. Son estimation suit donc une procédure analogue, en utilisant une log vraisemblance concentrée que l'on peut maximiser à partir des résidus $e^{o}(\phi)$ de la régression de y_{it}^{o} sur $\sum_{i\neq j} w_{ij}y_{jt}^{o}$ et x_{it}^{o} :

$$LogL_C = -\frac{NT}{2}log\left[(e^{o}(\phi))'(e^{o}(\phi))\right] + \frac{N}{2}log(\phi^2) \tag{7.28}$$

De la même manière que précédemment, il faut fixer des valeurs initiales des paramètres inconnus puis utiliser une procédure itérative jusqu'à obtenir des résultats qui convergent numériquement.

Dans le cas d'un modèle avec erreur spatialement autocorrélée (SEM), l'écriture la plus générale de la vraisemblance est assez complexe (ELHORST 2014b) et la méthode de résolution mise en œuvre dépend de la forme de la matrice de variances-covariances des erreurs qui découle de l'hypothèse formulée sur la structure de corrélation spatiale des erreurs.

Dans le cadre de la spécification SEM-RE (seul le terme d'erreur idiosyncratique est spatialement corrélé) la vraisemblance s'écrit :

$$LogL = -\frac{NT}{2}log(2\pi\sigma^2) - \frac{1}{2}log|V| + (T-1)\sum_{i=1}^{N}log|B| - \frac{1}{2\sigma^2}e'(\bar{J}_T \otimes V^{-1})e - \frac{1}{2\sigma^2}e'(E_T \otimes (B'B))e \tag{7.29}$$

où $V = T\phi' I_N + (B'B)^{-1}$, $\quad e = y - x\beta$, $\quad B = (I_N - \lambda W)$, $\quad \phi' = \frac{\sigma^2}{\sigma_\alpha}$
avec $J_T = i_T i_T^{'}$ une matrice (T,T) de 1, $\quad \bar{J}_T = \frac{J_T}{T}$, $\quad E_T = I_T - \bar{J}_T$

Compte tenu de cette structure complexe, l'algorithme de filtrage spatial suggéré par ELHORST 2003 est particulièrement adapté à la spécification dans laquelle le terme autorégressif spatial affecte la totalité du terme d'erreur. Dans le cadre de la spécification considérée par KAPOOR et al. 2007 (KKP), la matrice de variance covariance a une forme spécifique plus simple que dans le cas précédent ce qui facilite considérablement la mise en œuvre de l'estimation par le MV en deux étapes (MILLO et al. 2012).

Cette même procédure peut être mise en œuvre pour de nombreuses autres spécifications mixant des hypothèses sur la structure d'autocorrélation spatiale. Ces méthodes d'estimation sont implémentées *via* la fonction `spreml` qui permet d'estimer par le MV plus de spécifications que la fonction `spml` (MILLO 2014).

Encadré 7.2.4 — Estimation d'un modèle à effets aléatoires par la méthode des moments généralisée. Comme dans le modèle à effets fixes, la mise en œuvre de l'estimation par la méthode des moments généralisée repose sur la stratégie proposée par KELEJIAN et al. 1999 pour les données en coupe et étendue aux données de panel par KAPOOR et al. 2007 et MUTL

et al. 2011. Par exemple, dans le modèle SEM-RE, afin d'estimer le paramètre autorégressif λ et les variances des termes d'erreurs $\sigma_1^2 = \sigma_v^2 + T\sigma_\alpha^2$ et σ_v^2, ils définissent un ensemble de 6 conditions sur les moments. MILLO et al. 2012 détaillent les différentes variantes de cet estimateur selon les conditions formulées sur les moments. Ensuite, pour les paramètres du modèle, un estimateur des moindres carrés généralisés réalisables est défini basé sur une transformation de type Cochrane-Orcutt du modèle initial.

7.3 Tests de spécification

Nous présentons dans un premier temps le test de spécification d'Hausman qui permet d'arbitrer entre un modèle où les effets individuels ne sont pas corrélés avec les variables explicatives et un modèle où une telle corrélation existe. Ce test permet de définir quelle méthode d'estimation retenir. Dans un deuxième temps, nous présentons les autres tests de spécification permettant de choisir la spécification la plus adéquate.

7.3.1 Choisir entre effet fixe et effet aléatoire

Pour que le modèle à effets aléatoires soit valide, une hypothèse cruciale est que les caractéristiques inobservables ne soient pas corrélées avec les variables explicatives observables. L'hypothèse nulle du test peut se mettre sous la forme générale $\mathbb{E}[\alpha|X] = 0$. Si cette hypothèse n'est pas rejetée, les deux estimateurs MCG et *within* seront convergents. Dans le cas contraire, l'estimateur MCG ne sera pas convergent alors que l'estimateur *within* restera convergent.

Le test de spécification d'Hausman (HAUSMAN 1978) peut s'appliquer pour tester le modèle à effets aléatoires contre le modèle à effets fixes. Dans notre cas, ce test se construit en mesurant l'écart (pondéré par une matrice de variance covariance) entre les estimations produites par les estimateurs *within* (modèle à effets fixes) et MCG (modèle à effets aléatoires) dont on sait que l'un des deux (*within*) est convergent quelle que soit l'hypothèse faite sur la corrélation entre variables et caractéristiques inobservables tandis que l'autre (MCG) n'est pas convergent dans le seul cas où cette hypothèse n'est pas vérifiée. Par conséquent, une différence significative des deux estimations implique une mauvaise spécification du modèle à effet aléatoire.

MUTL et al. 2011 ont montré que ces propriétés restent valides dans un cadre spatial si l'on remplace chaque estimateur *within* et MCG par son "analogue" spatial (prenant en compte les termes d'autocorrélation spatiale). Le test d'Hausman robuste à l'autocorrélation spatiale s'écrit :

$$S_{hausman} = NT(\hat{\beta}_{MCG} - \hat{\beta}_{within})'(\hat{\sum}_{within} - \hat{\sum}_{MCG})^{-1}(\hat{\beta}_{MCG} - \hat{\beta}_{within}) \tag{7.30}$$

où $\hat{\beta}_{MCG}$ et $\hat{\beta}_{within}$ sont les estimations des paramètres obtenus respectivement par MCG et *within*, $\hat{\sum}_{within}$ et $\hat{\sum}_{MCG}$ correspondent aux éléments des matrices de variances-covariances des deux estimations.

7.3.2 Tests de spécification des effets spatiaux

Il s'agit ici de présenter certains des tests permettant de retenir la spécification la plus adéquate de la prise en compte de la dépendance spatiale. Nous insistons sur les tests mis en œuvre dans la package R *splm*. Les tests de spécification de l'autocorrélation spatiale les plus couramment utilisés reposent sur le test du multiplicateur de Lagrange. Ils permettent de tester l'absence de chacun des termes spatiaux sans avoir à estimer le modèle non contraint. Un ensemble de tests a été développé par DEBARSY et al. 2010 dans le cadre d'un modèle à effets fixes.

