7. Spatial econometrics on panel data

BOUAYAD AGHA SALIMA
GAINS (TEPP) and 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

Abstract

This chapter offers a summary presentation of the spatial econometric methods applied to panel data. We focus primarily on the specifications and methods implemented in the *splm* package available in R. We illustrate our presentation with an analysis of Verdoorn's second "law" before presenting recent extensions to the spatial models on panel data.

R Prior reading of Chapter 6: "Spatial econometrics: common models" is recommended.

Introduction

Panel data is a data structure consisting of a set of individuals (firms, households, local authorities) observed on multiple time periods (Hsiao 2014). With respect to cross-section data, the access to information on the individual and temporal dimensions offers three main advantages. The additional information related to the use of the individual dimension of the data makes it possible to account for the presence of unobservable heterogeneity. The larger sample sizes improves the accuracy of the estimates. Lastly, panel data can be used to model dynamic relations.

After a first generation of spatial models specified for cross-sectional data (Elhorst 2014b), many applications in spatial econometrics are currently based on panel data. While the a-spatial specifications on panel data make it possible to control a certain form of unobserved heterogeneity, the dependency of the cross-sections is not taken into account. In a way similar to cross-section models, the introduction of spatial effects in panel data models makes it possible to better take into account the interdependence between individuals.

In this chapter, we present the main specifications of the spatial panels, starting from standard panel data specifications (section 7.1). Section 7.2 is dedicated to the presentation of estimation methods, while section 7.3 describes the main specifications tests specific to spatial panels. We propose an empirical application by testing Verdoorn's second law as part of a panel of European regions (NUTS3) between 1991 and 2008 (section 7.4). Section 7.5 presents a number of recent extensions of spatial panels.

7.1 Specifications

This section presents the main specifications used for static models on panel data, taking into account spatial interactions. We consider only the case of balanced panels — individuals are observed for all periods. Research on estimation methods for unbalanced spatial panels is still less developed. Dynamic models will be briefly discussed in section 7.5.1. After a brief review of what characterises standard panel data specifications (without spatial dependence) and what distinguishes specific fixed effects from random effects, we present the different ways of taking spatial autocorrelation into account in the context of these models.

7.1.1 Standard model: modelling individual specific effects

Regarding cross-section data, the panel data, *i.e.* multiple observations for the same individuals, make it possible to take into account the influence of some non-observed characteristics invariant over time for these individuals.

For a sample with information on a set of individuals indexed by $i = 1,...,N$ that are assumed to be observable throughout the study period $t = 1,...,T$ (*i.e.* there is no attrition or missing observations), the standard (*a-spatial*) model is written:

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

The k explanatory variables of the model are grouped in k vectors x_{it} with dimension $(1,k)$ (which does not include a unit vector) and are assumed to be exogenous. The vector β dimension $(k,1)$ refers to the vector of unknown parameters to be estimated. Heterogeneity, or individual specific effect, is captured by the term $z_i\alpha$. Vector z_i includes a constant term and a set of variables specific to individuals that are invariant over time, whether observed (gender, education, etc.) or not observed (preferences, skills, etc.). The assumptions on the error terms ε_{it} depend on the type of model considered. Depending on the nature of the variables taken into account in vector z_i,

three model classes can be considered — the pooled data model, the fixed-effect model and the random-effect model.

The first type of model, based on pooled data, reflects a case in which z_i includes only one constant:

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

where $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. Individual heterogeneity is not modelled. The specification results in simple data pooling into cross-sections. In this case, a consistent and efficient estimator of β and α is obtained using Ordinary Least Squares (OLS).

In the second so-called "fixed effects" model, individual heterogeneity is modelled by taking into account specific individual effects that are constant over time. This model is written:

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

where the fixed effect α_i is a parameter (conditional average) to be estimated constant over time and $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. In this model, the unobservable differences are thus captured by these estimated parameters. This model is then particularly suitable when the sample is exhaustive with regard to the population to which it pertains, and the modeller wishes to restrict the results obtained to the sample that made it possible to obtain them. Individual effects α_i can be correlated with explanatory variables x_{it} and the estimator *within*, *i.e.* the estimated OLS derived from a model where the explanatory and explained variables are centered on their respective individual average, or Equation 7.20, remain consistent.

In the third model — the random effect model — individual heterogeneity is modelled by taking random individual specific effects into account (constant over time). We assume that this unobservable individual heterogeneity is not correlated with x_{it}:

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

where $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.
Unlike the fixed-effect model, individual effects are no longer parameters to be estimated, but realisations of a random variable. This model is therefore appropriate if individual specificities are linked to random causes. It is also preferable to the fixed effects model when the individuals in the sample are drawn from a larger population and the objective of the empirical study is to generalise to the population the results obtained. This model offers the advantage of providing more accurate estimates than those derived from the fixed effects model. It is usually estimated using the Generalised Least Squares (GLS) method.

In the rest of this chapter, we adopt a general presentation of the specification of the nature of individual effects by distinguishing fixed individual effects from random effects. We also present the usual specification tests used to choose the appropriate estimation method and therefore the most suitable specification for modelling heterogeneity. However, while these models make it possible to take individual heterogeneity into account, they are, like the standard cross-section model, based on the assumption that individuals are independent from one another. If the data relate to individuals for whom geolocated information is available, and if it is assumed that spatial interactions do exist, then this hypothesis is no longer acceptable. The specifications presented above therefore need to be extended, taking spatial autocorrelation into account.

7.1.2 Spatial effects in panel data models

As with cross-section models, spatial autocorrelation can be taken into account in multiple ways – by lagged, endogenous or exogenous spatial variables, or by spatial error autocorrelation.

Spatial effects in pooled data models

The pooled data model is used by incorporating these three potential spatial terms:

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

w_{ij} is part of a spatial weighting matrix W_N of dimension (N,N) in which neighbourhood relationships between sample individuals are defined. By convention, the diagonal elements w_{ii} are all set to zero. The weight matrix is generally row-standardised. Most academic research examines a spatial weighting matrix constant over time. Variable $\sum_{i \neq j} w_{ij} y_{jt}$ refers to the spatially offset endogenous variable. It is equal to the average value of the dependent variable taken by neigbours of observation i (within the context of the weight matrix). Parameter ρ captures the endogenous interaction effect. Spatial interaction is also taken into account by specifying a spatial autoregressive process in errors $\sum_{i \neq j} u_{jt}$ according to which unobservable shocks affecting individual i interact with shocks affecting the said individual's neighbours. Parameter λ captures a correlated effect of the unobservables. Lastly, a contextual effect (or exogenous interaction) is captured by vector θ with dimension $(k,1)$. As previously, it is assumed that $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$.

