6. Spatial econometrics - common models

JEAN-MICHEL FLOCH
INSEE

RONAN LE SAOUT
ENSAI

Abstract

This chapter describes how to do a spatial econometric study, drawing upon descriptive modelling of the unemployment rate by employment zone. However, spatial models can also be used more broadly, their approach being compatible with any problem in which "neighbourhood" relations come into play. Economic theory characterises many cases of interactions between agents - products, companies, individuals - that are not necessarily geographical in nature. The chapter focuses on the study of spatial correlation, and thereby on these different interactions, discussing the links with spatial heterogeneity, namely spatially differentiated phenomena. There are multiple forms of interaction related to the variable to be explained, the explanatory variables or the unobserved

variables. As a result, these many models end up in competition, all building from the same prior definition of neighbourhood relations. A methodology for selecting the best model (estimate and testing) is thus also detailed step by step. Due to feedback effects, results are to be interpreted in a distinct and more complex manner.

Prior reading of Chapters 1: “Descriptive spatial analysis” and 2: “Codifying the neighbourhood structure” and 3: “Spatial autocorrelation indices” is recommended.

Introduction

The relations between values observed on nearby territories have long been a focus for geographers. Waldo Tobler summed up the problematic in a statement often referred to as the first law of geography: “Everything interacts with everything, but two nearby objects are more likely to do so than two distant objects”. The availability of localised data, combined with the spatial statistics procedures now pre-programmed into multiple statistical software tools, raises the question of how this proximity can be modelled into economic studies. The first step, of course, still consists in characterising this proximity, drawing upon descriptive indicators and tests (Floch 2012a). Once the spatial autocorrelation of the data has been detected, it is time to proceed with modelling in a multi-variable setting. The purpose of this working document is to discuss the practical aspects of conducting spatial econometric studies, *i.e.* selecting the most appropriate model, interpreting the results and understanding the limits of the model.

We will illustrate our presentation with localised modelling of the unemployment rate, using a selection of explanatory variables that describe the characteristics of the labour force, the economic structure, the labour supply and the geographic neighbourhood. The aim is not to detail the results of an economic study [1] but to illustrate the techniques implemented. We will briefly review the definition of a neighbourhood matrix that describes proximity relations and spatial correlation tests (described in greater detail in Chapters 2: “Codifying the neighbourhood structure” and 4. “Spatial autocorrelation indices”). We will then explore specification, estimation and interpretation in detail, within the context of spatial econometric models.

The techniques presented apply to areas beyond the strictly geographical scope. There are many types of data that can be described as interconnected, *i.e.* that can interact with one another, points (individuals or companies the address of which has been identified), data by geographical or administrative zone (localised unemployment rate), physical networks (roads), relational networks (students in a single class) or continuous data (*i.e.* that exist at any point in space). The latter type of data is found mainly in physics, *e.g.* ground height, temperature, air quality, etc. and falls within the scope of geostatistics (see chapter 5: "Geostatistics"). It can nonetheless serve as an explanatory variable in the models presented in this document. It is important to note that we are dealing here with pre-existing proximity structures, which experience little if any change. Thus we will not deal with the characterisation of the formation or development of these neighbourhood relations. On the contrary, we will characterise to what extent spatial (or relational) proximity influences an outcome, by controlling multiple characteristics. Does the unemployment rate depend on neighbouring regions or the price of fuel at nearby stations ? Can non-response to a survey spread spatially? While the majority of applications have a geographical dimension (see Abreu et al. 2004 regarding convergence between regional GDP levels, Osland 2010 regarding determinants of real estate prices described using conventional examples), the fields of application are broader, including for example measuring peer effects in social networks (*cf.* Fafchamps 2015 for a summary view), ideological

1. Blanc et al. 2008 do so in detail, using a spatial econometric model for France, as Lottmann 2013 does, for Germany.

proximity in political science (Beck et al. 2006) or how to take into account proximity between products to study substitution effects in an industrial economy (Slade 2005). At INSEE, these methods have been used to study the relationship between real estate prices and industrial risks (Grislain-Letrémy et al. 2013), changes in places of residence (Guymarc 2015) or non-response to the Employment Survey (Loonis 2012).

Specific tools have been developed to estimate spatial econometrics models. Lesage et al. 2009 offer MatLab programmes. *GeoDa* is a spatial analysis freeware offered as part of a project initiated by Anselin in 2003 for spatial analyses. There are also complementary packages for Stata. However, R remaines the most complete software for estimating spatial econometrics models. All examples and codes herein will thus be presented using this software.

The sequence is organised as follows. Sections 6.1 and 6.2 lay out the economic and statistical rationale behind these models. Section 6.3 describes the stages of estimating a spatial econometrics model. Section 6.4 deals with more advanced technical points. Section 6.5 details implementation under R, as illustrated by a modelling exercise on the unemployment rate by employment zone, before moving on to the conclusion. Readers interested in exploring these methods in greater detail may refer to Lesage et al. 2009, Arbia 2014 or Le Gallo 2002, Le Gallo 2004 for a presentation in French.

6.1 What are the benefits of taking spatial, organisational or social proximity into account?

6.1.1 The economic reasons

Spatial, organisational or social interaction between economic agents has become common in economics. Anselin 2002a lists the following terms used to name these interactions: social norms, neighbourhood effects, peer group effects, social capital, strategic interaction, copy-catting, yardstick competition and race to the bottom, etc. In particular, he highlights two situations of competition between companies justifying the use of a spatial or interaction model.

In the first case, the decision of an economic agent (*e.g.* a company) depends on the decision of the other agents (his competitors). One example is provided by companies competing with each other by quantity (Cournot competition). Firm i wishes to maximise its profit function $\Pi(q_i, q_{-i}, x_i)$ by taking into account its competitors' production levels q_{-i} and its own characteristics x_i which determine its costs. The solution to this maximisation problem is a reaction function such as $q_i = R(q_{-i}, x_i)$.

In the second case, the decision of an economic agent depends on a scarce resource. Using the same example of an industrial firm, the profit function is written $\Pi(q_i, s_i, x_i)$ with s_i a scarce resource (which can be natural, for example uranium, or otherwise, for example, an electronic component manufactured by a single firm). Quantity s_i, which will then be consumed by the company, depends on the quantities consumed by the other companies and therefore on their production q_{-i}. This brings us back to the previous reaction function.

This example shows that the use of an interaction model is micro-founded and that the concept of neighbourhood is not necessarily spatial. Depending on the industrial sector, a company's competitors will be those that show proximity in terms of distance (services to individuals, supermarkets) or products sold (Coca-Cola and Pepsi). Anselin 2002a emphasises that these two situations lead to the implementation of the same spatial or interaction model. They are equivalent from an observational point of view. The data generating processes (DGP) are different but provide the same observations. Simple cross-section data are not enough to identify the source of the interaction (strategic quantity competition or resource competition in our example) but they can only confirm its presence and assess its strength. As with conventional econometrics, the effects identified by the model and the data still need to be considered.

In addition, externalities or neighbourhood effects are commonly taken into account (or controlled) using spatial variables such as distance (*e.g.* to the nearest competitor) or indicators aggregated by geographical zone (*e.g.* number of competitors). This type of variable can be interpreted as having spatial lag (*i.e.* function of observations in neighbouring zones), with an *a priori* definition of neighbourhood relations. Spatial econometrics therefore justifies and fosters the widespread use of these empirical choices.

6.1.2 Econometric reasons

The econometric reasons are rooted in the inadequacies of traditional linear modelling (and the associated estimate using the Ordinary Least Squares -OLS- method) when the assumptions necessary for its implementation are no longer valid. Lesage et al. 2009 thus present multiple technical arguments justifying the use of spatial methods. Spatial autocorrelations of residuals with spatial data, *i.e.* dependency between nearby observations are quite common. This dependency in the observations may either impair the OLS method (the estimators will be without bias but less precise, and the tests will no longer have the usual statistical properties), or produce biased estimators. If the model omits an explanatory variable spatially correlated to the variable of interest, then omitted variable bias is said to occur. In addition, comparing multiple spatial econometric models leads to discuss about the uncertainty of the data-generating process, which is never known, and verify the robustness of the results.

There are many econometric reasons for using spatial models, insofar as descriptive analyses highlight local effects and spatial correlations. In applied studies, it is sometimes difficult to link the econometric and economic aspects justifying consideration for spatial dependence, and economic causalities are difficult to establish from spatial econometrics models (Gibbons et al. 2012).

6.2 Autocorrelation, heterogeneity and weightings: a review of key points in spatial statistics

6.2.1 The nature of spatial effects in regression models

Waldo Tobler's famous assertion, quoted in the introduction, sums up the situation in an astute albeit perhaps simplified manner. Anselin et al. 1988, distinguish autocorrelation (spatial dependency) from heterogeneity (spatial non-stationarity). A variety of phenomena, in measurement (choice of territorial breakdown), externalities or *spillover* may cause observations (endogenous variable, exogenous variable or error term) to become spatially dependent. It is deemed that (positive) autocorrelation occurs when there is similarity between observed values and their location. This chapter deals mainly with the methods for taking this spatial correlation into account in regression models detailed in section 6.3. Spatial heterogeneity, meanwhile, refers to phenomena of structural instability in space. This other form of taking space into account is detailed in Chapter 9: "Geographically weighted regression". It is based on the idea that explanatory variables can be the same and yet not have the same effect at all points. The model's parameters are thus variable. The error term may differ by geographical zone. This is referred to as spatial heterogeneity. For example, to define the price index for old real estate in the INSEE-Notaries database, around 300 strata were defined according to the nature of the property (apartment or house) and the geographical area. The price per m^2 of an additional room or another characteristic is assumed to be different depending on the strata involved. The market is segmented.

This "pedagogical" sharing between autocorrelation and heterogeneity should not cause us to lose sight of the interactions between the two (Anselin et al. 1988 ; Le Gallo 2002 ; Le Gallo 2004). It is not always easy to distinguish between the two components, and poorly specifying one could cause the other to also be erroneous. The classic tests for heteroskedasticity (*i.e.* a particular form of heterogeneity on the error term) are affected by spatial autocorrelation, and

vice versa. The spatial autocorrelation tests are affected by heteroskedasticity. There is no simple solution for simultaneously integrating both these phenomena, apart from simply adding territorial indicators to the autocorrelation models. Moreover, the correlation between observed values means that the information provided by the data is less rich than it would be with independent data. In the event of autocorrelation, there is only one realisation of the data generating process. All of this pleads in favour of a preliminary exploratory approach to the data. Depending on the question, the methodology will first deal with the spatial autocorrelation of the observations (*i.e.* the links between nearby units) or the heterogeneity of behaviours (*i.e.* their variability depending on location).

6.2.2 The weight matrix

To measure the spatial correlation between agents or geographical areas, the first step consists in defining *a priori* neighbourhood relations between agents or geographical zones. These relationships cannot be estimated by the model. If we observe N regions, there are $N(N-1)/2$ different pairs of regions. It is therefore not possible to identify correlation relations between these N regions without making assumptions as to the structure of that spatial correlation. Given N agents or geographical zones, this means defining a square matrix with size $N \times N$, known as the neighbourhood matrix and listed as W, whose diagonal components are null (no element can be its own neighbour). The value of the non-diagonal elements is determined by expert analysis. Numerous neighbourhood matrices have been proposed in the literature. Their construction using software R is detailed in chapter 2: "Codifying Neighbourhood Structure".

