1

Econometrics and Spatial Dimensions

1.1. Introduction

Does a region specializing in the extraction of natural resources register slower economic growth than other regions in the long term? Does industrial diversification affect the rhythm of growth in a region? Does the presence of a large company in an isolated region have a positive influence on the pay levels, compared to the presence of small- and medium-sized companies? Does the distance from highway access affect the value of a commercial/industrial/residential terrain? Does the presence of a public transport system affect the price of property? All these are interesting and relevant questions in regional science, but the answers to these are difficult to obtain without using appropriate tools. In any case, statistical modeling (econometric model) is inevitable in obtaining elements of these answers.

What is econometrics anyway? It is a domain of study that concerns the application of methods of statistical mathematics and statistical tools with the goal of inferring and testing theories using empirical measurements (data). Economic theory postulates hypotheses that allow the creation of propositions regarding the relations between various economic variables or indicators. However, these propositions are qualitative in nature and provide no information on the intensity of the links that they concern. The role of econometrics is to test these theories and provide numbered estimations of these relations. To

summarize, econometrics, it is the statistical branch of economics: it seeks to quantify the relations between variables using statistical models.

For some, the creation of models is not satisfactory in that they do not take into account the entirety of the complex relations of reality. However, this is precisely one of the goals of models: to formulate in a simple manner the relations that we wish to formalize and analyze. Social phenomena are often complex and the human mind cannot process them in their totality. Thus, the model can then be used to create a summary of reality, allowing us to study it in part. This particular form obviously does not consider all the characteristics of reality, but only those that appear to be linked to the object of the study and that are particularly important for the researcher. A model that is adapted to a certain study often becomes inadequate when the object of the study changes, even if this study concerns the same phenomenon.

We refer to a model in the sense of the mathematical formulation, designed to approximately reproduce the reality of a phenomenon, with the goal of reproducing its function. This simplification aims to facilitate the understanding of complex phenomena, as well as to predict certain behaviors using statistical inference. Mathematical models are, generally, used as part of a hypothetico-deductive process. One class of model is particularly useful in econometrics: these are statistical models. In these models, the question mainly revolves around the variability of a given phenomenon, the origin of which we are trying to understand (dependent variable) by relating it to other variables that we assume to be explicative (or causal) of the phenomenon in question.

Therefore, an econometric model involves the development of a statistical model to evaluate and test theories and relations and guide the evaluation of public policies[1]. Simply put, an econometric model

1 Readers interested in an introduction to econometric models are invited to consult the introduction book to econometrics by Wooldridge [WOO 00], which is an excellent reference for researchers interested in econometrics and statistics.

formalizes the link between a variable of interest, written as y, as being dependent on a set of independent or explicative variables, written as $x_1, x_2, \ldots, x_K$, where K represents the total number of explicative variables (equation [1.1]). These explicative variables are then suspected as being at the origin of the variability of the dependent or endogenous variable:

$$y = f(x_1, x_2, \ldots, x_K) \qquad [1.1]$$

We still need to be able to propose a form for the relation that links the variables, which means defining the form of the function $f(\cdot)$. We then talk of the choice of functional form. This choice must be made in accordance with the theoretical foundation of the phenomena that we are looking to explain. The researcher thus explicitly hypothesizes on the manner in which the variables are linked together. The researcher is said to be proposing a data generating process (DGP). He/she postulates a relation that links the selected variables without necessarily being sure that the postulated form is right. In fact, the validity of the statistical model relies largely on the DGP postulated. Thus, the estimated effects of the independent variables on the determination of the dependent variables arise largely from the postulated relation, which reinfirce the importance of the choice of the functional form. It is important to note that the functional form (or the type of relation) is not necessarily known with certitude during empirical analysis and that, as a result, the DGP is postulated: it is the researcher who defines the form of the relations as a function of the *a priori* theoretical forms and the subject of interest.

Obviously, since all of the variables, which influence the behavior during the study, and the form of the relation are not always known, it is a common practice to include, in the statistical model, a term that captures this omission. The error of specification is usually designated by the term ϵ. Some basic assumptions are made on the behavior of the "residual" term (or error term). Violating these basic assumptions can lead to a variety of consequences, starting from imprecision in the

measurement of variance, to bias (bad measurement) of the searched for effect.