By pooling data for each period t, the previous model is written as follows:

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

where y_t is the vector with dimension $(N,1)$, observations of the variable explained for period t, x_t is the matrix (N,k) for observations on explanatory variables over period t. Lastly, pooling the data for all individuals, the model is written in matrix form as follows:

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

where $\otimes$ refers to the Kronecker product and $(I_T \otimes W_N)$ is a dimension matrix (NT,NT) with the following form:

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

As shown in the previous chapter: "Spatial Econometrics: Common Models", the parameters of this model are generally not identifiable (Manski 1993a). Choices must be made about the nature of the spatial terms to be preferred in the model. These choices can be based on theoretical modelling and/or a specification strategy ranging from the specific to the general, based on the results of the Lagrange multiplier tests used for cross-sectional models.

However, the interest of the pooled data model remains limited, as it does not allow for the presence of individual heterogeneity to be taken into account, whereas individuals are likely to differ due to characteristics that are unobservable or difficult to measure. Depending on how unobservable heterogeneity (fixed versus random) is modelled, omitting these characteristics may compromise the convergence of estimators for parameters β, θ and α. Consequently, models with specific fixed or random effects should be given priority. We now present the specifications involving one or two of the spatial terms presented above, for which we have estimators documented in the literature.

Spatial effects in fixed effects models

Various spatial specifications may be considered to take into account spatial autocorrelation in the fixed effects model. The first specification is the spatial autoregressive model (SAR) , which is written:

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

where $u_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. Spatial interaction here is modelled through the introduction of the spatially lagged dependent variable ($\sum_{i \neq j} w_{ij} y_{jt}$). As in cross-section models, introducing this variable entails global spillover effects: on average, the value of y in time t for observation i is explained not only by the values of the explanatory variables for this observation, but also by those associated with all the observations (neighbouring i or otherwise). This is the spatial multiplier effect. A global spatial spillover effect is also in play: a random shock in an observation i in time t affects not only the value of y from this observation at the same period, but also has an effect on the values of y from other observations.

The second model is known as the spatial error model (SEM) :

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

with $u_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. Spatial interaction is captured through spatial autoregressive specification of the error term ($\lambda \sum_{i \neq j} w_{ij} u_{jt}$). Only the spatial diffusion effect is found in the SEM model, but it remains global.

A third model recommended by Lesage et al. 2009 is the Durbin spatial model (DSM) which contains a spatially lagged dependent variable ($\sum_{i \neq j} w_{ij} y_{jt}$) and spatially lagged explanatory variables ($\sum_{i \neq j} w_{ij} x_{jt}$):

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

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

An alternative to this model is the Durbin spatial error model (SDEM), which consist in a spatially autocorrelated error term ($\sum_{i \neq j} w_{ij} u_{jt}$) and spatially lagged explanatory variables ($\sum_{i \neq j} w_{ij} x_{jt}$):

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

where $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. Through spatial autocorrelation of errors, there is indeed a global spatial diffusion effect but no spatial multiplier effect. Introducing lagged explanatory spatial variables induces local and non-global spatial spillover effects (see chapter 6: "Spatial Econometrics: Common Models").

Lastly, some authors use modelling that simultaneously calls upon a spatial autoregressive lag and error model (SARAR), with spatial weights (w_{ij} and m_{ij}) different for each of the processes

(Lee et al. 2010b; Ertur et al. 2015):

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

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

Spatial Error Model-Random Effect

In random effect models, unobserved individual effects are assumed to be uncorrelated with the other explanatory variables in the model and can therefore be treated as components of the error term. In this context, the SAR model is written in a way similar to that proposed in the fixed effects model, except for the individual effect:

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

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

Since the random effect is part of the error term, two SEM specifications are proposed in the literature. In the first (SEM-RE), the spatial diffusion effect is considered only for the idiosyncratic error term [1] and not for the random individual effect (Baltagi et al. 2003). We can write:

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

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

In a second specification (RE-SEM), suggested by Kapoor et al. 2007 (this specification is often referred to as KKP), it is considered that the spatial correlation structure applies both to the individual effects and to the remaining component of the error term:

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

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

These two specifications imply quite different spatial spillover effects governed by various structure of the variance-covariance matrices, which have implications in terms of estimation. Furthermore, as Baltagi et al. 2013 emphasise, these two models have different implications: in the first, only the component that varies over time diffuses spatially, while in the second it also characterises the permanent component.

Lastly, a more general specification as suggested by Baltagi et al. 2007 [2]:

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

1. *i.e.* the individual time error term.

2. can be considered. This model allows for the specification of Kapoor et al. 2007 as a special case for $\eta = \lambda$ and Baltagi et al. 2003 for $\eta = 0$.

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

The spatial autoregressive process on the individual effect is interpreted as a permanent spatial diffusion effect over the period.

7.1.3 Interpretation of coefficients in the presence of a spatial autoregressive term

As in cross-section regression models, based on the previous specifications, it is possible to derive the marginal effects of the explanatory variables, along with the direct, indirect and total impacts that facilitate the interpretation of coefficients in the estimated models. This is because, unlike a-spatial models, the marginal effect of a variation in an explanatory variable may be different between individuals. This is because, due to spatial interactions, the variation of an explanatory variable for a given individual directly affects its outcome and indirectly affects the outcome of all other zones. The `impacts.splm` function, of package *splm* in R, extends the impact calculation methods developed for the cross-section models taking into account the specificity of the dimension (NT, NT) of the spatial weighting matrix called upon in panel data specifications [3].

Regardless of the nature of the data taken into account, due to spatial interactions, any variation of an explanatory variable x_k for an individual i results in a change in the dependent variable for the same individual (direct effect) but also for the others (indirect effect). For the same unit variation, these effects may differ from one individual to another. The impact measures proposed by Lesage et al. 2009 are therefore average effects, the expression of which will depend on the spatial specification chosen.