6.2.3 Exploratory methods

Before specifying a spatial econometrics model, it is important to ensure that there is indeed a spatial phenomenon to be taken into account. This begins with characterising spatial autocorrelation using graphical representations (map) and statistical tests, as described in Chapter 3: "Spatial autocorrelation indices".

The main indicator [2] is the Moran indicator, which measures the overall association:

$I = \frac{N}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij}(y_i-\bar{y})(y_j-\bar{y})}{\sum_i (y_i-\bar{y})^2}$, with w_{ij} the weight of the coefficient located on the i-th line and j-th column of neighbourhood matrix W. The boundaries of the Moran indicator I are between -1 and 1 and depend on the weight matrix used. The upper limit is in particular equal to 1 where there is straight-line standardisation of the matrix, while the lower boundary remains different from -1. A positive correlation means that areas with high or low values for y group together, and a negative correlation that close geographical zones have very different y values. Under the assumption H_0 that there is no spatial autocorrelation ($I = 0$), the statistic $I^* = \frac{I-E(I)}{\sqrt{\mathbb{V}(I)}}$ asymptotically follows a normal law $\mathcal{N}(0,1)$. Rejecting the null hypothesis of the Moran test therefore amounts to finding spatial autocorrelation. This test of course depends on the choice of neighbourhood matrix W. In addition, rejecting H_0 does not mean that a spatial econometrics model is necessary but that it should be considered. It can only reflect the spatial distribution of an underlying variable. For example, if the underlying model is $Y = X \cdot \beta + \varepsilon$ with β a parameter to be estimated, $\varepsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ and X a spatially autocorrelated variable, a Moran test will show the spatial autocorrelation of variable Y. However, the linear model between Y and X is not a spatial model, and can be estimated conventionally using OLSs.

Local indicators (by geographic region) i, referred to as LISA for *Local Indicators of Spatial Association* have been defined to measure the propensity of a zone to group high or low values

2. The indicators put forth by Geary and Getis and Ord, as well as the other local indicators, are presented in Floch 2012a.

of y or, on the contrary, very diverse values. Their calculation is detailed in Chapter 3: "Spatial autocorrelation indices".

6.3 Estimating a spatial econometrics model

6.3.1 The galaxy of spatial econometrics models

Elhorst 2010 has established a classification of the main spatial econometrics models, based on the three types of spatial interaction derived from the founding model by Manski 1993b :

— an endogenous interaction, when the economic decision of an agent or geographical zone will depend on the decision of its neighbours;
— an exogenous interaction, when an agent's economic decision will depend on the observable characteristics of its neighbours;
— a spatial correlation of the effects due to the same unobserved characteristics.

This model is written in matrix form [3] :

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

with parameters β for exogenous explanatory variables, ρ for the endogenous interaction effect (of dimension 1) referred to as spatial autoregressive, θ for exogenous interaction effects (of dimension equal to the number of exogenous variables K) and λ for the spatial correlation effect of errors known as spatial autocorrelation. In the rest of the document, we will use the term *spatial correlation* to refer to one of these 3 types of spatial interaction.

The model offered by Manski 1993b is not identifiable in this form, *i.e.* β, ρ, θ, and λ cannot be estimated at the same time. We will use his example of peer effects to offer an intuition of this. Let us assume that the poor academic performance of a class can be explained by its social composition (exogenous interaction) as well the poor teaching quality (unobserved characteristics). While there will be a strong correlation between student performances within the class, this cannot be assumed to mean that being alongside pupils with lower academic performance levels (endogenous interaction) has an effect.

To make the model identifiable, a first solution is to assume that neighbourhood matrices W are not identical for all three spatial interactions. For example, some neighbourhood relations will be defined by W_ρ reflecting the autoregressive parameter and W_λ reflecting the spatial autocorrelation. Slade 2005 defines two separate neighbourhood matrices to study price effects in industrial economy: W_ρ being a function of the distance between competing companies and W_X a proximity indicator between the products sold. Another solution consists in removing one of the 3 forms of spatial correlation represented by parameters ρ, θ and λ. This is the preferred solution in the empirical literature.

3. For simplification purposes, the model constant is included here in the matrix of explanatory variables X. In the case of a contiguity matrix, $W \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ represents the number of neighbours of each observation. If the number of neighbours is the same for all individuals, the constant β_0 and the term $W \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \cdot \theta_0$ cannot be identified separately. Moreover, the number of neighbours (or the average number if the neighbourhood matrix is standardised by line) does not necessarily have a clear economic meaning. This is why the literature contains a presentation of the models where the constant is not included in the matrix of explanatory variables X.

The neighbourhood matrix must comply with multiple technical constraints (Lee 2004 ; Elhorst 2010) to ensure, in particular, the inversibility of matrices $I-\rho W$ and $I-\lambda W$, and the identification of models. It can be noted that the usual patterns of contiguousness or inverse distance comply with these constraints. This is not necessarily the case with "atypical" matrices created, for example, for social proximity relations. For example, it is not possible to have only islands (zones that have no neighbours) or, on the contrary, a model in which everyone is everyone else's neighbour. We must also assume that $|\rho|<1$ and $|\lambda|<1$ (criteria that can intuitively be likened to stationarity conditions for ARMA-type models).

Three main types of models can be deduced from the model proposed by Manski 1993b, depending on the constraint used, $\theta=0$, $\lambda=0$ or $\rho=0$.
The $\rho=0$ case (SDEM model, *Spatial Durbin Error Model*) can be considered if it is assumed that there is no endogenous interaction and that the emphasis is placed on neighbourhood externalities. This model nevertheless remains less frequently used (LeSage 2014).

If we assume that the model is such that $\theta=0$, we find the Kelejian-Prucha model (also referred to as SAC, *Spatial Autoregressive Confused*, Kelejian and Prucha 2010 for the heteroskedastic model):

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

The estimators of β in the Kelejian-Prucha model are flawed in that they are biased and not convergent when the real model includes exogenous interactions WX (Lesage et al. 2009). In this instance, there is omitted variable bias. In addition, Le Gallo 2002 emphasises that choosing the same neighbourhood matrix W for this model results in weak parameter identification.

In contrast, if we assume that the model is such that $\lambda=0$, $Y=\rho \cdot WY + X \cdot \beta + WX \cdot \theta + \varepsilon$, known as the SDM (*Spatial Durbin Model*), then the estimators will be unbiased (and the test statistics valid) even if, in reality, we are in the presence of spatially auto-correlated errors (SEMs). This model is therefore more robust in the face of a poor specification choice.

These two models - Kelejian-Prucha and SDM - include specific sub-models, *i.e.* the autoregressive spatial model (SAR, *AutoRegression spatial*): $Y=\rho \cdot WY + X \cdot \beta + \varepsilon$) and the model with spatially auto-correlated errors (SEM, *Spatial Error Model*) : $Y=X \cdot \beta + u$ and $u=\lambda \cdot Wu + \varepsilon$). To derive the latter from the Durbin spatial model, we establish $\theta=-\rho\beta$ (so-called common factor hypothesis). In this case, the SDM model is written: $Y=X \cdot \beta + \rho \cdot W(Y-X \cdot \beta) + \varepsilon$. By noting $u=Y-X \cdot \beta$, it results in the SEM model. The model with exogenous interactions (noted SLX, *Spatial Lag X*) reflects the case $\lambda=\rho=0$ and $\theta \neq 0$.

Furthermore, there are general versions of these models, which allow a variation in neighbourhood effects according to the order of the neighbourhood or according to the interactions taken into account. They are spatial versions of time models *ARMA(p,q)*.

Not all of these models are presented in an economic study. The statistical criteria and consistency with the economic question to be addressed help determine when one specification should be selected over another.

6.3.2 Statistical criteria for model selection

Two main approaches were used to determine the selection of models. These "practical" approaches are based on the assumption that the neighbourhood matrix is known and that the explanatory variables are exogenous. Under the normality assumption of residuals $\varepsilon \overset{i.i.d.}{\sim} \mathcal{N}\left(0,\sigma^2\right)$,

they are based on an estimate by maximum likelihood of the models and the related statistical tests [4]. The first so-called *bottom-up* approach (figure 6.1) consists in starting with the non-spatial model (see Le Gallo 2002 for a summary). The Lagrange multiplier tests (Anselin et al. 1996 for the SAR and SEM model specification tests, robust to the presence of other types of spatial interactions), then make it possible to choose between the SAR, SEM or non-spatial model. This approach was widely-favoured until the 2000s because the tests developed by Anselin et al. 1996 are based on the residuals of the non-spatial model. They are therefore inexpensive from a computational point of view. Florax et al. 2003 have also shown, using simulations, that this procedure was the most effective when the real model is a SAR or SEM model.

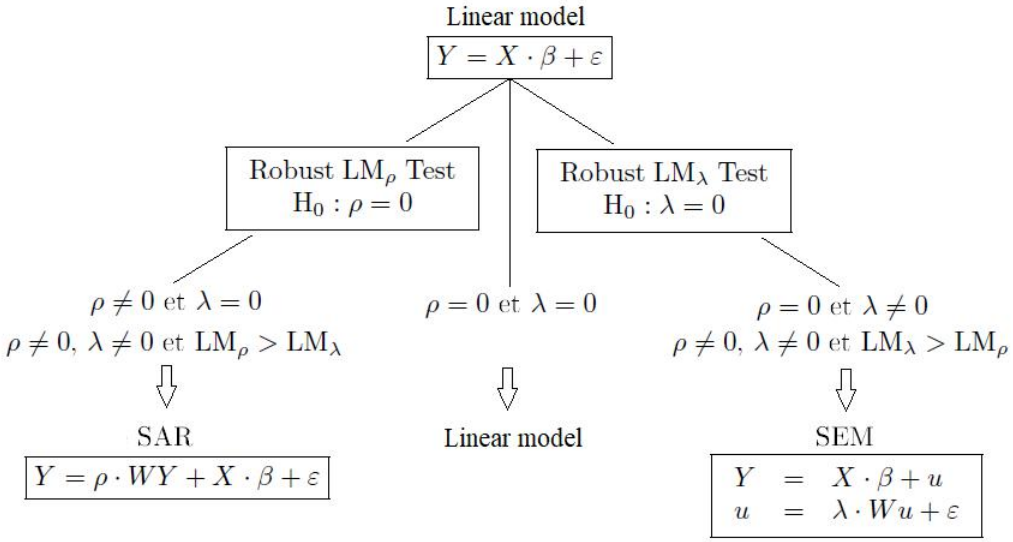

Figure 6.1 – The *bottom-up* approach
Source : *Florax et al. 2003*

The second so-called *top-down* approach (figure 6.2) consists in starting from the Durbin spatial model. Based on the tests of the likelihood ratio, the model most suitable for observations is deducted. Improving IT performance has made it easy to estimate these more complex models, including Durbin's spatial model, used as a reference in the book by Lesage et al. 2009.