On complète généralement ces deux tests par leur version robuste à la forme alternative de prise en compte de l'autocorrélation spatiale. Dans ce cas, il s'agit pour le RLMlag de tester l'absence de terme autorégressif spatial lorsque le modèle contient déjà un terme autorégressif spatial dans

les erreurs (RLMlag), ou inversement pour RLMerr de tester l'absence de terme autorégressif spatial dans les erreurs lorsque le modèle contient un terme autorégressif spatial. L'interprétation des résultats de ces tests est similaire à celle présentée dans le chapitre 6 "économétrie spatiale : modèles courants" sur les données en coupe.

BALTAGI et al. 2003 et BALTAGI et al. 2007 dérivent un ensemble de tests pour toutes les combinaisons d'effets aléatoires et d'autocorrélation spatiale dans les erreurs. Ces tests ont été complétés par BALTAGI et al. 2008 qui proposent un test joint d'absence de terme autorégressif spatial en présence d'effets individuels aléatoires. Les hypothèses de ces tests, également fondés sur le principe du multiplicateur de Lagrange, sont décrites dans la table 7.1.

Test	hypothèse nulle	hypothèse alternative
LMjoint	$\lambda = \sigma_\alpha^2 = 0$	$\lambda \neq 0$ ou $\sigma_\alpha^2 \neq 0$
SLM1	$\sigma_\alpha^2 = 0$ en posant $\lambda = 0$	$\sigma_\alpha^2 \neq 0$ en posant $\lambda = 0$
SLM2	$\lambda = 0$ en posant $\sigma_\alpha^2 = 0$	$\lambda \neq 0$ en posant $\sigma_\alpha^2 = 0$
CLMerr	$\lambda = 0$ en posant $\sigma_\alpha^2 >= 0$	$\lambda \neq 0$ en posant $\sigma_\alpha^2 >= 0$
CLMrandom	$\sigma_\alpha^2 = 0$ en posant $\lambda >= 0$	$\sigma_\alpha^2 \neq 0$ en posant $\lambda >= 0$

TABLE 7.1 – Test d'autocorrélation spatiale en présence d'effet aléatoire et/ou de corrélation sérielle

Enfin, comme dans les modèles en coupe transversale, il est possible de mettre en œuvre des tests de significativité sur les coefficients dans la mesure où certains des modèles présentés précédemment présentent un caractère emboîté. Ainsi, il est possible de retrouver le modèle SAR et le modèle SEM à partir du modèle SDM avec les contraintes testables suivantes sur les paramètres, respectivement $H_0 : \theta = 0$ (test de significativité du vecteur de paramètres θ) et $H_0 : \rho\beta - \theta = 0$ (test du facteur commun). De même à partir du modèle SDEM, on retrouve le modèle SEM si l'hypothèse $H_0 : \theta = 0$ ne peut pas être rejetée.

7.4 Application empirique

7.4.1 Le modèle

Notre application empirique porte sur la deuxième "loi" de VERDOORN 1949. Cette loi relie, de manière linéaire, les taux de croissance de la productivité du travail p à ceux de l'output q dans le secteur manufacturier pour un ensemble d'économies. La spécification de base est donnée par :

$$p_{it} = b_0 + b_1 q_{it} + \varepsilon_{it} \tag{7.31}$$

où b_0 et b_1 sont les paramètres inconnus à estimer et ε_{it} est un terme d'erreur pour lequel nous supposons dans un premier temps que $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. Le paramètre b_1 est appelé le coefficient de Verdoorn pour lequel une valeur positive traduit la présence de rendements croissants (FINGLETON et al. 1998). Cette spécification a été affinée par FINGLETON 2000, 2001 afin de caractériser l'endogénéité du progrès technique. Il suppose notamment un changement technique proportionnel à l'accumulation du capital par tête et une croissance du capital par tête égale à la croissance de la productivité et des effets de débordements géographiques, liés notamment à la diffusion des technologies et du capital humain entre unités spatiales. La spécification étendue de Verdoorn qui

découle de ces analyses est [4] :

$$p_{it} = b_0 + b_1 q_{it} + b_2 G_{it} + b_3 u_{it} + b_4 d_{it} + \varepsilon_{it} \tag{7.32}$$

où G correspond à l'écart technologique (approché par le différentiel de productivité du travail) au début de la période entre chaque unité et l'unité spatiale "leader". Dans le cadre des modèles de croissance endogène, les unités spatiales avec un retard technologique sont susceptibles de connaître une croissance de la productivité plus faible que celle des unités spatiales plus développées. u est une mesure de l'urbanisation, mesurée par la densité de population et a pour objectif de capter l'effet de la densité de l'activité économique. Enfin, d mesure le niveau initial de productivité du travail dans le secteur manufacturier (ANGERIZ et al. 2008).

Nous définissons cette spécification sous R :

```
# Spécifier le modèle à estimer
verdoorn<-p~q+u+G+d
```

La prise en compte des effets de débordements spatiaux nécessite d'estimer la spécification augmentée d'un terme autorégressif spatial (FINGLETON 2000, 2001) :

$$p_{it} = b_0 + \rho \sum_{i \neq j} w_{ij} p_{jt} + b_1 q_{it} + b_2 G_{it} + b_3 u_{it} + b_4 d_{it} + \varepsilon_{it} \tag{7.33}$$

Cette spécification est justifiée théoriquement par FINGLETON 2000 et 2001 et correspond à la spécification estimable d'un modèle inspirée de la Nouvelle Économie Géographique. À des fins d'illustration, nous considérons également une spécification alternative correspondant à un modèle autorégressif spatial dans les erreurs :

$$\begin{aligned} p_{it} &= b_0 + b_1 q_{it} + b_2 G_{it} + b_3 u_{it} + b_4 d_{it} + \varepsilon_{it} \\ \varepsilon_{it} &= \alpha_i + \lambda \sum_{i \neq j} w_{ij} \varepsilon_{jt} + v_{it} \end{aligned} \tag{7.34}$$

ou :

$$\varepsilon_{it} = \lambda \sum_{i \neq j} w_{ij} \varepsilon_{jt} + v_{it} \tag{7.35}$$

L'estimation de modèles sur données de panel avec R nécessite les packages *plm* (panel sans autocorrélation spatiale, gestion d'objets `pdata.frame` adaptée au panel) et *splm* (estimation et tests pour panels spatiaux). Il convient également de charger les packages *sp*, *maps* et *maptools* pour l'importation et la gestion des objets spatiaux.

```
# packages nécessaires
library(plm)
library(splm)
library(sp)
library(maps)
library(maptools)
```

4. L'analyse originale de FINGLETON 2000, 2001 est basée sur un modèle en coupe transversale, nous l'étendons au cas des données de panel.