The simplest econometric statistical model is the one which linearly links a dependent variable to a set of interdependent variables equation [1.2]. This relation is usually referred to as multiple linear regression. In the case of a single explicative variable, we talk of simple linear regression. The simple linear regression can be likened to the study of correlation[2]. The linear regression model assumes that the dependent variable (y) is linked, linearly in the parameter, β_k, to the K ($k = 1, 2, ..., K$) number of independent variables (x_k):

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \epsilon \qquad [1.2]$$

The linear regression model allows us not only to know whether an explicative variable x_k is statistically linked to the dependent variable ($\beta_k \neq 0$), but also to check if the two variables vary in the same direction ($\beta_k > 0$) or in opposite directions ($\beta_k < 0$). It also allows us to answer the question: "by how much does the variable of interest (explained variable) change when the independent variable (dependent variable) is modified?". Herein also lies a large part of the goal of regression analysis: to study or simulate the effect of changes or movements of the independent variable on the behavior of the dependent variable (partial analysis). Therefore, the statistical model is a tool that allows us to empirically test certain hypotheses certain hypotheses as well as making inference from the results obtained.

The validity of the estimated parameters, and as a result, the validity of the statistical relation, as well as of the hypotheses tests from the model, rely on certain assumptions regarding the behavior of the error term. Thus, before going further into the analysis of the results of the econometric model it is strongly recommended to check if the following assumptions are respected:

2 In fact, the link between correlation and the analysis of simple linear regression comes from the fact that the determination coefficient of the regression (R^2) is simply the square of the correlation coefficient between the variable y and x ($R^2 = \rho^2$).

– the expectation of error terms is zero: the assumed model is "true" on average:

$$E(\epsilon) = 0; \quad [1.3]$$

– the variance of the disturbances is constant for each individual: disturbance homoskedasticity assumption:

$$E(\epsilon^2) = \sigma_\epsilon^2 \quad \forall\, i = 1, \ldots, N; \quad [1.4]$$

– the disturbances of the model are independent (non-correlated) among themselves: the variable of interest is not influenced, or structured, by any other variables than the ones retained:

$$E(\epsilon_i \epsilon_j) = 0 \quad \forall\, i \neq j. \quad [1.5]$$

The first assumption is, by definition, globally respected when the model is estimated by the method of ordinary least squares (OLS). However, nothing indicates that, locally, this property is applicable: the errors can be positive (negative) on average for high (low) values of the dependent variable. This behavior usually marks a form of nonlinearity in the relation[3]. Certain simple approaches allow us to take into account the nonlinearity of the relation: the transformation of variables (logarithm, square root, etc.), the introduction of quadratic forms (x, x^2, x^3, etc.), the introduction of dummy variables and so on and so forth.

The second assumption concerns the calculation of the variance of the disturbances and the influence of the variance of the estimator of parameter β. Indeed, the application of common statistical tests largely relies on the estimated variance and when this value is not minimal, the measurement of the variance of parameter β is not correct and the application of classical hypothesis tests is not appropriate. It is then necessary to correct the problem of heteroskedasticity of the variance of the disturbances. The procedures to correct for the presence of heteroskedasticity are relatively simple and well documented.

3 Or even a form of correlation between the errors.

The third assumption is more important: if it is violated, it can invalidate the results obtained. Depending on the form of the structure between the observations, it can have an influence on the estimation of the variance of parameters or even on the value of the estimated parameters. This latter consequence is heavier since it potentially invalidates all of the conclusions taken from the results obtained. Once again, to ensure an accurate interpretation of the results, the researcher must correct the problem of the correlation between the error terms. Here the procedures to correct for correlation among the error terms are more complex and largely depend on the type of data considered.

1.2. The types of data

The models used are largely linked to the structure and the characteristics of the data available for the analysis. However, the violation of one or several assumptions on the error terms is equally a function of the type of data used. Without a loss in generality, it is possible to identify three types of data: cross-sectional data, time series data and spatio-temporal data. The importance of the spatial dimension comes out particularly in the cross-sectional and spatio-temporal data.

The first essential step when working with a quantitative approach is to identify the type of data available to make the analyses. Not only do these data have particular characteristics in terms of violating the assumptions about the structure of the error terms, but they also influence the type of model that must be used. The type of model depends largely on the characterization of the dependent variables. Specific models are drawn for dummy variables (logit or probit models), for positive discrete (count) data (Poisson or negative binomial models), for truncated data (Heckman or Tobit models), etc. For the most part, the current demonstration will be focused on the models adapted to the case where the dependent variable is continuous (linear regression model).