Elhorst 2010 proposes a "combined" approach represented in Figure 6.3. It consists in starting with the bottom-up approach but, in the event of spatial interaction ($\rho \neq 0$ or $\lambda \neq 0$), instead of directly choosing a SAR or SEM model, studying the Durbin spatial model. This approach then confirms, using multiple tests (Lagrange multiplier, likelihood ratio), the relevance of the chosen model. It also allows exogenous interactions to be integrated into the analysis. Lastly, if there is any doubt, the model that appears *a priori* the most robust (the Durbin spatial model) is chosen. Let consider the case where, from the residuals of the OLS model, the Lagrange multiplier tests (LM_ρ and LM_λ) [5] it is concluded that there is an autoregressive term, *i.e.*$\rho \neq 0$ and $\lambda = 0$ (left branch of Figure 6.1). The SDM model is then estimated, and, using a likelihood ratio test ($\theta = 0$), the choice is made between the SAR model and the SDM model. If the tests conclude that residual

4. Other estimation methods exist. In the case of endogenous explanatory variables, Fingleton et al. 2008 and Fingleton et al. 2012 propose an estimation by instrumental variables and the generalised method of moments. Lesage et al. 2009 propose a Bayesian estimate. Lastly, to relax the parametric framework, Lee 2004 suggests quasi maximum likelihood estimations.

5. There are two versions of these tests, one robust in the presence of other forms of spatial correlation, the other that is not (Anselin et al. 1996).

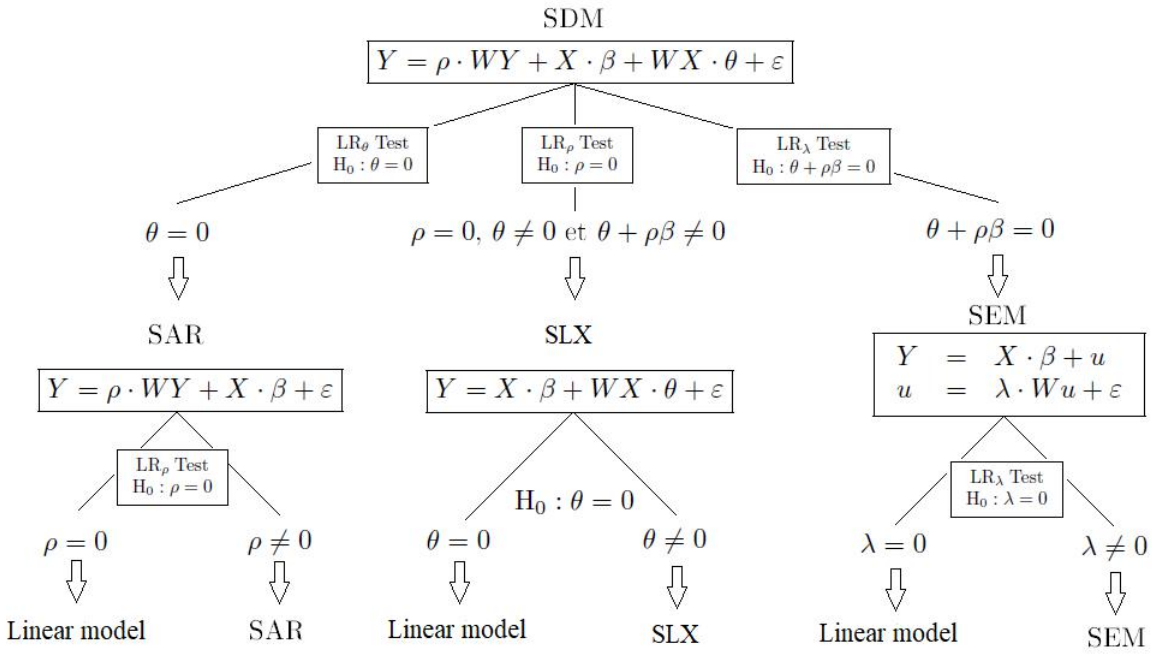

Figure 6.2 – The *top-down* approach
Source :*Lesage et al. 2009*

autocorrelation is present, *i.e.* $\rho = 0$ and $\lambda \neq 0$ (right branch of the Figure 6.2), then the SDM model can be brought back ($\rho \neq 0$ and $\theta \neq 0$), followed by a test of the likelihood ratio on the common factor hypothesis ($\theta = -\rho\beta$) to choose between SEM and SDM. If the tests point to the absence of a spatial correlation, *i.e.*$\rho = 0$ and $\lambda = 0$, then the exogenous interactions model (SLX) should be estimated. Likelihood ratio testing makes it possible to choose between the OLS, SLX and SDM models. Lastly, in the event that the tests conclude that there is both endogenous and residual correlation, *i.e.*$\rho \neq 0$ and $\lambda \neq 0$, the SDM model is estimated.

The dimension of neighbourhood matrix W is the square of the number of observations. However, calculating the likelihood of these spatial models in particular brings certain determinants into play, including this matrix. The computational cost can therefore be substantial when the number of observations becomes high. Lesage et al. 2009 devote a chapter to the computational issues at stake - and methods for successfully addressing them - associated with estimating these models. In practice, the number of observations is often limited to a few thousand.

These rules must not be considered as intangible [6], but rather as good practice. There is no point in directly estimating a SAR model, which is complex to interpret, if neither economic nor statistical analysis justify it.

6.3.3 When interpreting results, beware of feedback effects

Spatial econometrics deviates from the usual linear model framework when spatially shifted variables WY are found in the model. However, the conventional interpretation of linear models remains valid if only the spatial autocorrelation of errors is taken into account (SEM model).

In the presence of a spatially lagged variable WY, the parameters associated with the explanatory variables are not interpreted as in the usual framework of the linear model. This is because, due

6. The sequential testing approach can also lead to a bias as the rejection zone in the likelihood ratio (LR) tests should theoretically take into account the Lagrange multiplier (LM) pre-tests.

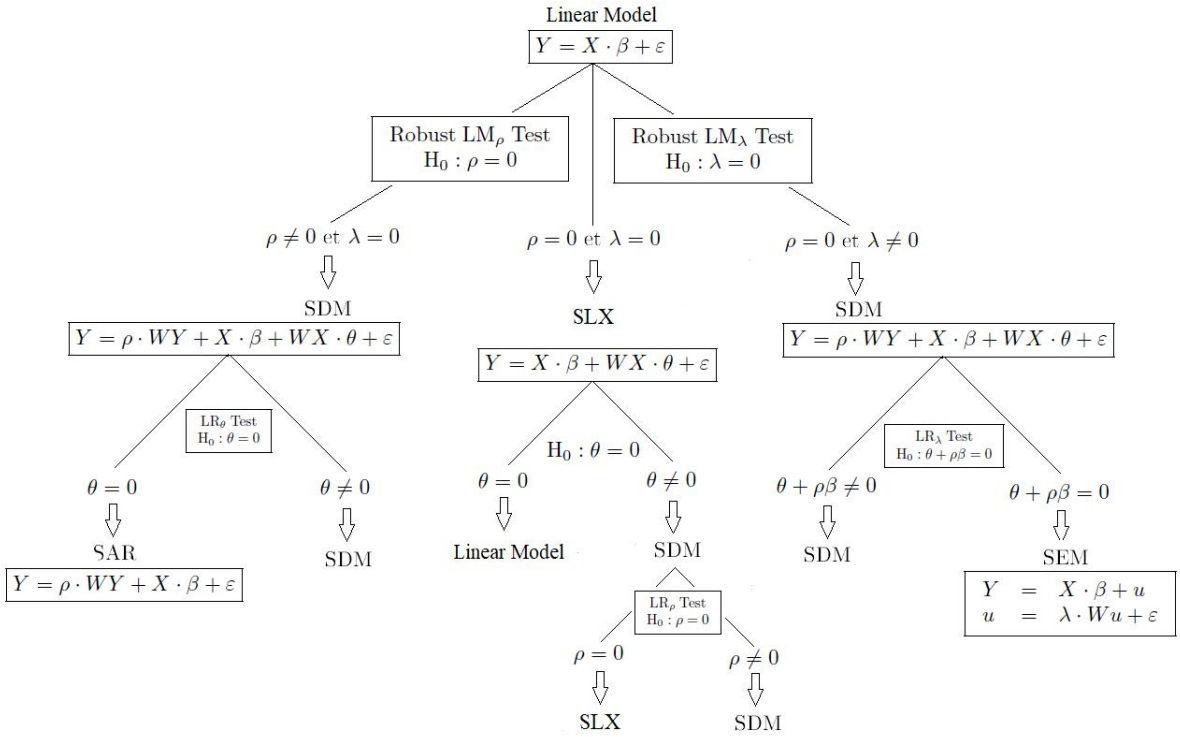

Figure 6.3 – Approach proposed by Elhorst 2010 for choosing a spatial econometric model
Source: Elhorst 2010

to spatial interactions, the variation of an explanatory variable for a given zone directly affects its result and indirectly affects the results of all other zones. The estimated parameters are then used to calculate a multiplier effect that is global in that it affects the whole of the sample.

In contrast, the interpretation of the parameters associated with the explanatory variables remains identical when the model includes only the autocorrelation of errors (SEM model). In this case, there is an overall diffusion effect stemming from spatially auto-correlated errors: the variation of an explanatory variable for a given zone directly affects its results and indirectly affects the results of all other zones, but without the value of this effect being multiplied.

When looking at models with spatially lagged explanatory variables (SLX), the parameters associated with the explanatory variables make it possible to calculate a local effect insofar as the variation of an explanatory variable directly affects its result and indirectly the result of the neighbouring zones, but not that of the neighbouring zones of those neighbours.

To formalise the various impacts, we use the framework defined by Lesage et al. 2009.

The SAR model is $Y = \rho \cdot WY + X\beta + \varepsilon$. It can be rewritten in several ways, writing r as the index for an explanatory variable and S_r as the square matrices of the size of the number of observations:

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

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

The predicted value is therefore $\hat{y} = (1-\hat{\rho}W)^{-1} X\hat{\beta}$ [7] and not $X\hat{\beta}$ as in a classic linear model.

Moreover, $E(y) = (1-\rho W)^{-1} X\beta$. The marginal effect (for a quantitative variable) of a change in variable X_r for individual i is not β_r but $S_r(W)_{ii}$, the diagonal rank value i of matrix S_r. Unlike the time series in which there is only one direction to consider (y_t depends on y_{t-1}, which is explained only by past values), spatial econometrics is multi-directional. A change in my territory affects my neighbours, which in turn affects me. This must be taken into account in the overall analysis of the results.

Furthermore, the marginal effect appears different for each zone [8]. The diagonal terms of matrix S_r are the direct effects, for each zone, of a change in variable X_r in the same zone. The other terms represent indirect effects, *i.e.* the impact changing variable X_r in one zone can have on another zone. For all zones (overall level) it is thus possible to calculate the direct and indirect effects found by averaging these effects (Lesage et al. 2009):

— The average direct effect is the average of the matrices' diagonal terms S_r, *i.e.* $\frac{1}{n}\text{trace}(S_r)$. This indicator can be interpreted in a way similar to that of the β coefficients of a non-spatial linear model calculated using the OLS method.

— The average total effect is an average of all the terms in matrix S_r, $\frac{1}{n}\sum_i [\sum_k S_r(W)_{ik}]$. It can be interpreted in two ways, *i.e.* as the average of n effects across a zone i due to the modification of a unit of variable X_r in all zones, *i.e.* $\sum_k S_r(W)_{ik}$ (the sum of the straight-line terms of matrix S_r), or as the average of the n effects from modifying a unit of variable X_r in a zone i across all zones, *i.e.* $\sum_k S_r(W)_{ki}$ (the sum of the terms in the column of matrix S_r).

— The average indirect effect is the difference between the average total effect and the average direct effect.