L'estimation des spécifications les plus courantes se fait à l'aide des commandes `spml` et `spreml` pour le maximum de vraisemblance et `spgm` pour la méthode des moments généralisée. Celles ci ont toutes une structure relativement identique avec des options additionnelles selon les cas :

```
# Maximum de Vraisemblance :
spml(formula, data, index=NULL, listw, listw2=listw, na.action,
                 model=c("within","random","pooling"),
                 effect=c("individual","time","twoways"),
                 lag=FALSE, spatial.error=c("b","kkp","none"),
                 ...)
```

Il faut dans un premier temps définir la spécification (`formula=...`) sans indiquer les effets spatiaux (qui seront définis par des options spécifiques), indiquer le nom du `pdata.frame` (`data=...`) et le `listw` nécessaire à la création des variables spatialement décalées (`listw=...`). La nature des effets spécifiques est déterminée par l'option `model` : l'utilisateur a le choix entre `pooling` pour un modèle sur données empilées, `within` pour un modèle à effets fixes ou `random` pour un modèle à effets aléatoires. On peut également définir si les effets concernent les individus ou/et les périodes grâce à l'option `effects` qui peut être posée égale à `individual`, `time` ou `twoways`. On peut également choisir si la spécification comporte des termes spatiaux : `lag=T` pour le modèle SAR ou `lag=F` dans le cas contraire. Pour finir, on peut choisir la nature de la spécification dans le modèle à effets aléatoires : `spatial.error="b"` pour une spécification à la Baltagi, `spatial.error="kkp"` pour la spécification à la KKP (KAPOOR et al. 2007) ou `spatial.error="none"` sinon.

La commande `spreml` permet d'estimer, par le maximum de vraissemblance, plus de spécifications avec effets aléatoires (`errors=`) avec la possibilité d'envisager différentes configurations parmi lesquelles celle d'introduire de la corrélation sérielle dans le terme d'erreur. Compte tenu des calculs matriciels que cela induit, elle comporte de nombreuses options pour paramétrer l'algorithme de calcul :

```
spreml(formula, data, index = NULL, w, w2=w, lag = FALSE,
          errors = c("semsrre", "semsr", "srre", "semre",
          "re", "sr", "sem","ols", "sem2srre", "sem2re"),
          pvar = FALSE, hess = FALSE, quiet = TRUE,
          initval = c("zeros", "estimate"),
          x.tol = 1.5e-18, rel.tol = 1e-15, ...)
```

Enfin, la commande `spgm` permet d'estimer les paramètres par la méthode des moments généralisée.

```
spgm(formula, data=list(), index=NULL, listw =NULL, listw2 = NULL,
         model=c("within","random"), lag = FALSE, spatial.error=TRUE,
         moments = c("initial", "weights", "fullweights"), endog = NULL,
         instruments= NULL, lag.instruments = FALSE, verbose = FALSE,
         method = c("w2sls", "b2sls", "g2sls", "ec2sls"), control = list(),
         optim.method = "nlminb", pars = NULL)
```

Les tests de spécification reprennent en grande partie ces options. Le test d'Hausman robuste à l'hétéroscédasticité est obtenu à l'aide de la commande `sphtest`. La commande `slmtest` permet de mettre en œuvre les tests de spécification de l'autocorrélation spatiale. Les tests de spécification sur le terme d'erreur (effet aléatoire, autocorélation spatiale, autocorrélation sérielle) s'effectuent à l'aide de la commande `bsjktest`. Ces tests sont facilement interprétables car l'hypothèse alternative est toujours rappelée dans l'output.

7.4.2 Les données et la matrice de poids

Notre analyse porte sur un échantillon de 1032 régions (NUTS3) européennes localisées dans 14 États membres de l'UE15 (seule la Grèce n'est pas présente dans notre échantillon). Les données sont disponibles pour la période 1991-2008. Nous agrégeons les données annuelles par période de 3 ans afin de contrôler des variations économiques de court terme (cycles). Nous obtenons un panel de 6 périodes pour lequel nous construisons les taux de croissance de la productivité du travail (p) et de la valeur ajoutée (q) dans le secteur manufacturier. Les estimations porteront donc sur 5 périodes. La figure 7.1 représente le périmètre d'étude de notre analyse.

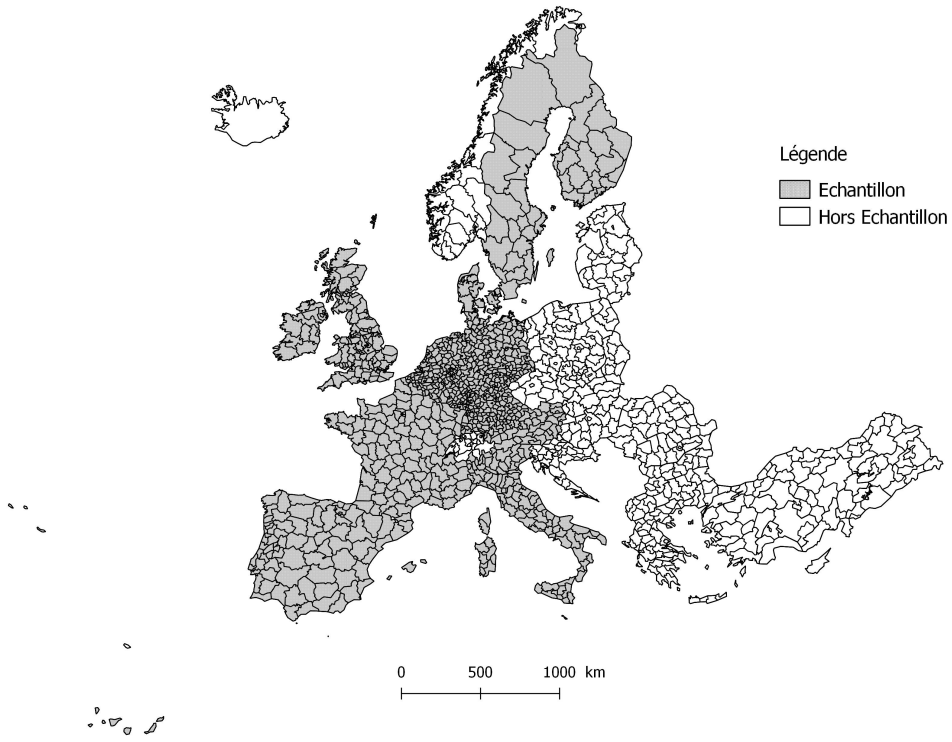

FIGURE 7.1 – Périmètre d'étude

```
# Importation des données
data_panel <- read.csv("panel_average_3_years_1991_2008.csv", sep=";")
# Importation du shapefile (Gisco) en objet "spatialpolygondataframe"
shape_nuts3<-readShapeSpatial("NUTS_RG_60M_2006")
# Sélection des NUTS3 (par niveau de NUTS)
shape_nuts3<- shape _nuts3[shape_nuts3$STAT_LEVL_== 3,]
# Sélection des NUTS3 de notre échantillon
data_panel_code<- data_panel[,"NUTS3"]
shape_nuts3<- shape _nuts3[shape_nuts3$NUTS_ID %in% data_panel_code,]
# Visualisation de l'échantillon
plot(shape_nuts3)
```