1.2.1. *Cross-sectional data*

Cross-sectional data rely on a large number of observations (individuals, firms, countries, etc.) at a given time period. Database are usually defined as a file containing characteristic information from a set of observations: in a sense it is a picture giving the portrait of individuals at a fixed date. It is common practice to introduce some subindices to mark the individual observations. This subindex is written as i and the total number of observations is usually designated by N: $i = 1, 2, \ldots, N$.

For this type of data, the sources of the variation are interobservations, i.e. between the observations. It is then possible that the variation of the dependent variable is linked to some characteristics that are unique to the individuals. In the case where we cannot identify the majority of the factors that influence the variation of the dependent variable, we are faced with a problem of non-homogeneous variance, or heteroskedasticity problem. This behavior violates the second assumption of the behavior of the error terms. The linear regression model must then be corrected so that the estimated variance respects the base assumption so that the usual tests have the correct interpretation.

The tests for the detection of heteroskedasticity that are the best known are certainly those by Breusch and Pagan [BRE 79] and White [WHI 80]. The former suggests verifying if there is a significant statistical relation between the error terms squared (an estimation of the variance) and the independent variables of the model. In the case where this relation proves to be significant, we say that the variance is not homogeneous and depends on certain values of the independent variables. The second test is based on a similar approach. The White test suggests regressing the error terms squared for the whole set of the independent variables of the model as well as the crossed terms and quadratic terms of the variables. This addition of the quadratic and crossed terms allows us to consider a certain form of nonlinearity in the explanation of the variance. As for the previous case, the tests aim to verify the existence of a significant relation between the variance of

the model and some independent variables or more complex terms, in which we must reject the homogeneity hypothesis of the variance.

This type of data is largely used in microeconomics and in all the related domains. The spatial data are cross-sectional data but incorporating another particularity: the error terms can be correlated among themselves in space since they share common localization characteristics. This behavior is then in violation of the third assumption, linked to the independence between the error terms. This is the heart and foundation of spatial econometrics (we will come back to it a bit later).

1.2.2. *Time series*

Time series rely on the accumulation of information, over time, of a given individual (a firm, an employee, a country, etc.). It is a continuous acquisition of information on the characteristics of an individual over time. Thus, it is quite common for the values of the observations to be dependent over time. As before, these series call upon the use of a subindex, marked t. The size of the database is given by the number of periods available to conduct the analysis, $T : t = 1, 2, \ldots, T$.

With this type of data the variation studied is intra-observation, i.e. over time, but for a unique observation. This type of data is likely to reveal a correlation between the error terms over time and thus be in violation of the third base assumption on the behavior of error terms. We then talk of temporal autocorrelation. In this case, the parameters obtained can be biased and the conclusions that we draw from the model can be wrong. The problems of temporal correlation between the error terms have been known for several centuries.

The most commonly used test to detect such a phenomenon is the Durbin and Watson statistic [DUR 50]. This test is inspired by a measurement of the correlation between the value of the residuals taken at a period, t, and one taken at the previous period, $t - 1$. It aims to verify that the correlation is statistically significant, in which case

we are in the presence of temporal (or serial) autocorrelation. Another simple test consists of regressing the values of the residuals of the model at the period t for the value of the previous period, $t-1$, and look to determine if the parameter associated with the time-lagged variable of the residuals is significant[4]. The correction methods are also largely documented and usually available in most software.

Time series can also bring additional complications such as a changing variance (increasing or decreasing) over time. The problem of non-homogeneous variance over time is in violation of the second assumption for the behavior of error terms and the modeling methods therefore become more complex.

This type of data is especially used in macroeconomics and related domains: the indicators of a spatial entity are followed for a certain number of periods. The data regarding the market indices of a company or of a bond also represent good examples of time series.

As we will see later on, the approach for the modeling of spatial data is largely inspired by models in time series. In fact, there exists an important parallel between the problems encountered in the analysis of time series (or temporal data) and the problems encountered in the analysis of spatial data. We will come back to this in Chapter 4.