The indicators are identical for the Kelejian-Prucha model. Such indicators can be defined for SDM model $Y = \rho \cdot WY + X\beta + WX \cdot \theta + \varepsilon$, but their calculations must take into account exogenous interactions $WX \cdot \theta$. In the case matrix $S_r(W)$ is written $(1-\rho W)^{-1}(I_n\beta_r + W\theta_r)$, instead of $(1-\rho W)^{-1}\beta_r$ in the case of the SAR model.

When an exogenous interaction $WX \cdot \theta$ is found but no endogenous interaction is (SLX and SDEM models), the direct effect of a variable X_r is β_r, while the indirect effect is θ_r.

In all cases, calculating the accuracy of these estimators is quite complex. In this regard, Lesage et al. 2009 draw upon Bayesian simulations of Markov Chain Monte Carlo methods (MCMC) [9].

Moreover, these effects depend first and foremost on the nearby neighbourhood. For the SAR model, it should be noted that the average direct effect is greater in absolute value than the marginal effect of the non-spatial linear model, $|S_r| > |\beta_r|$. The diagonal terms of neighbourhood matrix W are null. Decomposition into whole series $(1-\rho W)^{-1} = \left(I_n + \rho W + \rho^2 W^2 + \cdots\right)$ shows that the first feedback term (which dominates the other higher order terms) is proportional to ρ^2. The analysis of effects by neighbourhood order (distinguishing the direct effect, the effect of neighbours, neighbours of neighbours, etc.) is also elaborated upon byLesage et al. 2009.

7. This is not the optimal prediction, see Thomas-Agnan et al. 2014 for optimal prediction of a SAR model.

8. This characteristic is found for the marginal effect of a Probit model, for example. The model is $E(Y|X) = \mathbb{P}(Y=1|X) = \Phi(\beta X)$ with Φ the distribution function of a standard normal distribution. The marginal effect of a variable X_r is then $\beta_r \cdot \varphi(\beta X)$ and therefore differs for each individual. One solution thus consist in estimating the average marginal effect $\beta_r \cdot \overline{\varphi(\beta X)}$.

9. Markov Chain Monte Carlo methods are sampling algorithms that make it possible to generate samples of a complex probability law (to deduce the accuracy of a statistic, for instance). They are based on a Bayesian frame and a Markov chain, the boundary law of which is the distribution to be sampled.

In conclusion, for the overall interpretation of an endogenous interaction model, it is helpful to calculate, for each variable, the average direct effect ($\frac{1}{n}\text{trace}\,(S_r)$) and the average indirect effect ($\frac{1}{n}\left[\sum_j \sum_k S_r\,(W)_{kj} - \text{trace}\,(S_r)\right]$). Calculating the effect caused by space ($\frac{1}{n}\text{trace}\,(S_r) - \hat{\beta}_r$) also illustrates the impact of the feedback effects.

6.4 Econometric limits and challenges

6.4.1 What to do with missing data?

In conventional econometrics, a sample of n individuals is observed. If values are missing on some individuals, they are generally excluded from the analysis. If there are no selection issues due to non-response (the non-response process is independent from the variables in our model), this reduces the size of the sample but does not prevent the econometric methods from being implemented.

In spatial econometrics, there is only one realisation of the data-generating process (an analogy can be made with time series here, with the parameters of an ARMA model being estimated using a single time path). If the observation of the spatial distribution is incomplete (there are missing values), the model cannot be estimated. One solution consist in interpolating the missing values using geostatistical techniques (Anselin 2001). However, this leads to measuring variables with errors [10], or using an appropriate estimate (*e.g.* EM expectancy-maximisation algorithm, Wang et al. 2013b for the SAR model). However, these solutions are only possible when the percentage of missing values is small.

Another implication is that it is not easy to implement these techniques on individual survey data. In general, spatial econometrics is not suited to survey data. In this case, only partial neighbourly relations can be observed, and only for the individuals surveyed. We must then make the complementary and very strong hypothesis that the observations of the unsurveyed neighbours are exogenous, *i.e.* that they do not change the neighbourhood effects solely for the individuals surveyed. Lardeux and Marly-Alpa 2016 show that it is not possible to detect the spatial correlation generated by a SAR model only for a geographical cluster sampling plan. With low sampling rates and conventional sampling plans (stratified or systematic), only direct effects can otherwise be estimated. This point is elaborated upon in Chapter 10: “Spatial econometrics on survey data”.

6.4.2 Choosing the weight matrix

When defining a neighbourhood matrix, the constraints faced are strong, as the description sought must be simple - so that the model is identifiable - yet also accurately reflect the links between territories. Many authors emphasise how sensitive results are to the choice of matrix (Corrado et al. 2012 ; Harris et al. 2011), while Lesage et al. 2009 consider these findings to result from a poor interpretation of the models, stating that this assumed sensitivity to weight matrix is “the greatest myth” in spatial econometrics. They claim that direct and indirect effects are more robust to the choice of W than parameter estimators, which do not have an immediate interpretation. Nevertheless, we can subscribe to the remark from Harris et al. 2011: “Spatial econometrics emphasises the importance of selecting matrix W but gives us little information on the criteria for making this choice”. These difficulties that have contributed to the scepticism of several economists (Gibbons et al. 2012). These considerations show the complexity of matrix determination W, which remains a subject of scientific controversies.

We have seen that the models generally treat matrix W as exogenous. However, other methods draw upon the data used to determine the weight matrix. Aldstadt et al. 2006 define a matrix

10. Interpolation can also be useful when the geographical levels used to measure the variable to be explained and the explanatory variables are different, for example the known prices of housing at the address or municipality level and atmospheric pollution indicators measured using sensors whose location differs.

construction algorithm W from local indicators of spatial autocorrelation on variables of interest. Weights can also be estimated using econometric models with functional constraints that are low *a priori* (Bhattacharjee et al. 2013). The latter's approaches often entail calculation processes that are cumbersome and more difficult to implement. Moreover, a more realistic description that is more in line with economic reality may generate endogeneity. Research involving endogenous matrices has recently been proposed (Kelejian et al. 2014).

Lastly, the matrix W is considered fixed, which restricts the economic analysis framework. For example, in the case of a neighbourhood matrix measuring the distance between companies or products, Waelbroeck 2005 emphasises that the arrival (or departure) of a company or product is an endogenous event that should lead to changes in neighbourhood relations, which the usual methodology cannot take into account.

6.4.3 What if the phenomenon is spatially heterogeneous?

There are two forms of heterogeneity.

The first is heteroskedasticity. The model's parameters are the same but its individual variability (the variance of the error term) is not. Spatial autocorrelation of errors $(I-\lambda W)^{-1}\varepsilon$ (SEM model) can be interpreted as a spatial random effect (it is assumed that the individual effects within a neighbourhood are similar, as the fixed effects cannot be estimated) and therefore as a particular form of heteroskedasticity and spatial correlation (Lesage et al. 2009). An alternative solution to a spatial econometrics model would be to define the form of heteroskedasticity and the spatial correlation of the variance-covariance matrix (Dubin 1998) to define spatial clusters (Barrios et al. 2012) or adopt a Newey-West type spatial correction (Flachair 2005). Lastly, recent developments in spatial econometrics relax the hypothesis of homoskedasticity of the residuals ε from the models presented in this introduction. Kelejian et al. 2007Kelejian et al. 2010 proposed for instance a parametric HAC-type method (*Heteroskedasticity and Autocorrelation Consistent*), derived from time series, and a non-parametric method.

In the presence of heteroskedasticity, the estimators remain convergent but the test statistics are no longer distributed according to the usual laws. The spatial autocorrelation tests are therefore no longer reliable. *In contrast*, in the presence of spatial autocorrelation, the usual heteroskedasticity tests (*White*, *Breusch-Pagan*) are also no longer valid. Le Gallo 2004 presents joint spatial heteroscedasticity and autocorrelation tests.

The second form of heterogeneity relates to the spatial variability of the parameters or functional form of the model. When the territory of interest is well-known to researchers, it is often addressed in empirical literature by adding indicators of geographical zones in the model - possibly crossed with each explanatory variable - and thus estimating the model for different zones or by conducting tests of geographical stability of the parameters (known as the Chow test). When the number of these geographical zones increases, this treatment nevertheless reduces the number of degrees of freedom and therefore the accuracy of the estimators. More complex methods commonly used in geography have been developed (Le Gallo 2004). They remain to a large extent descriptive and exploratory (in particular through graphical representations), as their theoretical properties are partially known, and particularly as regards convergence properties and the inclusion of breaking points.

There are also geographical smoothing methods where the constant (or even each explanatory variable) is crossed with polynomials that are a function of geographical coordinates. Flachaire (2005) offers a partial (and alternative) linear model $Y_i = X_i\beta + f(u_i, v_i) + \varepsilon_i$, where f refers to a functional form dependent on geographical coordinates u_i and v_i (or even other explanatory variables if proximity is not spatial but social, or between products, for example). It shows that, like a SAR model, the f can be interpreted as a weighted sum of endogenous variables Y. This analysis thus highlights that spatial correlation and heterogeneity are linked.

There are also local regression methods whose extension to the spatial context is formalised within the framework of geographically weighted regression (Brunsdon et al. 1996). These methods are detailed in Chapter 9: "Geographically-weighted regression".

However, it remains difficult to distinguish spatial heterogeneity and correlation. To our knowledge, there is no method for distinctly identifying these two phenomena. Pragmatic approaches are therefore adopted. Le Gallo 2004 offers an application to crime in the United States. Using heteroskedasticity tests (robust to the presence of autocorrelation), it highlights the presence of distinct spatial regimes between two geographical zones, East and West. A SAR model is then estimated, for which the explanatory variables X are crossed with the two spatial regimes, and variances are assumed to be different between these two zones. Osland 2010 studies real estate prices in Norway using spatial econometric models, semi-parametric smoothing and weighted geographical regression models. The various approaches provide additional results but are not integrated into a single modelling.

6.4.4 The risk of "ecological" errors

The methods presented in this document are based on predefined geographical zonings (an employment zone in our example). Many economic variables are only available for the administrative divisions of the territory. However, this administrative division does not necessarily correspond to the economic reality of relations between agents. This geographical phenomenon is known as the MAUP (*Modifiable Areal Unit Problem*). It implies several consequences (Floch 2012). With different scales or breakdowns, the results of the models and interactions between agents are not identical. The spatial scope of the zones must also be taken into account: 1 000 economic agents do not interact in the same way in 1 km^2 or in 10 000 km^2. Where individual data are available (*e.g.* employment characteristics from population census rather than unemployment rates by employment zone), it is possible to disregard this administrative breakdown or build the geographical level *a priori* deemed most relevant. However, in general, there is no solution to the problem of the MAUP.

Moreover, the data used are often aggregated, in the sense that they represent the average of our variables of interest on a geographical zone. In conventional econometrics, the use of aggregated data, known as *ecological regression*, causes identification and heteroskedasticity problems. Anselin (2002) provides an example of a model where the decisions of an individual i, y_{ik}, are explained by that individual's characteristics x_{ik} as well as by the characteristics of group k, to which the individual belongs $\bar{x}_k = \sum_i x_{ik}/n_k$. The model is written $y_{ik} = \alpha + \beta \cdot x_{ik} + \gamma \cdot \bar{x}_k + \varepsilon_{ik}$ where β represents the individual effect and γ the context effect. If the only data available are per group (*e.g.* average scores of a class on a test, rather than individual results), the estimated model becomes $\bar{y}_k = \alpha + (\beta + \gamma) \cdot \bar{x}_k + \bar{\varepsilon}_k$. It is then no longer possible to separately identify parameters β and γ. The model is heteroskedastic because $\mathbb{V}(\bar{\varepsilon}_k) = \sigma^2/n_k$ in the case of initial disturbances independent and identically distributed of variance σ^2.