In the cross-section regression model, based on the reduced form of the spatial autoregressive model (SAR), the impact measurements of explanatory variable k are derived from the following equation:

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

By analogy, in a static spatial panel, to calculate direct and indirect effects, simply replace W_N, invariant over time, by diagonal block matrix $W_N = I_N \otimes W_N$. This matrix appears on the W_N diagonal in the previous equation (Piras 2014), or:

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

More generally, looking at a Durbin spatial model (DSM; Equation 7.10), the matrix of partial derivatives of the dependent variable, for each unit, relative to explanatory variable k at any given time t is written:

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

Lesage et al. 2009 define the direct effect as the average of the diagonal elements in the matrix in the right-hand term of Equation 7.19 and the indirect effect as the average of the sum of the items in rows (or columns) other than those located on the main diagonal.

In the case of the SEM model, the matrix of the right-hand term of Equation 7.19 is a diagonal matrix with elements equal to β_k. Accordingly, the direct effect of a variation in explanatory variable k is equal to β_k and the indirect effect is null, as in a-spatial models and cross-sectional spatial models.

3. Readers may refer to Piras 2014 for further details on calculating direct, indirect and total effects under R.

In the case of the SAR model, although the elements outside the main diagonal of the second matrix in the right-hand term of Equation 7.19 are null, due to the size of W, the calculation of direct and indirect effects requires that matrix calculations be implemented and that the trace of matrix Γ involving powers of W be calculated. Moreover, the statistics used to test the significance of these impact measurements are found by Monte Carlo simulation (for more details see Piras 2014).

7.2 Estimation methods

Two broad categories of methods for estimating spatial models using panel data are primarily used: methods based on the principle of maximum likelihood and methods based on the generalised method of moments (including instrumental variables). As before, we limit our presentation to the standard case of a cylinder panel and a spatial weighting matrix fixed over time. Generally, maximum likelihood estimators (MLE) are more effective, but require stronger conditions on the distribution of the error term. The generalised method of moments (GMM) is often preferred as it is less costly in calculation time and easier to implement. Furthermore, in the majority of cases, since these estimators are not based on the hypothesis of normality, the estimators found using this method are more robust to heteroskedasticity. Lastly, the flexibility allowed by the definition of conditions on moments also allows spatial models to be estimated in the presence of an endogenous explanatory variable. Both methods can be implemented under R.

This section presents the estimators of fixed-effect models (section 7.3.1), then random effect models (section 7.3.2).

7.2.1 Fixed effects model

Box 7.2.1 — Estimating a fixed effects maximum likelihood model. When the specific individual effect is considered fixed, the most commonly used procedure (direct approach) consists in transforming the model variables so as to remove the fixed effect and then directly estimate the model on these transformed variables. The most common transformation is intra-individual deviation (*within*). It consists in differentiating each variable from its intra-individual average:

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

Secondly, the estimate is based on the transformed variables. In a model without spatial autocorrelation, the likelihood function is written:

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

If the model includes a lagged endogenous variable ($\sum_{i\neq j} w_{ij}y_{jt}$), then the likelihood function must be derived by taking into account the endogenous nature of $\sum_{i\neq j} w_{ij}y_{jt} via$ a Jacobian term (Anselin et al. 2006) :

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

This function is very similar to that derived for the SAR cross-section model. Its estimate follows a similar procedure. As the estimators of β and σ^2 are a function of ρ, Elhorst 2003

proposes to use a concentrated log-likelihood function that can be maximised from residuals (u_0^* and u_1^*) of two regressions of y_{it}^* and $\sum_{i\neq j} w_{ij}y_{jt}^*$ of x_{it}^* :

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

An iteration procedure must be used, which requires that ρ be initially fixed to calculate $\hat{\beta}$ and $\hat{\sigma}^2$. Subsequently, $\hat{\rho}$ must be estimated, so as to maximise the concentrated log-likelihood function and re-calculate $\hat{\beta}$ and $\hat{\sigma}^2$ by fixing $\hat{\rho}$ until results converge numerically.

Modelling spatial autocorrelation through a spatially autocorrelated error term only modifies the estimate of σ^2 (the estimate of β is not affected). The generalised least squares method makes it possible to identify an estimator of σ^2 if λ was known. In general, this is not the case and the estimation needs to be carried out again iteratively β, λ followed by σ^2. The concentrated likelihood function can be maximised using residues (ε_{it}^*) of the regression of y_{it}^* on 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 have challenged this approach by showing that it does not necessarily make it possible to find consistent estimators of coefficients and standard deviations. The size of the bias and the parameters affected differs depending on the case. For example, when the model contains an individual fixed effect, σ^2 is biased for large N and fixed T. If the model includes both time and individual effects, β and σ^2 will be biased for N and large T. Based on these results, Lee et al. 2010b suggest corrections specific to each case to obtain consistent estimators from the direct approach. These corrections are available in the main econometrics software tools. We refer readers to Lee et al. 2010b and Elhorst 2014b for further details on this approach.

Box 7.2.2 — Estimating a fixed effects model using the generalised method of moments. An alternative estimation strategy is based on the generalised method of moments. In spatial models, the strategy proposed by Kelejian et al. 1999 for cross-sectional data is extended to panel data by Kapoor et al. 2007 and Mutl et al. 2011.

For a SAR model, the estimation strategy implemented is based on the instrumental variables method proposed by Kelejian et al. 1998 on the intra-individual deviation model (*within*). The instruments used are the exogenous variables of the model as well as their spatial lag.

In the case of a SEM model, the strategy for estimating the spatial autocorrelation parameter on errors is based on the three conditions on moments proposed by Kelejian et al. 1999 for cross-sectional data, these being extended to residues of the intra-individual deviation model. The other model parameters can then be estimated by the ordinary least squares, based on a model to which a Cochrane-Orcutt transformation has been applied.

7.2.2 Random effects model

Box 7.2.3 — Estimating a random effects maximum likelihood model. When considering a random effects model, it is assumed that unobserved individual effects are not correlated with the explanatory variables of the model. As in the case of the fixed effects model, a two-step method can be implemented using variables for which the transformation depends on ϕ such as

$\phi^2 = \sigma^2/(T\sigma_\alpha^2 + \sigma^2)$, or:

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

It can be noted that if $\phi = 0$, then the transformation *within* applies, and the random-effects model amounts to a fixed-effect model.
In a model without spatial autocorrelation, the likelihood function is written:

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

If the model includes a lagged endogenous variable, then the likelihood function is written:

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

For a given ϕ, this function is very close to that derived for the SAR fixed effects model. Its estimate therefore follows an analogous procedure, using a concentrated log-likelihood that can be maximised from residues $e^{o}(\phi)$ of the regression of y_{it}^{o} on $\sum_{i\neq j} w_{ij}y_{jt}^{o}$ and x_{it}^{o} :

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

In the same way as previously, initial values need to be set for unknown parameters, then an iterative procedure is used until the results found converge numerically.