Afin de générer un tableau de statistiques descriptives en format LaTeXde la variable expliquée et des variables explicatives du modèle, il est possible d'utiliser le package `stargazer` et d'appliquer la commande *stargazer* sur la base de données comprenant les variables du modèle. Le résultat est reporté dans la table 7.2.

```
library(stargazer)
variables <-data.frame(data_panel$p,data_panel$q,data_panel$u,data_panel$G,
    data_panel$d)
stargazer(variables, title="Statistiques descriptives")
```

Statistic	N	Mean	St. Dev.	Min	Max
p	5 160	0.402	0.078	0.000	0.888
q	5 160	0.399	0.081	0.000	0.900
u	5 160	51.761	110.371	0.187	2 084.284
G	5 160	45.667	12.054	0.000	90.055
d	5 160	3.801	0.335	1.746	5.405

TABLE 7.2 – Statistiques descriptives

Concernant la matrice de poids, la présence d'îles dans l'échantillon (Madère, Canaries entre autres) nécessite l'utilisation d'une matrice de poids basée sur un autre critère que la simple contiguïté liée à la présence d'une frontière commune (voir le chapitre 2 :" Codifier la structure de voisinage"). Nous construisons une matrice des 10 plus proches voisins afin de garantir une connexion entre les régions de la Grande Bretagne et de l'Europe continentale.

```
# Création d'une matrice k plus proches voisins, k = 10
map_crd <- coordinates(shape_nuts3)
Points_nuts3 <- SpatialPoints(map_crd)
nuts3.knn_10 <- knearneigh(Points_nuts3, k=10)
K10_nb <- knn2nb(nuts3.knn_10)
wknn_10 <- nb2listw(K10_nb, style="W")
```

7.4.3 Les résultats

Pour retenir la spécification la plus appropriée, nous partons du modèle sans autocorrélation spatiale et mettons en œuvre le test d'Hausman et des tests du multiplicateur de Lagrange.

La table 7.3 présente les résultats de l'estimation d'un modèle sans autocorrélation spatiale des erreurs. La colonne (1) correspond au modèle sur données empilées alors que les colonnes (2) et (3) prennent en compte l'hétérogénéité individuelle inobservée respectivement à travers des effets fixes et des effets aléatoires. Concernant le coefficient de Verdoorn, les résultats sont similaires : avec un coefficient significatif et positif supérieur à 0.5 dans les trois cas, la présence de rendements d'échelle croissants est confirmée pour notre échantillon. Le taux de croissance de l'emploi dans le secteur manufacturier d'une région est également d'autant plus grand que cette région est urbanisée (coefficient associé à u positif et significatif dans le premier et troisième cas), d'autant plus grand que l'écart avec la région leader en début de période est important (coefficient associé à G positif et significatif dans le premier et le troisième cas) et d'autant moins important que la productivité initiale est grande, ce qui traduit un phénomène de convergence des productivités du travail dans le secteur manufacturier (coefficient associé à d négatif et significatif dans les trois cas).

```
##### Table 7.3 : estimation sans prise en compte de l'autocorrélation
    spatiale
summary(verdoorn_pooled <- plm(verdoorn, data = data_panel, model = "
    pooling"))
```

```
summary(verdoorn_fe1<- plm(verdoorn, data = data_panel,
                           model = "within", effect="individual"))
summary(verdoorn_re1<- plm(verdoorn, data = data_panel,
                           model = "random", effect="individual"))
```

		p	
Modèle :	données empilées	effets fixes (*within*)	effets aléatoires (MCG)
	(1)	(2)	(3)
q	0.692^{***}	0.604^{***}	0.701^{***}
	(0.009)	(0.010)	(0.010)
u	0.0001^{***}	-0.0002	0.0001^{***}
	(0.00001)	(0.0002)	(0.00001)
G	0.0001	0.002^{***}	0.0003^{***}
	(0.0001)	(0.0001)	(0.0001)
d	-0.008^{***}	-0.182^{***}	-0.033^{***}
	(0.003)	(0.005)	(0.003)
Constante	0.146^{***}		0.228^{***}
	(0.012)		(0.014)
Observations	5 160	5 160	5 160
$R^2 ajust$	0.523	0.587	0.552

TABLE 7.3 – Estimations sans prise en compte de l'autocorrélation spatiale
Note : * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Les résultats du test d'Hausman standard et du test d'Hausman robuste à l'autocorrélation spatiale des erreurs conduisent au rejet de l'hypothèse nulle de l'absence de corrélation entre les effets individuels et les variables explicatives. Nous optons donc dans la suite de l'analyse empirique pour un modèle à effets fixes.

```
# Test d'Hausman (plm)
print(hausman_panel<-phtest(verdoorn, data = data_panel))
## Hausman Test
## data:  verdoorn
## chisq = 1040.8, df = 4, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent

# Test d'Hausman robuste à l'autocorrélation spatiale (splm)
print(spat_hausman_ML_SEM<-sphtest(verdoorn,data=data_panel,
                        listw =wknn_10, spatial.model = "error", method="ML
                           "))
## Hausman test for spatial models
## data:  x
```

```
## chisq = 1263.8, df = 4, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent

print(spat_hausman_ML_SAR<-sphtest(verdoorn,data=data_panel,
            listw =wknn_10,spatial.model = "lag", method="ML"))
## Hausman test for spatial models
## data:  x
## chisq = 1504, df = 4, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

Les résultats des tests du multiplicateur de Lagrange dans un modèle à effets fixes nous conduisent à privilégier une spécification SEM (code des tests ci-dessous). Si les statistiques de test de prise en compte de l'autocorrélation spatiale par un SAR (Test 1) ou par un SEM (Test 2) confirment le rejet de l'hypothèse que ces deux termes (pris indépendamment) sont nuls, la lecture simultanée ne nous permet pas de conclure sur la spécification la plus appropriée pour prendre en compte l'autocorrélation spatiale (ces deux tests n'étant pas emboîtés). On peut toutefois noter que la statistique de test pour une alternative SEM est supérieure à celle correspondant à une alternative SAR. Pour conclure de façon plus crédible, on utilise des tests robustes à la présence de la spécification alternative de l'autocorrélation spatiale (Tests 3 et 4). Autrement dit, il s'agit pour le RLMlag de tester l'absence de terme autorégressif spatial lorsque le modèle contient déjà un terme autorégressif spatial dans les erreurs (RLMlag), ou inversement pour RLMerr de tester l'absence de terme autorégressif spatial dans les erreurs lorsque le modèle contient un terme autorégressif spatial. La version robuste RLMerr est fortement significative (Test 4) alors que RLMlag ne l'est pas (Test 3). Nous estimons donc un modèle à effets fixes avec un processus autorégressif spatial dans les erreurs. Dans certains cas, ces deux derniers tests robustes ne permettent pas de discriminer entre un SAR et un SEM. Plusieurs possibilités sont envisageables. La première consiste à estimer un modèle comportant ces deux termes spatiaux (SARAR). La seconde consiste à discriminer entre les deux spécifications sur les bases des statistiques des tests RLMerr et RLMlag (en prenant la spécification dont la statistique associée est la plus élevée) ou de comparer les critères d'Akaike des deux spécifications.