The problem is even more complex in the case of spatial models. It is not possible to aggregate a neighbourhood matrix W defined at the individual level. With individual data, an individual i of group k may have neighbours in group k but also in another group k'. If we now consider an aggregated neighbourhood matrix at group level, intra-group relations will no longer be taken into account (the diagonal is hypothetically null). In addition, there may be many individuals in group k who are neighbours to individuals in group k' but very remote neighbours to another group k''. With a matrix of contiguousness aggregated at the group level, the strength of individual relationships will no longer be taken into account (each neighbour has the same weight). Beyond problems identifying an *ecological regression*, a SAR model defined at the individual level cannot be aggregated to match a SAR model defined at a higher level. There are no simple relations between the parameters.

To understand this issue, let us take the example of the real estate market. The observation deals

with cities in which prices are very high in the centre, then gradually decline. There are also very different price levels between cities. If we only consider average prices per urban centre (grouping nearby cities), the disparity in prices within cities will be hidden. These interlocking scales can generate results that at first appear paradoxical.

In practice, this means that the interpretation of the results is only valid for the chosen geographical breakdown. Studying economic relations at an aggregate level with a spatial model, it is impossible to draw any conclusions about individual relations between agents. To take into account this entanglement between geographical zones (regions, departments, cantons, individuals) and make the analyses consistent between them, one solution consist in carrying out multi-level analyses (Givord et al. 2016). In the case of macroeconomic studies such as regional growth, this problem is less prominent. The aggregate level is the most relevant level.

6.5 Practical application under R

In this section, we detail the practical implementation of a spatial econometric study, modelling the localised unemployment rate (by employment zone, excluding Corsica) using the structural characteristics relating to the characteristics of the labour force (proportion of low-skilled workers and those under 30 in the labour force), the economic structure (proportion of jobs in the industrial sector and the public sector) and the labour market (activity rate). The purpose of this section is not to detail the results of an economic study but to illustrate the techniques implemented, *i.e.* the definition of a neighbourhood matrix that describes local relations between territories, spatial correlation and specification tests, estimation, and the interpretation of spatial econometric models. Other variables can of course explain local unemployment rates (Blanc and Hild 2008, Lottmann 2013). The economic variables are assumed to be structural and with little variability in the short term. To limit endogeneity problems, the unemployment rate is calculated for Year 2013 and the explanatory variables are the 2011 data from the CLAP (Local Knowledge of the Productive Apparatus) and the RP (Population Census). A causal interpretation nonetheless remains impossible. Many variables have been omitted from the analysis, such as the supply of jobs. The explanatory variables taken into account can thus include the effect of such omitted variables, as opposed to only their own effect. Lastly, the time lag between explanatory variables and the unemployment rate does not completely do away with the simultaneous nature of phenomena (for example between the activity rate and the unemployment rate), which are structurally stable in the short term.

Examples and codes are presented using R, the most comprehensive software for estimating spatial econometrics models. Some useful packages in R are listed below :

— *sp* and *rgdal* for importing and defining spatial objects, *maptools* for the definition of cards ;
— functions similar to those of GIS (Geographic Information System) such as distance calculation or geostatistical methods : *fields*, *raster* and *gdistance* ;
— spatial econometrics:*spdep* (spatial dependencies) for all conventional models, and *spgwr* for geographically weighted regression.

6.5.1 Mapping and testing

After importing the data and defining a neighbourhood matrix using the methods presented in section 6.2, the data can be mapped out and an initial analysis carried out on spatial autocorrelation.

Figure 6.4 depicts unemployment rates by employment zone in 2013. Polarised zones appear, which could be a sign of spatial heterogeneity. The North of France and Languedoc-Roussillon thus have higher unemployment rates, while the regions bordering Switzerland have lower ones. The zones continguous to these regions also show similar unemployment rates, which is characteristic of a spatial autocorrelation. As to explanatory variables, a strong polarisation can be seen in particular in the percentage of industrial employment. Employment rates show a spatial structure similar to

the labour force participation rate.

Table 6.1 describes the distribution of variables. The average unemployment rate is 10%, with a labour force participation rate of 73%. 22% of the population consist in low-skilled workers and young workers under the age of 30. Apart from the percentage of industrial employment and public employment, the interquartile gaps are low, below 5%. The percentage of industrial employment appears the most polarised variable.

	N	Mean	Std dev.	Min	Q25	Median	Q75	Max
Unemployment rate (%)	297	10.0	2.4	4.9	8.3	9.6	11.4	17.5
Labour force participation rate (%)	297	72.8	2.6	65.9	71.3	72.8	74.2	81.6
Working-age Low-Skilled Graduates (%)	297	22.1	3.6	13.0	19.5	22.2	24.8	32.2
Working-age Adults 15-30 y.o. (%)	297	21.8	2.0	16.7	20.4	21.8	23.2	27.7
Industrial Employment (%)	297	19.7	8.8	3.7	13.3	18.2	24.8	52.0
Public Employment (%)	297	33.5	6.2	15.0	29.5	33.2	36.9	51.0

Table 6.1 – Sample Description
Note: The geographical zone is the employment zone. Statistics are not weighted.

Spatial autocorrelation tests and advanced graphical representations

The near-null p-value of the Moran test indicates that the null hypothesis assuming no spatial autocorrelation should be rejected (see Chapter 3: “Spatial autocorrelation indices”). The result is robust to the choice of neighbourhood matrix. The raw data autocorrelation can be illustrated graphically using the Moran graph. It links the observed value at one point with that observed in the neighbourhood determined by the weight matrix.

Figure 6.5 is consistent with the results of the Moran test. A linear relationship appears between the unemployment rate of one zone and that of its neighbourhood. A map can be associated so that employment zones be located according to their characteristics (HH means high unemployment in a high environment, HB a high rate in a lower environment). It shows that this relationship is not homogeneous across the territory. The north and south have high unemployment rates. In contrast, a "middle" France will show lower unemployment rates.

6.5.2 Estimation and model selection

The descriptive analysis showed that space was not neutral in characterising local unemployment rates. However, it is not certain that an econometric model taking space into account is needed. The scatter plot showing unemployment and labour force participation rates shows a strong linear relationship between the two variables. The unemployment and activity rates are both spatially correlated. The unemployment rate could therefore be linked to the activity rate, without any form of spatial correlation other than that present in the two variables. First of all, we begin by estimating a non-spatial linear model using the OLSs. A Moran test adapted to the situation of residuals confirms the residual presence of spatial autocorrelation (potentially associated with spatial heterogeneity), regardless of the neighbourhood matrix.

To determine the form of spatial correlation (endogenous, exogenous or unobserved), a pragmatic approach must be taken. The Elhorst 2010 approach would result in adopting the SDM model.

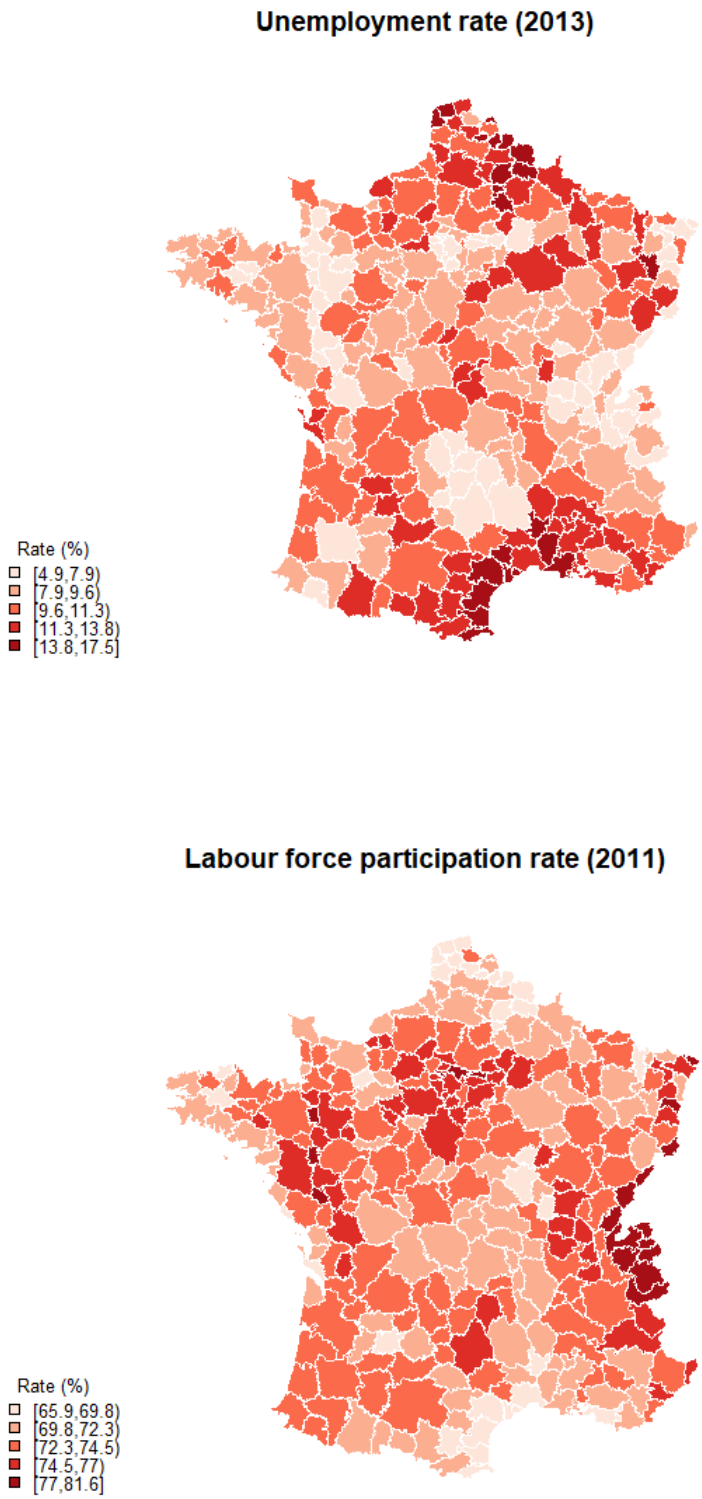

Figure 6.4 – Distribution of unemployment and labour force participation rate, by employment zone