In the case of a spatially auto-correlated error model (SEM), the most general way of deriving the likelihood is quite complex (Elhorst 2014b) and the resolution method used depends on the form of the variance-covariance matrix of errors that results from the hypothesis put forward on the spatial correlation structure of errors.

In the context of the SEM-RE specification (only the idiosyncratic error term is spatially correlated) the likelihood is written as follows:

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

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

Given this complex structure, the spatial filtering algorithm suggested by Elhorst 2003 is particularly suited to the specification in which the spatial autoregressive term affects the entire error term. Within the scope of the specification considered by Kapoor et al. 2007 (KKP), the variance covariance matrix has a specific form that is simpler than in the previous case, making it considerably easier to implement the two-step estimation by the MV (Millo et al. 2012).

This same procedure can be implemented for many other specifications combining hypotheses on the spatial autocorrelation structure. These estimation methods are implemented *via* the `spreml` function which makes it possible to estimate – using the MV – more specifications than the `spml` function (Millo 2014).

Box 7.2.4 — Estimating a random effects model using the generalised method of moments. As in the fixed effects model, implementing the estimation process using the generalised method of moments relies on the strategy proposed by Kelejian et al. 1999 for cross-sectional data, and extended to panel data by Kapoor et al. 2007 et Mutl et al. 2011. For example, in the SEM-RE model, in order to estimate autoregressive parameter λ and variances of error terms $\sigma_1^2 = \sigma_v^2 + T\sigma_\alpha^2$ and σ_v^2, a set of 6 conditions is defined on moments. Millo et al. 2012 detail the different variants of this estimator according to the conditions formulated on the moments. Secondly, for the parameters of the model, an estimator of realisable generalised least squares is defined based on a Cochrane-Orcutt transformation of the initial model.

7.3 Specification tests

We first present the Hausman specification test which makes it possible to arbitrate between a model where the individual effects are not correlated with the explanatory variables and a model where such a correlation exists. This test determines which estimation method to use. Secondly, we present the other specification tests that can be used to choose the most appropriate specification.

7.3.1 Choosing between fixed and random effects

The random effect model is valid since the unobservable characteristics are not correlated with observable explanatory variables. The null hypothesis of the test can be stated in the general form $E[\alpha|X] = 0$. If this hypothesis is not rejected, both GLS and *within* estimators will be consistent. Otherwise, the GLS estimator will not converge while the estimator *within* will remain consistent.

The Hausman specification test (Hausman 1978) may apply to test the random effects model against the fixed effects model. In our case, this test is constructed by measuring the gap (weighted by a covariance variance matrix) between the estimates produced by the estimators *within* (fixed effects model) and GLS (random effect model) of which it is known that one of the two (*within*) is converging regardless of the hypothesis made regarding the correlation between variables and unobservable characteristics, while the other (GLS) is not converging in the sole case where this hypothesis is not verified. Therefore, a significant difference in both estimates implies a poor specification of the random effect model.

Mutl et al. 2011 have shown that these properties remain valid in a spatial setting when replacing each estimator *within* and GLS by its spatial "analogue" (taking the terms of spatial autocorrelation into account). Hausman's robust test of spatial autocorrelation is written:

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

where $\hat{\beta}_{MCG}$ and $\hat{\beta}_{within}$ are the estimates of the parameters obtained respectively by GLS and *within*, $\hat{\sum}_{within}$ and $\hat{\sum}_{MCG}$ correspond to the elements of the variance-covariance matrices of the two estimates.

7.3.2 Specification tests for spatial effects

In this section, we present some of the tests that can be used to retain the most appropriate specification for taking spatial dependency into account. We insist on the tests implemented in package *splm* in R. The most commonly used spatial autocorrelation specification tests are based

on the Lagrange multiplier test. They test the absence of each spatial term without having to estimate the unconstrained model. A set of tests was developed by Debarsy et al. 2010 as part of a fixed-effect model.

These two tests are generally complemented by their robust version in the alternative form taking into account spatial autocorrelation. In this case, the aim is for the RLMlag to test for the absence of a spatial autoregressive term when the model already contains a spatial autoregressive term in the errors (RLMlag), or vice versa for RLMerr to test for the absence of a spatial autoregressive term in the errors when the model contains a spatial autoregressive term. The interpretation of the results of these tests is similar to that presented in Chapter 6 "Spatial econometrics: common models" on cross-section data.

Baltagi et al. 2003 and Baltagi et al. 2007 derive a set of tests for all random effect and spatial autocorrelation combinations in the errors. These tests were completed by Baltagi et al. 2008 offering a joint test on the absence of a spatial autoregressive term in the presence of random individual effects. The assumptions of these tests, also based on the Lagrange multiplier principle, are described in Table 7.1.

Test	null hypothesis	alternative hypothesis
LMjoint	$\lambda = \sigma_\alpha^2 = 0$	$\lambda \neq 0$ or $\sigma_\alpha^2 \neq 0$
SLM1	$\sigma_\alpha^2 = 0$ by stating that $\lambda = 0$	$\sigma_\alpha^2 \neq 0$ by stating that $\lambda = 0$
SLM2	$\lambda = 0$ by stating that $\sigma_\alpha^2 = 0$	$\lambda \neq 0$ by stating that $\sigma_\alpha^2 = 0$
CLMerr	$\lambda = 0$ by stating that $\sigma_\alpha^2 >= 0$	$\lambda \neq 0$ by stating that $\sigma_\alpha^2 >= 0$
CLMrandom	$\sigma_\alpha^2 = 0$ by stating that $\lambda >= 0$	$\sigma_\alpha^2 \neq 0$ by stating that $\lambda >= 0$

Table 7.1 – Spatial autocorrelation test in the presence of random effects and/or serial correlation

Lastly, as in cross-section models, it is possible to implement significance tests on the coefficients insofar as some of the models presented above are interlinked. Thus, it is possible to find the SAR model and the SEM model based on the DSM model with the following testable constraints on the parameters, respectively $H_0 : \theta = 0$ (significance test on parameter vector θ) and $H_0 : \rho\beta - \theta = 0$ (common factor test). Similarly, using the SDEM model, the SEM model can be found if the hypothesis $H_0 : \theta = 0$ cannot be rejected.