```
# Modèle effets fixes
# Test 1
slmtest(verdoorn, data=data_panel, listw = wknn_10, test="lml",
        model="within")
## LM test for spatial lag dependence
## data:  formula (within transformation)
## LM = 326.41, df = 1, p-value < 2.2e-16
## alternative hypothesis: spatial lag dependence

# Test 2
slmtest(verdoorn, data=data_panel, listw = wknn_10, test="lme",
        model="within")
## LM test for spatial error dependence
## data:  formula (within transformation)
## LM = 1115.5, df = 1, p-value < 2.2e-16
## alternative hypothesis: spatial error dependence

# Test 3
slmtest(verdoorn, data=data_panel, listw = wknn_10, test="rlml",
```

```
        model="within")
## Locally robust LM test for spatial lag dependence sub spatial error
## data:  formula (within transformation)
## LM = 0.0025551, df = 1, p-value = 0.9597
## alternative hypothesis: spatial lag dependence

# Test 4
slmtest(verdoorn, data=data_panel, listw = wknn_10, test="rlme",
        model="within")
## Locally robust LM test for spatial error dependence sub spatial lag
## data:  formula (within transformation)
## LM = 789.08, df = 1, p-value < 2.2e-16
## alternative hypothesis: spatial error dependence
```

		p		
Modèle :	données empilées	effets fixes (MV)		effets fixes (MMG)
		erreur Baltagi	erreur KKP	
	(1)	(2)	(3)	(4)
q	0.716***	0.650***	0.650***	0.836***
	(0.017)	(0.008)	(0.008)	(0.009)
u	0.0001***	0.0001	0.0001	0.0001
	(0.00001)	(0.0002)	(0.0002)	(0.0002)
G	-0.0004***	0.001***	0.001***	0.0003***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
d	-1.70***	-0.163***	-0.163***	-0.164***
	(0.003)	(0.0005)	(0.0005)	(0.005)
Constante	0.2***			
	(0.02)			
λ		0.566***	0.566***	0.513***
		(0.02)	(0.02)	(0.02)
Observations	5 160	5 160	5 160	5 160

TABLE 7.4 – Estimations du modèle sur données empilées et du modèle à effets fixes avec autocorrélation spatiale des erreurs
Note : * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

La table 7.4 synthétise les résultats de l'estimation du modèle avec prise en compte de l'autocorrélation spatiale sous la forme d'une autocorrélation spatiale des erreurs. Contrairement au modèle SAR, les paramètres estimés d'un SEM s'interprètent de manière classique [5]. La première colonne

5. Il n'est pas nécessaire de calculer les effets directs, indirects et totaux dans le cadre d'un SEM en raison de

correspond au modèle sur données empilées alors que les trois colonnes suivantes présentent les résultat correspondant au modèle à effets fixes avec différentes méthodes d'estimation (maximum de vraisemblance pour les colonnes (2) et (3) ; MMG pour la colonne (4)) et différentes spécifications du terme d'erreur (Baltagi pour la colonne (2) et KKP pour la colonne (3)). Dans tous les cas, le coefficient d'autocorrélation est positif et significatif. Concernant le coefficient de Verdoorn, il reste positif et significatif et d'une ampleur plus importante que précédemment. L'impact de l'urbanisation n'est plus significatif lorsque l'on introduit un effet fixe : les variations temporelles de densité de population n'affectent pas significativement le taux de croissance de de la productivité du travail. L'effet de l'urbanisation observé sur données empilées provient certainement de caractéristiques inobservables favorables à l'urbanisation (par exemple les avantages de localisation de première nature, KRUGMAN 1999).