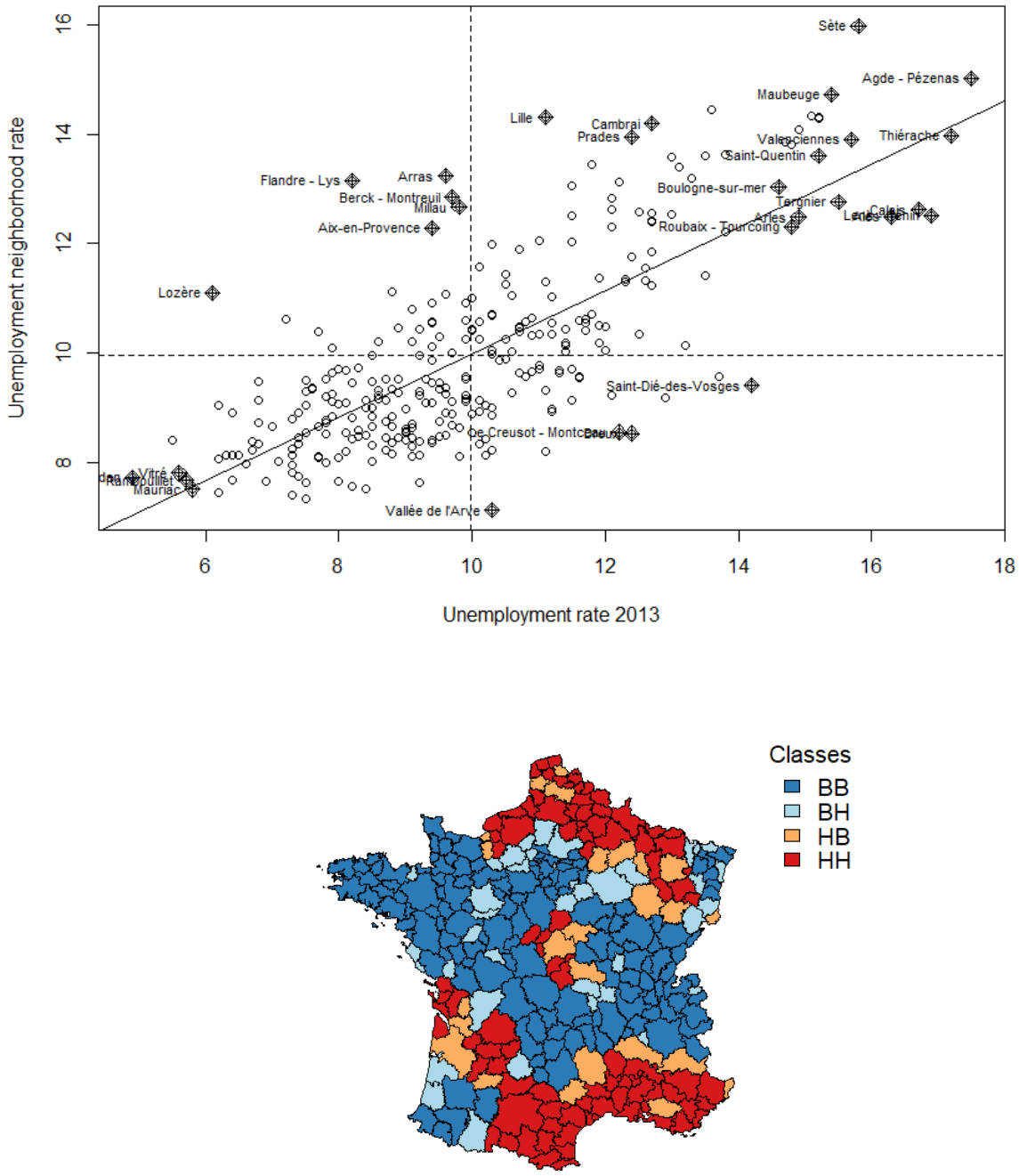

Figure 6.5 – Moran unemployment rate graph and associated map

Only the OLS and SDM models would then be estimated. For educational purposes, all spatial the models are nevertheless estimated for 6 neighbourhood matrices - contiguous, closer neighbours (2, 5 or 10), inverse distance, and proportional to commutes (known as the endogenous matrix). Regressions are estimated using the *spdep* package. The computational cost of estimating these models is also low.

```
### Estimated model
model <- txcho_2013 ~ tx_act+part_act_peudip+part_act_1530+part_emp_ind+
    part_emp_pub
### Neighbourhood Matrix
matrix <- dist.w

### OLS model
ze.lm <- lm(model, data=donnees_ze)
summary(ze.lm)

### Moran test adapted to residuals
lm.morantest(ze.lm,matrix)

### LM-Error and LM-Lag test
lm.LMtests(ze.lm,matrix,test="LMerr")
lm.LMtests(ze.lm,matrix,test="LMlag")
lm.LMtests(ze.lm,matrix,test="RLMerr")
lm.LMtests(ze.lm,matrix,test="RLMlag")

### SEM model
ze.sem<-errorsarlm(model, data=donnees_ze, matrix)
summary(ze.sem)
### Hausman test
Hausman.test(ze.sem)

### SAR Model
ze.sar<-lagsarlm(model, data=donnees_ze, matrix)
summary(ze.sar)

### SDM Model
ze.sardm<-lagsarlm(model, data=donnees_ze, matrice, type="mixed")
summary(ze.sardm)
### Common factor hypothesis test
# ze.sardm: Constraint-free model
# ze.sem: Constrained model
FC.test<-LR.sarlm(ze.sardm,ze.sem)
print(FC.test)
```

Only the results associated with the reverse distance matrix are presented here, because this matrix is the one with the strongest explanatory character (the lowest AICs) and whose economic interpretation is the most intuitive. As the employment zones have various sizes, contiguousness or nearest neighbours may have unexpected effects. The endogenous matrix may, by construction, trigger a bias in estimators. The results on the choice of model nevertheless remain consistent, regardless of the neighbourhood matrix selected.

Here we expect a negative relationship between the unemployment rate and the labour force participation rate, but a positive one for the percentage of low-skilled workers and young workers. The unemployment halo is less prominent in dynamic zones in terms of employment. Less educated people and young people are deemed to be more affected by unemployment. The zones of high industrial employment are *a priori* more affected by unemployment (reaction of employment to economic conditions and of factories closing down). On the contrary, as public jobs are more stable, the percentage of public employment should be negatively correlated to the unemployment rate. Let us remember that this model is designed to illustrate spatial econometric techniques, and no economic conclusion can be drawn from it.

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

Table 6.2 – Determinants of the unemployment rate by employment zone, based on an inverse spatial distance matrix

Note: All models are estimated with an inverse spatial distance matrix (with a threshold of 100 km). Standard deviations are shown in brackets. For tests, the p-value is indicated. Significant: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Regarding the choice of model, the following points can be derived from table 6.2.

— Elhorst's sequential approach (shown in 6.3.2) would result in adopting an SDM model (column 4). It has the lowest AIC (960). All spatial autocorrelation tests implemented from

OLS model residuals are rejected (column 1). Similarly, the common factor hypothesis in the SDM model is rejected (p-value of 0.004). Several exogenous interaction effects are significantly non-zero (the percentage of non-qualified workers at the 1 % threshold). Lastly, for the model with exogenous interactions (SLX, column 6), we do not reject the hypothesis of no residual autocorrelation under the hypothesis of endogenous correlation (strong LM-Error test, p-value of 0.787).

— Selecting a SAR model (column 3) would not be advisable here. A test shows that a residual spatial autocorrelation remains present (p-value - LM residual auto test - of 0.003). The consequences are significant for interpreting the results. The "percentage of industrial employment" variable remains significant at 1 % (regardless of the neighbourhood matrix), while the negative sign may appear counter-intuitive.

— The Manski model (column 8) provides divergent results according to the neighbourhood matrix (not shown here), certainly due to the lack of identification of this model. Similarly, the SAC model (endogenous and residual correlation, column 5) estimates an endogenous correlation that is low and not significant compared to the residual autocorrelation. This result is difficult to interpret and may result from poor model specification (Le Gallo 2002).

Finally, for reasons of parsimony, the choice of a SEM model (table 6.2, column 2) or even a SDEM model (column 7) could be considered, after verifying the consistency of the results with those of the SDM model. The interpretation of this SEM model is easier but is limited to direct effects. The AIC criterion (967) is close to the SDM model, and for weight matrices of the 5 or 10 closest neighbours (table 6.3, columns 4 and 5), the common factor hypothesis is not rejected at 1 %. The divergence in results between OLS and SEM could lead to the conclusion that the SEM model specification is not accurate, *i.e.* that it suffers from an omitted variable bias. A Hausman test (LeSage and Pace 2009 p.61-63) between the OLS and SEM models is based on the null hypothesis of the validity of both models, with the SEM model being more effective. The hypothesis is not rejected at the 1 % threshold, except as concerns the weight matrix of the 2 closest neighbours (table 6.3).

Differences in results (for different neighbourhood matrices) are analyzed for SEM and SDM models. The SEM model can be interpreted as the OLS model. The marginal effect matches the model parameters. This comparison is consistent with a bias in the OLS model. As to the activity rate, the effect is overvalued by 0.09 to 0.12 point compared with the SEM model. Concerning the percentage of industrial employment, the OLS model concludes that there is a significant negative effect whereas it is considered null with the SEM model in the case of a reverse distance matrix, or lower with the other matrices. The effect of the labour force participation rate could be overestimated with a matrix of contiguity or a small number of closest neighbours. The effect of the percentage of young working-age adults appears to be underestimated with an endogenous matrix. For the SDM model (table 6.7 in the appendix to this chapter), a direct interpretation is not possible because the effects must take into account the effects of endogenous interaction. The effects of exogenous interaction vary according to the neighbourhood matrix.

The results for the SEM model are not always robust to the choice of the neighbourhood matrix, as the “percentage of industrial employment” may or may not be significant. There is no obvious choice of a neighbourhood matrix, which would favour the results obtained with an inverse spatial distance matrix, for example. The choice should not, of course, under any circumstance, be dictated by an argument of significance of the results, but instead be based on an analysis associated with the economic question.

6.5.3 Interpreting the results

For the SDM model, in order to allow an interpretation with regard to the OLS and SEM models, the direct and indirect effects are computed as described in section 6.4 (tables 6.4 and

	(1) MCO	(2) SEM Contiguity	(3) SEM 2 Neighbours	(4) SEM 5 Neighbours	(5) SEM 10 Neighbours	(6) SEM Distance	(7) SEM Neighbours
Participation rate	-0.622***	-0.518***	-0.517***	-0.530***	-0.507***	-0.498***	-0.515***
	(0.039)	(0.040)	(0.040)	(0.040)	(0.040)	(0.041)	(0.041)
% Low educated working-age adults	0.186***	0.188***	0.204***	0.185***	0.181***	0.184***	0.184***
	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)	(0.027)	(0.026)
% Working-age adults 15-30 y.o.	0.138***	0.179***	0.195***	0.201***	0.198***	0.196***	0.139***
	(0.043)	(0.045)	(0.044)	(0.045)	(0.046)	(0.045)	(0.044)
% Industrial employment	-0.062***	-0.023*	-0.027**	-0.023*	-0.024**	-0.018	-0.026**
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
% Public employment	-0.068***	-0.042**	-0.039**	-0.047***	-0.048***	-0.044***	-0.050***
	(0.019)	(0.017)	(0.017)	(0.017)	(0.017)	(0.016)	(0.016)
$\hat{\lambda}$		0.687***	0.506***	0.681***	0.763***	0.747***	0.700***
		(0.050)	(0.047)	(0.051)	(0.053)	(0.051)	(0.044)
Intercept	51.653***	41.535***	40.672***	42.166***	40.685***	39.729***	42.414***
	(3.635)	(3.681)	(3.643)	(3.639)	(3.644)	(3.685)	(3.745)
Observations	297	297	297	297	297	297	297
AIC	1072	977	996	972	973	967	995
Hausman test		0.030	0.000	0.042	0.114	0.029	0.115
Common factor test		0.002	0.001	0.040	0.035	0.004	0.000