7.4 Empirical application

7.4.1 The model

Our empirical application pertains to Verdoorn's second law Verdoorn 1949. This law links, in linear fashion, labour productivity growth rates p with those of output q in the manufacturing sector for a range of economies. The basic specification is given by:

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

where b_0 and b_1 are the unknown parameters to be estimated and ε_{it} is an error term for which we initially assume that $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. Parameter b_1 is called the Verdoorn coefficient for which a positive value reflects the presence of increasing yields (Fingleton et al. 1998). This specification has been refined by Fingleton 2000, 2001 in command to characterise the endogeneity of the technical progress observed. It presupposes, in particular, a technical change proportional to the accumulation of per capita capital and growth in per capita capital equal to productivity growth and geographical spillover effects, linked in particular to the dissemination of technologies and human

capital between spatial units. The extensive specification of Verdoorn resulting from these analyses is [4]:

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

where G corresponds to the technological gap (approached by the labour productivity differential) at the beginning of the period between each unit and the "leader" spatial unit. In the context of endogenous growth models, spatial units with a technological lag are likely to experience lower productivity growth than that of more developed spatial units. u is a measure of urbanisation, measured by population density and is aimed at capturing the effect of economic activity density. Lastly, d measures the initial level of labour productivity in the manufacturing sector (Angeriz et al. 2008).

This specification is defined with R as follows:

```
## Specify the model to be estimated
verdoorn<-p~q+u+G+d
```

Taking into account spatial spillover effects requires estimating the specification augmented by a spatial autoregressive term (Fingleton 2000, 2001):

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

This specification is theoretically warranted by Fingleton 2000 and 2001and reflects the estimable specification of a model inspired by the New Geographic Economy. For illustration purpose, we also consider an alternative specification that can be linked to a spatial autoregressive error model:

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

where:

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

The estimation of panel data models with R requires the *plm* (panel without spatial autocorrelation, object management `pdata.frame` adapted to the panel) and *splm* (estimate and tests for spatial panels) packages. Packages *sp*, *maps* and *maptools* also must be loaded for importing and managing spatial objects.

```
# Packages needed
library(plm)
library(splm)
library(sp)
library(maps)
library(maptools)
```

4. The original analysis of Fingleton 2000, 2001 is based on a cross-section model, we extend it to the case of panel data.

The most common specifications are estimated using `spml` and `spreml` orders for maximum likelihood and `spgm` for the generalised method of moments. These all have a relatively identical structures with additional options depending on the case:

```
# Maximum Likelihood
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"),
                 ...)
```

The first step consists in defining the specification (`formula=...`) without indicating spatial effects (which are defined by the specific options), indicating the name of the `pdata.frame(data=...)` and the `listw` needed to create spatially lagged variables (`listw=...`). The nature of the specific effects is determined by the option `model` — the user may choose between `pooling` for a pooled data model, `within` for a fixed-effect model or `random` for a randomised model. It is also possible to define whether the effects relate to individuals or/and periods using the option `effects` that can be established as equal to `individual`, `time` or `twoways`. We can also choose whether the specification includes spatial terms: `lag=T` in the SAR model, or `lag=F`in all other cases. Lastly, it is possible to choose the nature of the specification in the random effects model: `spatial.error="b"`for a Baltagi specification, `spatial.error="kkp"`for the KKP-style specification (Kapoor et al. 2007) or `spatial.error="none"` in all other cases.

The `spreml` command makes it possible to estimate, by maximum likelihood, more specifications with random effects (`errors=`) with the possibility of considering different configurations including the possibility of introducing serial correlation in the error term. Given the matrix calculations which this entails, it includes multiple options for configuring the calculation algorithm:

```
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, ...)
```

Lastly, the `spgm` command makes it possible to estimate the parameters using the generalised method of moments.

```
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)
```

The specification tests have largely incorporated these options. The Hausman test, which is robust to heteroskedasticity, is activated using the `sphtest` command. The `slmtest` command triggers the implementation of the specification tests for spatial autocorrelation. Specification tests on the error term (random effect, spatial autocorrelation, serial autocorrelation) are run using the `bsjktest` command. These tests are easily interpretable since the alternative hypothesis is always recalled in the output.

7.4.2 Data and spatial weights matrix

Our analysis is based on a sample of 1,032 European regions at the NUTS3 level in 14 member states of the EU15 (only Greece is not present in our sample). The data are available for the period 1991-2008. We aggregate the annual data by periods of 3 years in order to control for short-term economic variations (cycles). We obtain a panel of 6 periods for which we construct growth rates of labor productivity (p) and of gross added value (q) in the manufacturing sector. The estimations are therefore done for 5 periods. Figure 7.1 displays the perimeter of our analysis.