```
##### Table 7.4 : Estimations du modèle sur données
empilées et du  modèle à effets fixes
avec autocorrélation spatiale des erreurs
# Estimation par le Maximum de Vraisemblance
summary(verdoorn_SEM_pool <- spml(verdoorn, data = data_panel,
listw = wknn_10, lag=FALSE,model="pooling"))
# SEM à effets fixes
summary(verdoorn_SEM_FE<- spml(verdoorn, data = data_panel,
listw = wknn_10, lag=FALSE,model="within", effect="individual", spatial.
    error="b"))
summary(verdoorn_SEM_FE<- spml(verdoorn, data = data_panel,
listw = wknn_10, lag=FALSE,model="within", effect="individual", spatial.
    error="kkp"))
# Estimation par la méthode des moments généralisée
summary(verdoorn_SEM_FE_GM <- spgm(verdoorn, data=data_panel,
       listw = wknn_10, model="within", moments="fullweights",
       spatial.error = TRUE))
```

7.5 Extensions

Dans cette section nous présentons certaines extensions des modèles spatiaux sur données de panel. Les méthodes présentées dans ces extensions ne sont pas implémentées dans R à l'heure actuelle.

7.5.1 Modèles dynamiques spatiaux

Les modèles considérés dans les sections précédentes sont des modèles statiques. Cependant, les interactions spatiales peuvent présenter un caractère dynamique. Ainsi, les valeurs prises pour une observation i à une période de temps t peuvent dépendre des valeurs prises par les observations voisines de i à la période précédente. Le même type de schéma peut s'appliquer pour les termes d'erreurs. Le caractère dynamique peut être pris en compte en repartant de l'équation 7.6, où nous introduisons des retards temporels sur la variable expliquée et son décalage spatial :

$$y_t = \tau y_{t-1} + \rho W_N y_t + \eta W_N y_{t-1} + x_t \beta + W_N x_t \theta + \alpha + u_t \tag{7.36}$$

l'absence d'effet multiplicateur spatial. Toutefois, nous renvoyons le lecteur à (PIRAS 2014) pour le calcul de ces effets dans un SAR en panel statique.

Ce modèle peut s'interpréter comme un modèle de Durbin spatial dynamique (DEBARSY et al. 2012 ; LEE et al. 2015). Dans ce modèle la valeur de la variable expliquée prise pour une observation i au temps t dépend de la valeur de la variable expliquée pour l'observation i à la période précédente (retard temporel), de la valeur de la variable expliquée pour les observations voisines à i à la période t (décalage spatial simultané) et enfin de la valeur de la variable expliquée pour les observations voisines de i à la période précédente $t-1$ (décalage spatial retardé). Pour ce dernier terme, on peut par exemple penser à des effets de diffusion spatiale : un choc se produisant en une zone i à une période t qui se diffuse aux zones voisines dans les périodes suivantes. On pourrait également incorporer des retards temporels sur les variables explicatives X_t ou le terme d'erreur u_t mais comme le montrent ANSELIN et al. 2008 et ELHORST 2012, les paramètres d'un tel modèle ne sont pas identifiables. Enfin, en toute généralité, ce modèle peut inclure un effet individuel, fixe ou aléatoire. DEBARSY et al. 2012 détaillent la nature des impacts (directs, indirects, totaux) dans ce modèle. Pour donner l'intuition de ces impacts, on réécrit le modèle décrit par l'équation 7.36 sous la forme suivante :

$$y_t = (I_N - \rho W_N)^{-1}(\tau y_{t-1} \eta W_N y_{t-1}) + (I_N - \rho W_N)^{-1}(x_t\beta + W_N x_t\theta) + (I_N - \rho W_N)^{-1}(\alpha + u_t) \quad (7.37)$$

La matrice des dérivées partielles de la valeur espérée de y_t par rapport à la k^{me} variable explicative de X à la période t est alors :

$$\left[\frac{\partial q\mathbb{E}(y)}{\partial x_{1k}} \quad \ldots \quad \frac{\partial q\mathbb{E}(y)}{\partial x_{nk}}\right]_t = (I_N - \rho W_N)^{-1}(\beta_k I_N + \theta_k W_N) \quad (7.38)$$

Ces dérivées partielles reflètent l'effet d'un changement affectant une variable explicative pour une observation i sur la variable expliquée de toutes les autres observations dans le court terme uniquement. Les effets de long terme sont définis par :

$$\left[\frac{\partial q\mathbb{E}(y)}{\partial x_{1k}} \quad \ldots \quad \frac{\partial q\mathbb{E}(y)}{\partial x_{nk}}\right]_t = [(1-\tau)I_N - (\rho+\eta)W_N]^{-1}(\beta_k I_N + \theta_k W_N) \quad (7.39)$$

Les effets directs sont constitués des éléments de la diagonale du terme à droite de l'équation 7.38 ou de l'équation 7.39 et les effets indirects comme la somme des lignes ou des colonnes des éléments non diagonaux de ces matrices. Ces effets sont indépendants de la période t. Il n'y a donc pas d'effet indirect de court terme si $\rho = \theta_k = 0$ et il n'y a pas d'effet indirect de long terme si $\rho = -\eta$ et si $\theta_k = 0$.

Deux grandes catégories de méthodes ont été proposées pour estimer ce modèle. D'une part, en se basant sur le principe du maximum de vraisemblance, YU et al. 2008 construisent un estimateur pour le modèle décrit par l'équation 7.36 incluant des effets fixes individuels. Cet estimateur est étendu par LEE et al. 2010a pour un modèle incluant en outre des effets fixes temporels. L'intuition est d'estimer le modèle par la méthode du maximum de vraisemblance conditionnellement à la première observation. Ils proposent également une correction lorsque le nombre d'unités spatiales et le nombre de périodes tend vers l'infini. D'autre part, LEE et al. 2010a proposent un estimateur des Moments Généralisés optimal basé sur des conditions linéaires et des conditions quadratiques. Cet estimateur est convergent, même si le nombre de périodes est petit par rapport au nombre d'observations spatiales.

Le lecteur pourra se reporter à ELHORST 2012 ou LEE et al. 2015 pour une présentation plus détaillée des modèles de panels spatiaux dynamiques.

7.5.2 Modèles multidimensionnels spatiaux

Dans certains cas, les données de panel présentent une structure multidimensionnelle plus complexe. Par exemple, dans les modèles gravitaires, des flux économiques (flux de commerce, d'IDE, etc.) entre des objets spatiaux (des pays ou des régions) sont modélisés dans des modèles de panel à trois dimensions en introduisant des effets fixes individuels, temporels, voire des effets bilatéraux d'interaction. L'introduction de l'autocorrélation spatiale dans ces modèles de type gravitaire est abordée par exemple par Arbia (2015). La structure multidimensionnelle peut également être de nature hiérarchique. Ainsi, les données régionales européennes sont disponibles à plusieurs échelles spatiales : NUTS3, NUTS2, NUTS1, les régions NUTS3 étant imbriquées dans les régions NUTS2, ces dernières étant elles-mêmes imbriquées dans les régions NUTS1. Dans le cas des modèles de panel a-spatiaux, une série d'articles des années 2000 (par exemple BALTAGI et al. 2001) modélisent cette structure hiérarchique à travers une spécification particulière des effets aléatoires. Récemment, des auteurs ont étendu cette littérature sur les modèles hiérarchiques à l'analyse des panels spatiaux (voir LE GALLO et al. 2017 pour une revue de littérature récente). Nous présentons ici la logique générale de ces modélisations.