Table 6.3 – SEM model for different neighbourhood matrices
Note: The SEM model is estimated with 6 different neighbourhood matrices. Standard deviations are shown in brackets. Significant: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

6.5). Empirical confidence intervals are found using 1 000 simulations from empirical distribution. For the direct effects, the interpretation of the SEM model can be applied. For indirect effects, only the percentage of industrial employment has a significant negative effect. These indirect effects have a greater variability, making it impossible to conclude on any effects. The SDM model highlights the particular role of the percentage of industrial employment, which alone would have an indirect (negative) effect associated with a low or zero direct (negative) effect depending on the neighbourhood matrix selected. Yet, it is difficult to understand such an outcome from an economic point of view. The SDM model can lead us to incorrectly interpret the endogenous correlation, which does not have a clear economic interpretation here. In view of these results, the SEM model could thus be favoured, on the principle of parsimony.

```
### Estimating the direct and indirect effects of the SDM model >
    impactssdm<-impacts(ze.sardm, listw=matrix, R=1000)
summary(impactssdm)
```

	(1) MCO	(2) SDM Contiguity	(3) SDM 2 Neighbours	(4) SDM 5 Neighbours	(5) SDM 10 Neighbours	(6) SDM Distance	(7) SDM Endogenous
Participation rate	-0.622	-0.509	-0.510	-0.529	-0.505	-0.490	-0.508
	[-0.700,-0.545]	[-0.588,-0.435]	[-0.589,-0.434]	[-0.611,-0.451]	[-0.583,-0.422]	[-0.574,-0.409]	[-0.588,-0.429]
% Low educated working-age adults	0.186	0.178	0.208	0.183	0.177	0.180	0.178
	[0.136,0.237]	[0.122,0.232]	[0.154,0.261]	[0.132,0.235]	[0.125,0.230]	[0.122,0.230]	[0.129,0.232]
% Working-age adults 15-30 y.o.	0.138	0.194	0.223	0.213	0.212	0.207	0.184
	[0.054,0.223]	[0.102,0.288]	[0.135,0.312]	[0.123,0.309]	[0.119,0.306]	[0.119,0.299]	[0.092,0.279]
% Industrial employment	-0.062	-0.026	-0.032	-0.027	-0.027	-0.022	-0.033
	[-0.087,-0.038]	[-0.048,-0.003]	[-0.053,-0.008]	[-0.051,-0.005]	[-0.050,-0.005]	[-0.045,0.001]	[-0.055,-0.011]
% Public employment	-0.068	-0.045	-0.048	-0.052	-0.051	-0.049	-0.052
	[-0.106,-0.030]	[-0.078,-0.010]	[-0.081,-0.011]	[-0.084,-0.017]	[-0.083,-0.018]	[-0.081,-0.014]	[-0.084,-0.019]

Table 6.4 – Direct impacts of the SDM model, for different neighbourhood matrices
Note: The SDM model is estimated with 6 different neighbourhood matrices. The empirical confidence intervals (quantiles at 2.5 % and 97.5 % of 1000 MCMC simulations) are shown in brackets.

	(1) SDM Contiguity	(2) SDM 2 Neighbours	(3) SDM 5 Neighbours	(4) SDM 10 Neighbours	(5) SDM Distance	(6) SDM Endogenous
Participation rate	-0.323	-0.200	-0.241	-0.306	-0.357	-0.351
	[-0.587,-0.091]	[-0.337,-0.068]	[-0.488,0.007]	[-0.700,0.030]	[-0.658,-0.073]	[-0.638,-0.107]
% Low educated working-age adults	-0.015	-0.059	-0.032	-0.050	-0.053	-0.079
	[-0.161,0.142]	[-0.146,0.032]	[-0.205,0.124]	[-0.291,0.158]	[-0.254,0.137]	[-0.251,0.085]
% Working-age adults 15-30 y.o.	-0.016	-0.079	-0.082	0.016	-0.023	0.047
	[-0.321,0.249]	[-0.214,0.058]	[-0.334,0.174]	[-0.321,0.390]	[-0.352,0.301]	[-0.230,0.332]
% Industrial employment	-0.130	-0.064	-0.100	-0.135	-0.136	-0.111
	[-0.208,-0.055]	[-0.105,-0.022]	[-0.170,-0.030]	[-0.244,-0.041]	[-0.229,-0.059]	[-0.187,-0.043]
% Public employment	-0.120	-0.078	-0.113	-0.098	-0.130	-0.037
	[-0.274,0.017]	[-0.140,-0.011]	[-0.257,0.031]	[-0.345,0.132]	[-0.335,0.046]	[-0.186,0.106]

Table 6.5 – Indirect impacts of the SDM model for different neighbourhood matrices
Note: The SDM model is estimated using 6 different neighbourhood matrices. Empirical confidence intervals (quantiles at 2.5 % and 97.5 % of 1000 MCMC simulations) are shown in brackets.

6.5.4 Other spatial modelling

Descriptive analysis showed the model's possible spatial heterogeneity. It would be possible to integrate and test the presence of this phenomenon, either by authorising the heteroscedastic model (*via* the *sphet* package, citepiras2010sphet), or by modelling spatial variability in the parameters or functional form of the model. This second form of heterogeneity is obtained by including geographical zone indicators in the model, using a geographical smoothing model (*via* the *McSpatial* package , which includes semi-parametric or spline spatial models) or by conducting a weighted geographical analysis.

The practical implementation procedures for geographically weighted regression is detailed in Chapter 9: "Geographically weighted regression". Here we present the results of the geographically weighted estimate of the linear model linking unemployment rate with the structural characteristics presented above.

Table 6.6 provides the minimum, maximum and quartile values of the resulting coefficients. This makes it possible to assess the variability of the coefficients, and compare these results with those of the OLSs. The use of geographical weighted regression results in coefficients that are not always of the same sign. This may lead us to question the validity of the specification. Coefficients can vary significantly, particularly for working-age adults aged 15 to 30, with the median coefficient deviating very significantly from that of OLSs.

The first step consist in collecting a table containing, for each of the estimation points (here, the centroids of the employment zones), the value of the coefficients, the value predicted by the model, the residuals and the local value of the R^2. This makes it possible in particular to map local variations in parameters. This mapping dimension is important for assessing spatial trends. We can also check whether the residuals remain auto-correlated spatially, using suitable maps and Moran tests. There is no spatial structure of residuals in this case. Distribution of spatial parameters for industrial employment and public employment (figure 6.6) emphasises regional specificities, which can make it possible to understand surprising results, for example the null (or negative) relationship between industrial employment and the unemployment rate. This negative relationship is present mainly in the southern part of France (as well as a few regions in the north), while regions in the centre and east that have undergone major industrial restructuring show a positive correlation between the unemployment rate and the proportion of industrial employment. Concerning public employment, there is a negative correlation with the unemployment rate for

	(1) MCO	(2) Min	(3) P1	(4) Median	(5) P3	(6) Max
Participation rate	-0.622	-1.492	-0.653	-0.508	-0.379	-0.133
% Low educated working-age adults	0.186	-0.116	0.081	0.188	0.250	0.607
% Working-age adults 15-30 y.o.	0.138	-0.753	-0.040	0.183	0.340	0.875
% Industrial employment	-0.062	-0.233	-0.066	-0.029	0.006	0.184
% Public employment	-0.068	-0.318	-0.098	-0.048	-0.002	0.218
Intercept	51.650	-7.485	29.940	40.440	52.310	130.500

Table 6.6 – Weighted geographical regression results

part of southern and northern France, while the correlation is positive in Brittany, for example. Our model includes a limited number of variables, and the effect of certain regional peculiarities (industrial restructuring, employment supply characteristics, etc.) could thus be wrongly captured by our explanatory variables, a classic source of endogeneity bias. It is also possible that behaviours are heterogeneous between zones of employment. In any case, this analysis should spur us to change our model, by including other variables or spatial correlation parameters by geographical zone. We are limiting our analysis here, reiterating that the results presented are intended only to illustrate the approach for choosing and estimating a spatial model. Considering both spatial heterogeneity and correlation remains challenging.

We carried out the tests to verify non-stationarity, and therefore to assess whether weighted geographical regression is preferable to the linear model estimated by OLS (Brunsdon et al. 2002 ; Leung et al. 2000). Stationarity is rejected here regardless of the test, at the overall level and for each explanatory variable (results not shown here). Geographically weighted regression is considered to be a good exploratory method, in particular because it enables the visualization of non-stationarity phenomena. However, it has also attracted a certain amount of criticism. Wheeler et al. 2009 emphasise that the results are not robust to a high correlation between explanatory variables or the joint presence of spatial autocorrelation. In addition, as in all non-parametric statistical methods, the distance introduced (*i.e.* window selection) is not neutral. A long distance, introducing many points, will lead to coefficients that have little local variation. Conversely, a short distance will introduce a great deal of variability. The choice made may have consequences on the tests assessing the choice of the geographically weighted regression with respect to OLSs. The *GWmodel* package (Brunsdon *et al.* 2015) aims to respond to these criticisms.

Conclusion

The spatial econometric models define a consistent (and parametric) framework for modelling any type of interaction between economic agents - not only geographical zones but also products, companies or individuals. They are based on an *a priori* definition of neighbourhood relations. The main criticisms addressed to them are their lack of robustness in choosing the neighbourhood matrix and their lack of identification of the data-generating process. However, these criticisms seem exaggerated to us. As with any empirical work, that may always be questioned, choices are required in terms of specification. The strength of these models lies in their highlighting whether a "spatial" problem arises and in what form. *In contrario*, estimating a spatial econometric model as soon as "spatial" data are available is not always necessary. Methodological refinement