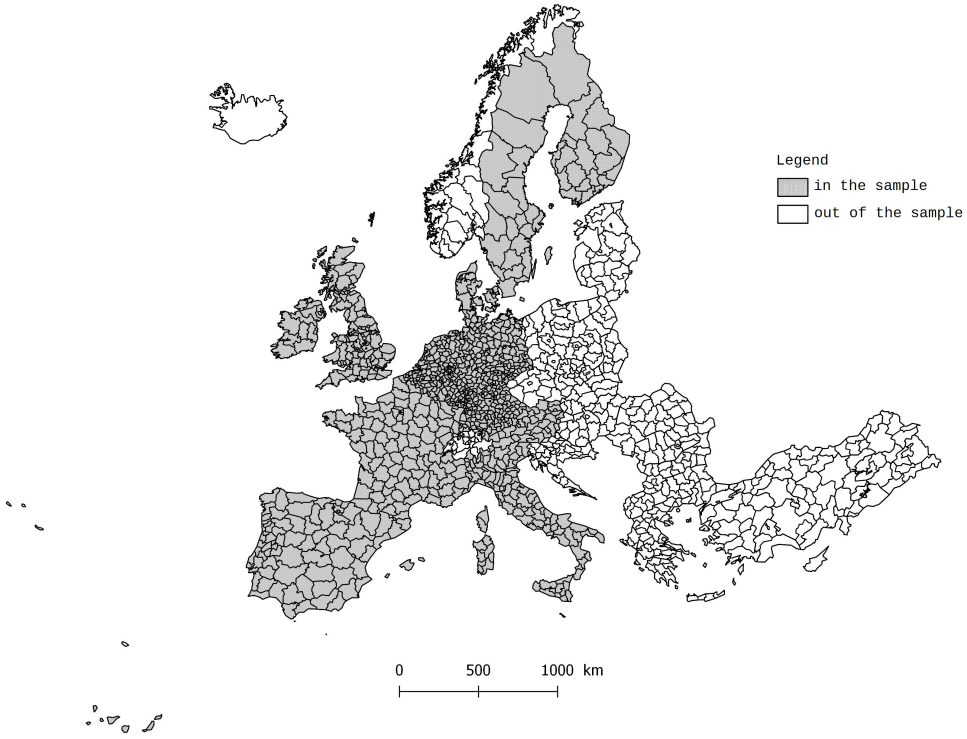

Figure 7.1 – Perimeter of the study

```
# Import data
data_panel <- read.csv("panel_average_3_years_1991_2008.csv", sep=";")
# Import shapefile (Gisco) as a "SpatialPolygonDataFrame"
shape_nuts3<-readShapeSpatial("NUTS_RG_60M_2006")
# Select NUTS3 (by NUTS3 level)
shape_nuts3<- shape _nuts3[shape_nuts3$STAT_LEVL_== 3,]
# Select NUTS3 from the sample
data_panel_code<- data_panel[,"NUTS3"]
shape_nuts3<- shape _nuts3[shape_nuts3$NUTS_ID %in% data_panel_code,]
# Visualising the sample
plot(shape_nuts3)
```

In order to generate a table of descriptive statistics in LaTeXformat of the explained variables and the explanatory variables of the model, it is possible to use package `stargazer` and apply the *stargazer* command on the database including the model variables. The result is shown in 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="Descriptive statistics ")
```

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 – Descriptive statistics

Regarding the spatial weight matrix, as there are islands in the sample (Madeira, Canaries, etc.), a weight matrix based on a criterion other than simple contiguousness due to the presence of a common boundary is required (see chapter 2 : “Codifying neighbourhood structure”). We build a matrix of the 10 closest neighbours to ensure a connection between the regions of Great Britain and continental Europe.

```
# Creation of a k matrix plus close neighbours, 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 The results

To select the most appropriate specification, we start from the model without spatial autocorrelation and implement the Hausman test and the Lagrange multiplier tests.

Table 7.3 shows the results of the estimation of a spatial error autocorrelation model. Column (1) shows the pooled data model while columns (2) and (3) take into account the unobserved individual heterogeneity, respectively, through fixed effects and random effects. Regarding the Verdoorn coefficient, the results are similar: with a significant and positive coefficient greater than 0.5 in all three cases, the presence of increasing returns to scale is confirmed for our sample. Employment growth rate in the manufacturing sector of a region is also all the greater as this region is urbanised (the coefficient associated with u positive and significant in the first and third cases), especially as the gap with the leading region at the beginning of the period is significant (the coefficient associated with G positive and significant in the first and third cases) and even less important as initial productivity is high, which reflects a phenomenon of convergence of labour productivity in the manufacturing sector (the coefficient associated with d is negative and significant in all three cases).

```
# Table 7.3: estimation without consideration for spatial autocorrelation
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	
Model:	pooled data	fixed effects (*within*)	random effects (GLS)
	(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)
Constant	0.146^{***}		0.228^{***}
	(0.012)		(0.014)
Observations	5 160	5 160	5 160
$R^2 ad\,justed$	0.523	0.587	0.552

Table 7.3 – Estimates without consideration for spatial autocorrelation
Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

The results of the standard Hausman test and the Hausman test robust to spatial autocorrelation of errors leads to rejection of the null hypothesis on absence of correlation between individual effects and explanatory variables. For the rest of the empirical analysis, a fixed effects model is thus chosen.

```
# Hausman test (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

# Hausman test robust to spatial autocorrelation (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
```

The results of the Lagrange multiplier tests in a fixed effects model encourages favouring a SEM specification (code to tests below). If the test statistics for taking spatial autocorrelation into account by SAR (Test 1) or SEM (Test 2) confirm the rejection of the hypothesis that these two terms (taken independently) are null, the simultaneous reading does not make it possible to conclude on the most appropriate specification to take spatial autocorrelation into account (these two tests are not included). However, it should be noted that the test statistic for a SEM alternative is higher than that for a SAR alternative. To conclude in a more credible way, robust tests are used in the presence of the alternative specification of spatial autocorrelation (Tests 3 and 4). In other words, the aim is for the RLMlag to test for the absence of a spatial autoregressive term when the model already contains a spatial autoregressive term in the errors (RLMlag), or vice versa for RLMerr to test for the absence of a spatial autoregressive term in the errors when the model contains a spatial autoregressive term. The robust RLMerr version is highly significant (Test 4) while RLMlag is not (Test 3). We therefore estimate a fixed-effect model with an autoregressive spatial process in the errors. In some cases, these last two robust tests do not make it possible to discriminate between a SAR and a SEM. Several possibilities are possible. The first consists in estimating a model containing both these spatial terms (SARAR). The second consists in discriminating between the two specifications on the basis of RLMerr and RLMlag test statistics (by using the specification with the highest associated statistics) or comparing the two specifications' Akaike criteria.

```
# Fixed effects model
# 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		
Model:	pooled data	fixed effects (MV)		fixed effects (MMG)
		Baltagi error	KKP error	
	(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)
Constant	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 of the pooled data model and fixed effects model with spatial autocorrelation of errors
Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 7.4 displays model estimation results taking spatial autocorrelation into account in the form of a spatial autocorrelation of errors. In contrast to the SAR model, the estimated parameters of an SEM are interpreted in traditional manner [5]. The first column shows the pooled data model, while the following three columns show the results from the fixed effects model with different estimation methods (maximum likelihood in columns (2) and (3); MMG in column (4)) and different specifications for the error term (Baltagi in column (2) and KKP in column (3)). In all cases, the autocorrelation coefficient is positive and significant. Regarding the Verdoorn coefficient, it remains

5. It is not necessary to calculate direct, indirect and total effects in an SEM as there is no spatial multiplier effect. However, readers may refer to (Piras 2014) on the calculation of these effects in a static panel SAR.

positive and significant and of greater magnitude than previously. The impact of urbanisation is no longer significant when a fixed effect is introduced. Temporary variations in population density do not significantly affect the growth rate in labour productivity. The effect of urbanisation observed on pooled data is likely due to unobservable characteristics conducive to urbanisation (for instance, first-nature location benefits, Krugman 1999).

```
# Table 7.4: Estimates of pooled-data model and fixed-effect
# model with spatial errors autocorrelation

# Likelihood Maximum estimation
summary(verdoorn_SEM_pool <- spml(verdoorn, data = data_panel,
listw = wknn_10, lag=FALSE,model="pooling"))
# Fixed-effect SEM
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"))
# Generalised moments method estimation
summary(verdoorn_SEM_FE_GM <- spgm(verdoorn, data=data_panel,
       listw = wknn_10, model="within", moments="fullweights",
       spatial.error = TRUE))
```

7.5 Extensions

In this section, we present some extensions of spatial models on panel data. The methods presented in these extensions are not implemented in R at present.