Formellement, soit un panel multidimensionnel à 3 dimensions où la variable dépendante est observée selon trois indices : y_{ijt} avec $i = 1,2,\ldots,N$, $j = 1,2,\ldots,M_i$ et $t = 1,2,\ldots,T$. N est le nombre de groupes. M_i est le nombre d'individus dans le groupe i, de telle sorte qu'il y a $S = \sum_{i=1}^{N} M_i$ individus. T représente le nombre de périodes. En toute généralité il peut y avoir un nombre différent d'individus entre les N groupes, cependant le panel reste cylindré dans la dimension temporelle. Dans le cas d'une structure hiérarchique spatiale, on suppose que l'indice j se réfère aux individus (par exemple, les régions NUTS3) qui sont imbriqués dans N groupes (par exemple, les régions NUTS2). En supposant que l'autocorrélation spatiale se produit au niveau des individus et que les coefficients sont homogènes, on peut écrire le modèle SDM suivant :

$$y_{ijt} = \rho \sum_{g=1}^{N} \sum_{h=1}^{M_g} w_{ij,gh} y_{ght} + x_{ijt}\beta + \sum_{g=1}^{N} \sum_{h=1}^{M_g} w_{ij,gh} x_{ght}\theta + \varepsilon_{ijt}, \tag{7.40}$$

où y_{ijt} est la valeur de la variable dépendante pour l'individu j dans le groupe i à la période t. x_{ijt} est un vecteur $(1,K)$ de variables explicatives exogènes, alors que β et θ sont des vecteurs $(K,1)$ de paramètres inconnus à estimer. ε_{ijt} est le terme d'erreur avec des propriétés détaillées plus bas. La pondération spatiale $w_{ij,gh} = w_{k,l}$ est l'élément $(k = ij; l = gh)$ de la matrice de pondération spatiale W_S avec ij dénotant l'individu j dans le groupe i, et de façon similaire pour gh. Ainsi, $k,l = 1,\ldots,S$ et W_S est une matrice de pondération de dimension (S,S) avec les propriétés habituelles. ρ est le paramètre de décalage spatial. En toute généralité, on peut également spécifier une autocorrélation spatiale des erreurs, sous la forme d'un modèle autorégressif au niveau individuel :

$$\varepsilon_{ijt} = \lambda \sum_{g=1}^{N} \sum_{h=1}^{M_g} m_{ij,gh} \varepsilon_{ght} + u_{ijt}. \tag{7.41}$$

Le poids $m_{ij,gh}$ est un élément de la matrice de poids M_S. Par simplicité, on peut supposer que $M_S = W_S$. λ est le paramètre spatial à estimer. u_{ijt} est un terme aléatoire composé qui capte la structure hiérarchique des données. À cette fin, on suppose que u_{ijt} est la somme d'une composante spécifique au groupe et invariante dans le temps α_i, une composante spécifique au couple individu-groupe invariante dans le temps μ_{ij} et un terme résiduel v_{ijt} :

$$u_{ijt} = \alpha_i + \mu_{ij} + v_{ijt}, \tag{7.42}$$

avec les hypothèses suivantes : (i) $\alpha_i \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_\alpha^2\right)$, (ii) $\mu_{ij} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_\mu^2\right)$, (iii) $v_{ijt} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_v^2\right)$ et (iv) les trois termes sont indépendants les uns des autres. Le lecteur pourra consulter (LE GALLO et al. 2017) pour les méthodes d'estimation (maximum de vraisemblance, méthode des moments généralisés), d'inférence statistique et de prévision adaptées à ces modèles.

7.5.3 Modèles de panels à facteurs communs

L'apport majeur des données de panel réside dans la modélisation de l'hétérogénéité inobservée. Les modèles présentés précédemment se proposent de modéliser l'hétérogénéité inobservée en utilisant une transformation des variables (modèle à effets fixes) ou en posant des hypothèses sur la structure du terme d'erreur (modèle à effets aléatoires). Dans les deux cas, une restriction est faite sur la forme de l'hétérogénéité : pour chaque individu, elle est constante dans la dimension temporelle. Autrement dit, il y a une séparation totale des deux dimensions individuelle et temporelle : les effets spécifiques individuels variant entre individus mais restant constants dans le temps et les effets spécifiques temporels variant dans le temps mais constants dans la dimension individuelle. Si cette hypothèse reste crédible dans le cadre des panels courts, elle est trop restrictive pour les panels composés d'une dimension temporelle importante.

Dans certains cas, les bases de données comprennent également une dimension temporelle importante. Les modèles à facteurs commun ont été développés pour exploiter cette configuration des données. Cette nouvelle classe de modèles permet de modéliser l'effet de facteurs communs qui affectent différemment les individus, en résumant l'information présente dans les données en un nombre réduit de facteurs communs :

$$y_{it} = x_{it}\beta + \sum_{l=1}^{d} \lambda_{il} f_{lt} + \varepsilon_{it} \tag{7.43}$$

où $\sum_{l=1}^{d} \lambda_{il} f_{lt}$ correspond aux facteurs communs du modèle. Nous renvoyons à BAI et al. 2016 pour une présentation plus précise de cette classe de modèles, et nous nous intéressons à ce qui les relie aux panels spatiaux.

Par définition, les facteurs communs et panels spatiaux permettent de capter les interactions entre individus. Ils adoptent toutefois des logiques différentes. Les modèles d'économétrie spatiale reposent sur une structure donnée des interactions entre les individus d'un panel. Cette structure est généralement construite à partir d'une métrique géographique (distance entre les individus). Dans les panels à facteurs communs, la structure des interactions n'est pas contrainte *a priori* (seul le nombre de facteurs communs est contraint).

Initialement, les panels spatiaux étaient utilisés pour des panels comprenant un grand nombre d'individus (relativement à la dimension temporelle) et l'utilisation des modèles à facteurs communs employés lorsque la dimension temporelle suffisamment grande pour construire correctement les facteurs communs. Récemment, une série de travaux a mis en avant, à travers des applications, les synergies entre les deux approches (BHATTACHARJEE et al. 2011 ; ERTUR et al. 2015) et a proposé des méthodes combinant effets spatiaux et facteurs communs (PESARAN et al. 2009 ; 2011 ; SHI et al. 2017a ; 2017b). Une application récente est proposée par VEGA et al. 2016 qui étudie l'évolution des disparités de chômage entre régions néerlandaises à l'aide d'un modèle prenant en compte les dépendances spatiales et temporelles mais également la présence de facteurs communs. Leur étude met l'accent sur l'importance de prendre en compte simultanément ces trois dimensions (et non à l'aide de méthodes en plusieurs étapes) sous risque d'obtenir des résultats biaisés. L'analyse de leurs résultats suggère que la dépendance spatiale reste un élément important pour comprendre les dispersions de taux de chômage régionaux, même une fois prise en compte la dépendance temporelle et la présence de facteurs communs.

Conclusion