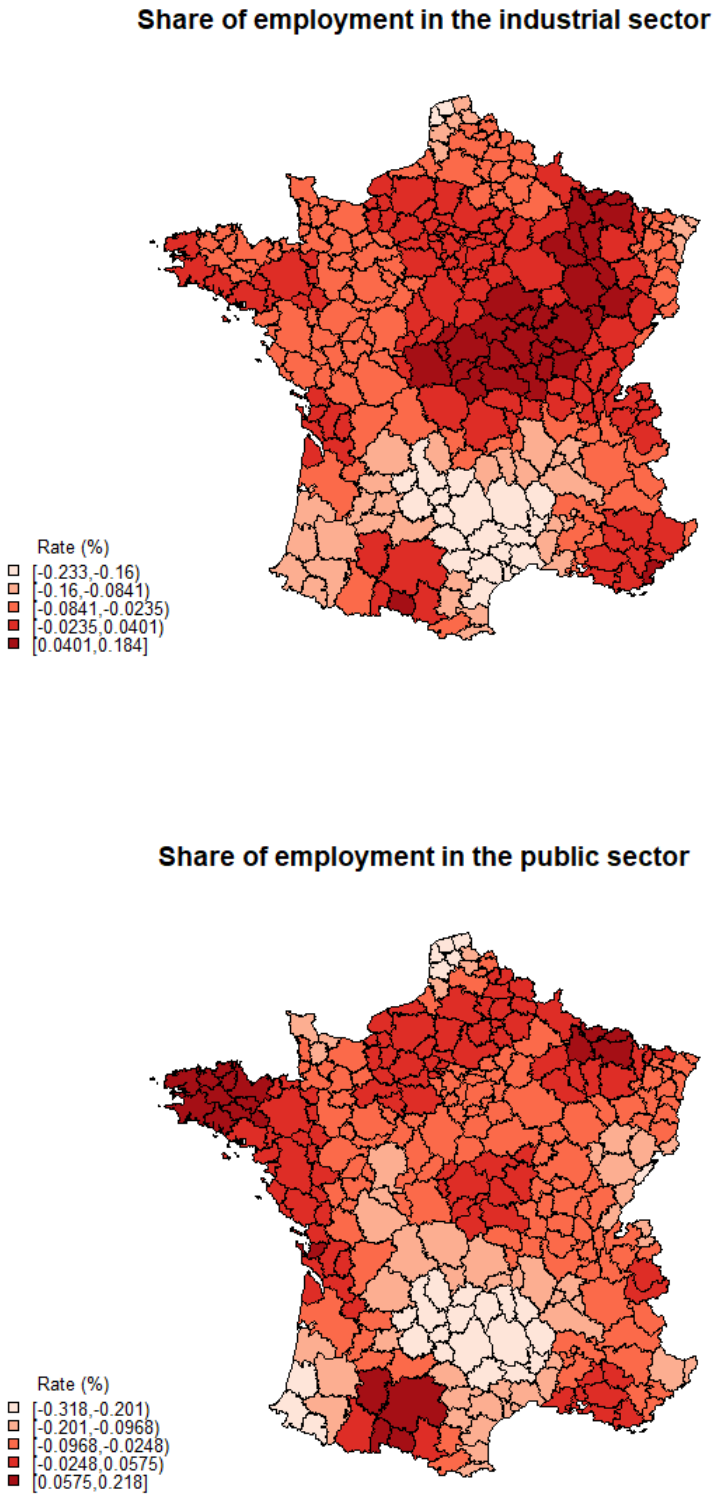

Figure 6.6 – Distribution of local parameters

must be considered with regard to the economic issue and the complexity of these new models, particularly in terms of interpretation.

It is tricky to choose between the modelling spatial correlation or heterogeneity, or even both simultaneously. In our example, taking into account the spatial correlation for modelling the localised unemployment rate seems necessary according to the statistical tests. This corrects some erroneous interpretations from the classic linear model. Here, one would opt for a spatial Durbin model (SDM) or even a model with spatially autocorrelated errors (SEM model). However, analysis of spatial heterogeneity based on weighted geographical regression also highlights that the specification should be improved, with some surprising results possibly coming from an omitted-variable bias and poor consideration of the spatial heterogeneity of labour markets. This uncertainty about the choice of the model should lead us to remain cautious about the interpretation of the direct and indirect effects of the SDM model. Moreover, it is not because the model is more complicated that it solves the problem of endogeneity of explanatory variables or the direction of causality between model variables. No causal interpretation is possible here.

The theoretical issues at stake with these methods, and in particular the links between spatial correlation and heterogeneity, are not fully controlled. The spatial econometrics models allow spatial or agent relationships to be taken into account, which is often preferable to doing nothing. Geographically weighted regression and geographical smoothing allow, in addition to descriptive approaches, defining large homogeneous regional clusters and complementary analyses to regional failure tests. Nevertheless, estimating these models requires comprehensive data. In general, therefore, they are not suitable for survey data.

Appendices

Appendix 1: Additional R codes

Creation of an endogenous neighbourhood matrix, based on commuter travel

```
## Reading the SAS file, commuting data library flows ("sas7bdat") flux<-
    read.sas7bdat("flux.sas7bdat") ## Numbering of zeo zones<-unique(flux
    [,1]) zed<-unique(flux[,1]) lig<-c(rep(1:297)) col<-c(rep(1:297)) dzeo
    <-data.frame(zeo,lig) dzed<-data.frame(zed,col) flux$zeo<-flux$
    ZEMPL2010_RESID flux$zed<-flux$ZEMPL2010_TRAV flux<-merge(flux,dzeo,by
    ="zeo") flux<-merge(flux,dzed,by="zed")
## Construction of the link weight matrix<-matrix(0,nrow=297,ncol=297)
               for (i in 1:297) {   for (j in 1:297)
                 {ze<-flux$IPONDI[flux$lig==i & flux$col==j]
                                 if(length(ze)>0)
                                 lien[i,j]<-ze
                 }
}
mig.w<-mat2listw(lien,style="W")
```

spatial linear models : additional estimates

```
### SAC Model
ze.sac<-sacsarlm(model, data=donnees_ze, matrix)
summary(ze.sac)
```

```
### SLX model
ze.slx<-lmSLX(model, data=donnees_ze, matrix)
summary(ze.slx)

### SDEM model
ze.sdem<-errorsarlm(model, data=donnees_ze, matrix, etype="emixed")
summary(ze.sdem)

### Manski model
ze.manski<-sacsarlm(model, data=donnees_ze, matrix, type="sacmixed")
summary(ze.manski)
```

Appendix 2: SDM model for different neighbourhood matrices

	(1) SDM Contiguity	(2) SDM 2 Neighbours	(3) SDM 5 Neighbours	(4) SDM 10 Neighbours	(5) SDM Distance	(6) SDM Endogenous
Participation rate	-0.486***	-0.485***	-0.513***	-0.494***	-0.472***	-0.485***
	(0.042)	(0.042)	(0.041)	(0.041)	(0.042)	(0.042)
% Low educated working-age adults	0.180***	0.215***	0.186***	0.179***	0.182***	0.184***
	(0.027)	(0.028)	(0.028)	(0.027)	(0.027)	(0.028)
% Working-age adults,15-30 y.o.	0.196***	0.232***	0.219***	0.211***	0.209***	0.181***
	(0.047)	(0.046)	(0.047)	(0.048)	(0.046)	(0.047)
% Industrial employment	-0.016	-0.024**	-0.020*	-0.022*	-0.015	-0.026**
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
% Public employment	-0.037**	-0.038**	-0.044***	-0.048***	-0.042**	-0.050*
	(0.017)	(0.017)	(0.017)	(0.017)	(0.016)	(0.016)
$\hat{\rho}$	0.601***	0.448***	0.606***	0.647***	0.629***	0.609***
	(0.057)	(0.050)	(0.057)	(0.068)	(0.064)	(0.051)
$\hat{\theta}$, Participation rate	0.153**	0.094	0.209***	0.207**	0.157*	0.149**
	(0.075)	(0.057)	(0.072)	(0.087)	(0.083)	(0.075)
$\hat{\theta}$, % Low educated working-age adults	-0.114***	-0.133***	-0.126***	-0.134***	-0.135***	-0.145***
	(0.040)	(0.034)	(0.041)	(0.047)	(0.045)	(0.040)
$\hat{\theta}$, % Working-age adults 15-30 y.o.	-0.124*	-0.153***	-0.167***	-0.131*	-0.140*	-0.090
	(0.069)	(0.053)	(0.065)	(0.078)	(0.072)	(0.068)
$\hat{\theta}$, % Industrial employment	-0.046**	-0.029**	-0.030	-0.035	-0.044**	-0.031*
	(0.021)	(0.015)	(0.019)	(0.022)	(0.020)	(0.018)
$\hat{\theta}$, % Public employment	-0.029	- 0.031	-0.020	- 0.005	-0.024	0.015
	(0.033)	(0.022)	(0.031)	(0.043)	(0.037)	(0.031)
Intercept	28.582***	33.848***	26.710***	24.504***	27.456***	27.662***
	(6.184)	(4.814)	(5.844)	(7.372)	(6.766)	(6.312)
Observations	297	297	297	297	297	297
AIC	968	985	970	971	960	987
TCommon factor test	0.002	0.001	0.040	0.035	0.004	0.000
LM residual auto. test	0.054	0.263	0.071	0.715	0.572	0.135

Table 6.7 – SDM Model for different neighbourhood matrices
Note: The SDM model is estimated with 6 different neighbourhood matrices. Standard deviations are shown in brackets. Significance: $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

References - Chapter 6

Abreu, Maria, Henri De Groot, and Raymond Florax (2004). « Space and growth: a survey of empirical evidence and methods ».

Aldstadt, Jared and Arthur Getis (2006). « Using AMOEBA to create a spatial weights matrix and identify spatial clusters ». *Geographical Analysis* 38.4, pp. 327–343.

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

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

Anselin, Luc and Daniel A Griffith (1988). « Do spatial effects really matter in regression analysis? » *Papers in Regional Science* 65.1, pp. 11–34.

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

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

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

Beck, Nathaniel, Kristian Skrede Gleditsch, and Kyle Beardsley (2006). « Space is more than geography: Using spatial econometrics in the study of political economy ». *International studies quarterly* 50.1, pp. 27–44.

Bhattacharjee, Arnab and Chris Jensen-Butler (2013). « Estimation of the spatial weights matrix under structural constraints ». *Regional Science and Urban Economics* 43.4, pp. 617–634.

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

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

Brunsdon, Chris, A Stewart Fotheringham, and Martin E Charlton (1996). « Geographically weighted regression: a method for exploring spatial nonstationarity ». *Geographical analysis* 28.4, pp. 281–298.

Corrado, Luisa and Bernard Fingleton (2012). « Where is the economics in spatial econometrics? » *Journal of Regional Science* 52.2, pp. 210–239.

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

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

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

Fingleton, Bernard and Julie Le Gallo (2008). « Estimating spatial models with endogenous variables, a spatial lag and spatially dependent disturbances: finite sample properties ». *Papers in Regional Science* 87.3, pp. 319–339.

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

Floch, Jean-Michel (2012a). « Détection des disparités socio-économiques. L'apport de la statistique spatiale ». *Document de travail INSEE H* 2012.

Florax, Raymond JGM, Hendrik Folmer, and Sergio J Rey (2003). « Specification searches in spatial econometrics: the relevance of Hendry's methodology ». *Regional Science and Urban Economics* 33.5, pp. 557–579.

Gibbons, Stephen and Henry G Overman (2012). « Mostly pointless spatial econometrics? » *Journal of Regional Science* 52.2, pp. 172–191.

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

Grislain-Letrémy, Céline and Arthur Katossky (2013). « Les risques industriels et le prix des logements ». *Economie et statistique* 460.1, pp. 79–106.

Harris, Richard, John Moffat, and Victoria Kravtsova (2011). « In search of 'W' ». *Spatial Economic Analysis* 6.3, pp. 249–270.

Kelejian, Harry H and Gianfranco Piras (2014). « Estimation of spatial models with endogenous weighting matrices, and an application to a demand model for cigarettes ». *Regional Science and Urban Economics* 46, pp. 140–149.

Kelejian, Harry H and Ingmar R Prucha (2007). « HAC estimation in a spatial framework ». *Journal of Econometrics* 140.1, pp. 131–154.

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

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

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

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

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

Leung, Yee, Chang-Lin Mei, and Wen-Xiu Zhang (2000). « Statistical tests for spatial nonstationarity based on the geographically weighted regression model ». *Environment and Planning A* 32.1, pp. 9–32.

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

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

Manski, Charles F (1993b). « Identification of endogenous social effects: The reflection problem ». *The review of economic studies* 60.3, pp. 531–542.

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

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

Thomas-Agnan, Christine, Thibault Laurent, and Michel Goulard (2014). « About predictions in spatial autoregressive models ».

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

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

Wheeler, D and A Páez (2009). *Geographically weighted regression. 1er MM, Getis A (eds) Handbook of applied spatial analysis*.