7.5.1 Dynamic spatial models

The models studied at in the previous sections are static models. However, spatial interactions can also be dynamic in nature. For instance, the values used for an observation i at a given point in time t may depend on the values taken by the observations close to i in the previous period. The same type of process may apply for error terms. The dynamic nature can be taken into account by building from Equation 7.6, where time lags are introduced on the explained variable and its spatial lag:

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

This model can be interpreted as a dynamic spatial Durbin model (Debarsy et al. 2012; Lee et al. 2015). In this model, the value of the explained variable used for an observation i over time period t depends on the value of the variable explained for observation i during the previous period (time lag), the value of the variable explained for observations neighbouring i in period t (simultaneous spatial lag) and lastly the value of the variable explained for observations neighbouring i in previous period $t-1$ (delayed spatial offset). For the latter term, one possible route is that of spatial spillover effects — a shock occurring in a zone i at a time period t which spreads to neighbouring zones in subsequent periods. Time lags on explanatory variables X_t or the error term u_t could also be incorporated. However, as Anselin et al. 2008 and Elhorst 2012 show, the parameters of such a

model are not identifiable. Finally, in all generality, this model may include an individual, fixed or random effect. Debarsy et al. 2012 detail the nature of the impacts (direct, indirect, total) in this model. To give an idea of these impacts, the model described is re-written into Equation 7.36 in the following form:

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

The matrix showing the partial derivatives of the expected value of y_t with respect to the k^{th} explanatory variable of X in period t is thus:

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

These partial derivatives reflect the effect of a change affecting an explanatory variable for an observation i on the explained variable of all other observations in the short term only. Long-term effects are defined by:

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

The direct effects consist of diagonal elements of the term to the right of Equation 7.38 or Equation 7.39 and indirect effects such as the sum of the lines or columns of the non-diagonal elements of these matrices. These effects are independent of period t. There is therefore no indirect short-term effect if $\rho = \theta_k = 0$ and there is no indirect long-term effect if $\rho = -\eta$ and if $\theta_k = 0$.

Two main categories of methods have been proposed to estimate this model. On the one hand, based on the principle of maximum likelihood, Yu et al. 2008 build an estimator for the model described by Equation 7.36 including individual fixed effects. This estimator is extended by Lee et al. 2010a for a model that also includes temporal fixed effects. Intuition recommends estimating the model using the maximum likelihood method conditional upon first observation. They also propose a correction when the number of spatial units and the number of periods tend towards infinity. On the other hand, Lee et al. 2010a propose an optimal Generalised Moments estimator based on linear conditions and quadratic conditions. This estimator is convergent, even if the number of periods is small compared to the number of spatial observations.

Readers may refer to Elhorst 2012 or Lee et al. 2015 for a more detailed presentation of the dynamic spatial panel models.

7.5.2 Multidimensional spatial models

In some cases, panel data show a more complex multidimensional structure. For example, in gravity models, economic flows (trade flows, FDI, etc.) between spatial objects (countries or regions) are modelled in three-dimensional panel models by introducing fixed individual, temporal, or even bilateral interaction effects. The introduction of spatial autocorrelation in these gravitational-type models is discussed by such authors as Arbia (2015). The multidimensional structure can also be hierarchical in nature. For instance, European regional data are available on multiple spatial scales: NUTS3, NUTS2, NUTS1, as the NUTS3 regions are intermeshed in the NUTS2 regions, the latter being themselves intermeshed in the NUTS1 regions. In the case of a-spatial panel models, a series of articles from the 2000s (*e.g.* Baltagi et al. 2001) models this hierarchical structure through a distinct specification of random effects. Recently, authors have extended this literature on hierarchical models to the analysis of spatial panels (see Le Gallo et al. 2017 for a review of the recent literature). We present here the general logic of these models.

Formally, given a 3-dimensional panel where the dependent variable is observed according to three indices: y_{ijt} with $i = 1,2,\ldots,N$, $j = 1,2,\ldots,M_i$ and $t = 1,2,\ldots,T$. N is the number of groups. M_i is the number of individuals in group i, such that there are $S = \sum_{i=1}^{N} M_i$ individuals. T represents the number of periods. In general, there may be a different number of individuals between N groups, however, the cylinder structure remains in the panel as regards the time dimension. In the case of a spatial hierarchical structure, it is assumed that index j refers to individuals (for example, in the NUTS3 regions) that are intertwined in N groups (for example, in the NUTS2 regions). Assuming that spatial autocorrelation occurs at the individual level and that the coefficients are homogeneous, the following DSM model can be used:

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

where y_{ijt} is the value of the dependent variable for individual j in group i over period t. x_{ijt} is a vector $(1,K)$ of exogenous explanatory variables, whereas β and θ are $(K,1)$ vectors of unknown parameters, waiting to be estimated. ε_{ijt} is the error term with properties as detailed hereafter. Spatial weight $w_{ij,gh} = w_{k,l}$ is the element $(k = ij; l = gh)$ of the spatial weighting matrix W_S with ij denoting individual j in group i, and similarly for gh. For instance, $k,l = 1,\ldots,S$ and W_S are a dimension weighting matrix (S,S) with the usual properties. ρ is the spatial lag parameter. In general, spatial error autocorrelation can also be specified as an autoregressive model at the individual level:

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

Weight $m_{ij,gh}$ is an element of weight matrix M_S. For the purpose of simplicity, we can assume that $M_S = W_S$. λ is the spatial parameter to be estimated. u_{ijt} is a random composite term that captures the hierarchical structure of the data. To this end, it is assumed that u_{ijt} is the sum of a specific group component α_i that is invariable over time, an individual-group specific component μ_{ij} that is invariable over time and a residual term v_{ijt} :

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

with the following assumptions: (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) $v_{ijt} \overset{i.i.d.}{\sim} \mathcal{N}\left(0,\sigma_v^2\right)$ and (iv) the three terms are independent from one another. Readers may refer to Le Gallo et al. 2017 for estimation methods (maximum likelihood, generalised method of moments), statistical inference and forecasting appropriate for these models.