L'économétrie spatiale sur données de panel est aujourd'hui l'un des domaines les plus actifs de l'économétrie spatiale, tant sur le plan théorique qu'empirique. Dans ce contexte, ce chapitre a présenté les principaux modèles d'économétrie spatiale sur données de panel. Il n'a pas pour vocation d'être exhaustif sur l'ensemble des spécifications, méthodes d'estimation et d'inférence, mais il s'est concentré sur les procédures implémentables actuellement dans le logiciel R. Ces procédures concernent les modèles spatiaux de panel statiques, pour données cylindrées, avec des matrices de poids invariantes dans le temps. Des librairies ou scripts existent également pour des logiciels propriétaires comme Matlab (commandes proposés par ELHORST 2014a) et Stata (module `XSMLE`, BELOTTI et al. 2017b) et permettent de compléter les procédures proposées sous R.

Références - Chapitre 7

ANGERIZ, Alvaro, John MCCOMBIE et Mark ROBERTS (2008). « New estimates of returns to scale and spatial spillovers for EU Regional manufacturing, 1986—2002 ». *International Regional Science Review* 31.1, p. 62–87.

ANSELIN, Luc, Julie LE GALLO et Hubert JAYET (2006). « Spatial panel econometrics ». *The econometrics of panel data, fundamentals and recent developments in theory and practice*. Sous la dir. de Dordrecht KLUWER. 3e éd. T. 4. The address of the publisher : Matyas L, Sevestre P, p. 901–969.

— (2008). « Spatial panel econometrics ». *The econometrics of panel data*. Springer, p. 625–660.

BAI, Jushan et Peng WANG (2016). « Econometric analysis of large factor models ». *Annual Review of Economics* 8, p. 53–80.

BALTAGI, Badi H, Peter EGGER et Michael PFAFFERMAYR (2013). « A Generalized Spatial Panel Data Model with Random Effects ». *Econometric Reviews* 32.5, p. 650–685.

BALTAGI, Badi H et Long LIU (2008). « Testing for random effects and spatial lag dependence in panel data models ». *Statistics & Probability Letters* 78.18, p. 3304–3306.

BALTAGI, Badi H, Heun Song SEUCK et Won KOH (2003). « Testing panel data regression models with spatial error correlation ». *Journal of econometrics* 117.1, p. 123–150.

BALTAGI, Badi H, Seuck Heun SONG et Byoung Cheol JUNG (2001). « The unbalanced nested error component regression model ». *Journal of Econometrics* 101.2, p. 357–381.

BALTAGI, Badi H et al. (2007). « Testing for serial correlation, spatial autocorrelation and random effects using panel data ». *Journal of Econometrics* 140.1, p. 5–51.

BELOTTI, Federico, Gordon HUGHES, Andrea Piano MORTARI et al. (2017b). « XSMLE : Stata module for spatial panel data models estimation ». *Statistical Software Components*.

BHATTACHARJEE, Arnab et Sean HOLLY (2011). « Structural interactions in spatial panels ». *Empirical Economics* 40.1, p. 69–94.

DEBARSY, Nicolas et Cem ERTUR (2010). « Testing for spatial autocorrelation in a fixed effects panel data model ». *Regional Science and Urban Economics* 40.6, p. 453–470.

DEBARSY, Nicolas, Cem ERTUR et James P LESAGE (2012). « Interpreting dynamic space–time panel data models ». *Statistical Methodology* 9.1, p. 158–171.

ELHORST, J Paul (2003). « Specification and estimation of spatial panel data models ». *International regional science review* 26.3, p. 244–268.

— (2012). « Dynamic spatial panels : models, methods, and inferences ». *Journal of geographical systems* 14.1, p. 5–28.

— (2014a). « Matlab software for spatial panels ». *International Regional Science Review* 37.3, p. 389–405.

— (2014b). « Spatial panel data models ». *Spatial Econometrics*. Springer, p. 37–93.

ERTUR, Cem et Antonio MUSOLESI (2015). « Weak and Strong cross-sectional dependence : a panel data analysis of international technology diffusion ». *SEEDS Working Papers 1915*.

FINGLETON, Bernard (2000). « Spatial econometrics, economic geography, dynamics and equilibrium : a 'third way' ? » *Environment and planning A* 32.8, p. 1481–1498.

— (2001). « Equilibrium and economic growth : spatial econometric models and simulations ». *Journal of regional Science* 41.1, p. 117–147.

FINGLETON, Bernard et John SL MCCOMBIE (1998). « Increasing returns and economic growth : some evidence for manufacturing from the European Union regions ». *Oxford Economic Papers* 50.1, p. 89–105.

HAUSMAN, Jerry (1978). « Specification Tests in Econometrics ». *Econometrica* 46.6, p. 1251–1271.

HSIAO, Cheng (2014). *Analysis of panel data*. 54. Cambridge university press.

KAPOOR, Mudit, Harry H KELEJIAN et Ingmar R PRUCHA (2007). « Panel data models with spatially correlated error components ». *Journal of Econometrics* 140.1, p. 97–130.

KELEJIAN, Harry H et Ingmar PRUCHA (1998). « A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances ». *Journal of Real Estate Finance and Economics* 17, p. 99–121.

— (1999). « A generalized moments estimator for the autoregressive parameter in a spatial model ». *International Economic Review* 40.2, p. 509–533.

KRUGMAN, Paul (1999). « The role of geography in development ». *International regional science review* 22.2, p. 142–161.

LE GALLO, Julie et Alain PIROTTE (2017). « Models for Spatial Panels ».

LEE, Lung-fei et Jihai YU (2010a). « A spatial dynamic panel data model with both time and individual fixed effects ». *Econometric Theory* 26.2, p. 564–597.

— (2010b). « Some recent developments in spatial panel data models ». *Regional Science and Urban Economics* 40.5, p. 255–271.

— (2015). « Spatial panel data models ».

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

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

MILLO, Giovanni (2014). « Maximum likelihood estimation of spatially and serially correlated panels with random effects ». *Computational Statistics and Data Analysis* 71, p. 914–933.

MILLO, Giovanni et Gianfranco PIRAS (2012). « splm : Spatial panel data models in R ». *Journal of Statistical Software* 47.1, p. 1–38.

MUTL, Jan et Michael PFAFFERMAYR (2011). « The Hausman test in a Cliff and Ord panel model ». *The Econometrics Journal* 14.1, p. 48–76.

PESARAN, M Hashem et Elisa TOSETTI (2009). « Large panels with spatial correlations and common factors ». *Journal of Econometrics* 161.2, p. 182–202.

— (2011). « Large panels with common factors and spatial correlation ». *Journal of Econometrics* 161.2, p. 182–202.

PIRAS, Gianfranco (2014). « Impact estimates for static spatial panel data models in R ». *Letters in Spatial and Resource Sciences* 7.3, p. 213–223.

SHI, Wei et Lung-fei LEE (2017a). « Spatial dynamic panel data models with interactive fixed effects ». *Journal of Econometrics* 197.2, p. 323–347.

— (2017b). « A spatial panel data model with time varying endogenous weights matrices and common factors ». *Regional Science and Urban Economics*.

VEGA, Solmaria Halleck et J Paul ELHORST (2016). « A regional unemployment model simultaneously accounting for serial dynamics, spatial dependence and common factors ». *Regional Science and Urban Economics* 60, p. 85–95.

VERDOORN, JP (1949). « On the factors determining the growth of labor productivity ». *Italian economic papers* 2, p. 59–68.

YU, Jihai, Robert DE JONG et Lung-fei LEE (2008). « Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and T are large ». *Journal of Econometrics* 146.1, p. 118–134.