7.5.3 Panel models with common factors

The major benefit of panel data lies in its modelling unobserved heterogeneity. The models presented above are intended to represent unobserved heterogeneity by using a transformation of variables (fixed effects model) or by setting out assumptions about the structure of the error term (random effects model). In both cases, a restriction is made on the form of the heterogeneity — for each individual, it is constant in the temporal dimension. In other words, there is a total separation of the two individual and temporal dimensions: the individual specific effects vary between individuals but remain constant over time and the specific temporal effects vary over time but remain constant in the individual dimension. While this hypothesis remains credible in the context of short panels, it is too restrictive for panels with a significant time dimension.

In some cases, the databases also include an important time dimension. Common factor models have been developed to take advantage of this data configuration. This new class of models allows to model the effect of unobserved common factors which affect individuals differently, by summarising the information found in the data into a limited number of common factors:

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

where $\sum_{l=1}^{d} \lambda_{il} f_{lt}$ are the common factors in the model. Readers are referred to Bai et al. 2016 for a more precise presentation of this class of models, while our focus is on that which links them to the spatial panels.

By definition, common factors and spatial panels make it possible to capture interactions between individuals. However, they adopt different strategies for this purpose. The spatial econometric models are based on a given structure of interactions between individuals in a panel. This structure is generally constructed from a geographical metric (distance between individuals). In common factor panels, the structure of interactions is not constrained *a priori* (only the number of common factors is constrained).

Initially, spatial panels were used for panels comprising a large number of individuals (relative to the temporal dimension), while the use of the common factor models was preferred when the temporal dimension was large enough to adequately build common factors. Recently, a series of studies has highlighted, through applications, the synergies between the two approaches (Bhattacharjee et al. 2011; Ertur et al. 2015) and proposed methods combining spatial effects and common factors (Pesaran et al. 2009; 2011; Shi et al. 2017a; 2017b). A recent application is proposed by Vega et al. 2016 which studies the development of unemployment disparities between Dutch regions using a model that takes into account spatial and temporal dependencies but also the presence of common factors. Their study emphasises the importance of simultaneously considering these three dimensions (and not using multi-step methods) at the risk of ending up with skewed results. Their results suggest that spatial dependence remains an important factor in understanding the dispersion of regional unemployment rates, even once time dependency and the presence of common factors are taken into account.

Conclusion

Spatial econometrics on panel data is now one of the most active fields in spatial econometrics, both theoretically and empirically. In this context, this chapter has presented the main spatial econometric models on panel data. It is not intended to be exhaustive on all specifications, estimation and inference methods, but has focused on the procedures that can currently be implemented in software R. These procedures concern static panel spatial models, for cylindrical data, with invariable weight matrices over time. Libraries or scripts also exist for proprietary software such as Matlab (commands put forward by Elhorst 2014a) and Stata (module `XSMLE`, Belotti et al. 2017b) and can beneficially supplement the procedures proposed under R.

References - Chapter 7

Angeriz, Alvaro, John McCombie, and Mark Roberts (2008). « New estimates of returns to scale and spatial spillovers for EU Regional manufacturing, 1986—2002 ». *International Regional Science Review* 31.1, pp. 62–87.

Anselin, Luc, Julie Le Gallo, and Hubert Jayet (2006). « Spatial panel econometrics ». *The econometrics of panel data, fundamentals and recent developments in theory and practice*. Ed. by Dordrecht Kluwer. 3rd ed. Vol. 4. The address of the publisher: Matyas L, Sevestre P, pp. 901–969.

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

Bai, Jushan and Peng Wang (2016). « Econometric analysis of large factor models ». *Annual Review of Economics* 8, pp. 53–80.

Baltagi, Badi H, Peter Egger, and Michael Pfaffermayr (2013). « A Generalized Spatial Panel Data Model with Random Effects ». *Econometric Reviews* 32.5, pp. 650–685.

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

Baltagi, Badi H, Heun Song Seuck, and Won Koh (2003). « Testing panel data regression models with spatial error correlation ». *Journal of econometrics* 117.1, pp. 123–150.

Baltagi, Badi H, Seuck Heun Song, and Byoung Cheol Jung (2001). « The unbalanced nested error component regression model ». *Journal of Econometrics* 101.2, pp. 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, pp. 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 and Sean Holly (2011). « Structural interactions in spatial panels ». *Empirical Economics* 40.1, pp. 69–94.

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

Debarsy, Nicolas, Cem Ertur, and James P LeSage (2012). « Interpreting dynamic space–time panel data models ». *Statistical Methodology* 9.1, pp. 158–171.

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

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

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

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

Ertur, Cem and 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, pp. 1481–1498.

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

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

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

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

Kapoor, Mudit, Harry H Kelejian, and Ingmar R Prucha (2007). « Panel data models with spatially correlated error components ». *Journal of Econometrics* 140.1, pp. 97–130.

Kelejian, Harry H and 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, pp. 99–121.

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

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

Le Gallo, Julie and Alain Pirotte (2017). « Models for Spatial Panels ».

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

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

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

Lesage, James and Robert K Pace (2009). *Introduction to spatial econometrics*. Chapman and Hall/CRC.

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

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

Millo, Giovanni and Gianfranco Piras (2012). « splm: Spatial panel data models in R ». *Journal of Statistical Software* 47.1, pp. 1–38.

Mutl, Jan and Michael Pfaffermayr (2011). « The Hausman test in a Cliff and Ord panel model ». *The Econometrics Journal* 14.1, pp. 48–76.

Pesaran, M Hashem and Elisa Tosetti (2009). « Large panels with spatial correlations and common factors ». *Journal of Econometrics* 161.2, pp. 182–202.

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

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

Shi, Wei and Lung-fei Lee (2017a). « Spatial dynamic panel data models with interactive fixed effects ». *Journal of Econometrics* 197.2, pp. 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 and J Paul Elhorst (2016). « A regional unemployment model simultaneously accounting for serial dynamics, spatial dependence and common factors ». *Regional Science and Urban Economics* 60, pp. 85–95.

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

Yu, Jihai, Robert De Jong, and 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, pp. 118–134.