1.2.3. *Spatio-temporal data*

There are also data that possess the two characteristics: individuals that are observed over time. We then talk of spatio-temporal data. Without loss of generality, there exist two types of spatio-temporal data: panel data (or longitudinal data) and the cross-section pooled over time. In the first case, these are the same individuals that are observed at each (or nearly) time period, while in the second case, these are different individuals that are observed in each of the periods. The distinction is small, but real and important (we will come back to it in Chapter 5). In both cases, the notation relies on the introduction of

4 This regression is usually estimated without a constant (or y-intercept).

two sub-indices: an index identifying the individual observation, i, and an index identifying the time period at which each of these observations are collected t. In the case of the panel, the subindices i are the same in each of the periods, while in the case of cross-sectional data pooled over time, these indices are different in each of the periods.

For this type of data, several problems are likely to arise: persistence of the behaviors over time, a non-homogeneous variance between the individuals and a correlation of the responses in space and time. The problems are potentially very important because the information contained in this data is a lot richer. This type of data is currently increasingly popular, notably because it enables not only the studying of variations between individuals and across time, but also the evolution of given individuals over time. This is certainly the data that provides the most information. Nevertheless, the introduction of the spatial dimension in this type of data is relatively new.

These types of data have recently captured a particular amount of attention and they are currently the object of numerous theoretical advances. Several pieces of software now allow the accurate modeling of this type of data. Spatio-temporal data is also likely to reveal several problems that invalidate the base postulations with regard to the behavior of error terms. There can exist not only a spatial correlation between the error terms but also a serial correlation. Moreover, the variance can depend on the relative situation in space, just like the behavior of the independent variable. Therefore, the richness of the source of the variation can result in several problems with the assumptions on the behavior of the residuals of the model, and that the use of appropriate models to take into account these phenomena is essential.

In summary, no matter what type of data considered, it is vital to verify that all three basic assumptions for the behavior of residuals are respected if we want to be sure of an accurate interpretation of the results. However, these postulations are largely linked to the type of data that the researcher is using.

1.3. Spatial econometrics

Why spatial econometrics? The simplest answer is that the spatial dimension of the mobilized data must be taken into account. As we saw previously, this type of data is susceptible to not respecting the assumptions relative to the disturbances of the linear regression model which, as a result, could invalidate the conclusions that could be made from the analysis. Thus, it is important to adopt a model that takes account of the correlation that could exist between the error terms. This correlation between the error terms can come from different sources (we will come back to this in Chapter 4). The use of statistical tools requires that the researcher be able to verify the assumptions formulated regarding the error terms of the models retained.

To summarize, we use spatial econometrics because we are working with data that possess information on the location of the observations and this location is an additional source of variation. Therefore, it is necessary to use quantitative tools that take into account the characteristics unique to each of the observations as well as their location. The location can hide several pieces of information: a grouped (or localized) heterogeneity, or even spilling or spillover effects, or the effects of externalities. Therefore, spatial econometric models aim to take into account these characteristics.

Ignoring the sources of variation can cause bias in the quantitative analyses since the basic assumptions concerning the behavior of the error terms are no longer respected. It is therefore essential, regardless of the type of data used, to use the appropriate modeling tool. This is the case when we are working with discrete or binary variables, with truncated data, counting data, etc. This is also the case for spatial data.

Obviously, there are several types of spatialized data. The most common case consists of geographical regions that describe towns, regions (states), countries, etc. For example, two neighboring regions, that share a common border, with a similar economic structure have high chances of sharing economic climates. Or even, a economic crash that is exogenous in the first region would influence the economic

conditions in the second region. We then talk of spatial spillover effect (we will return to this in Chapter 4).

The spatial data can also rely on the observation of points. We then talk of spatial microdata: consumers, companies, residences, crimes, etc. This data is then defined by its precise geographical coordinates. We do not talk of the spatial relations based on a shared common boundary, but rather of the distance to be defined that separates the observations.

An example of a study that uses spatial microdata is the choice of location of companies. Since the companies export their production that can be used as input by another company and since they seek to minimize costs associated with shipping merchandise, it is highly likely that a company will locate close to the source of its main input or even the market that it serves. The localization decision of the company will in turn influence the localization decision of another firm that uses the production of the first firm as its principal input in its transformation process, and so on and so forth. The location decision can depend not only on the production process, but also on the location decisions of the other firms. Therefore, there exists a form of spatial dependence in the process of the localization of firms.

We can also cite the case of the price determination process of real estate as a typical example where space plays a crucial role. An old saying states that three factors influence real-estate price: the location, the location and the location. The view, local amenities and other spatial factors are likely to influence the price of a building. However, since these factors are fixed in space, it is possible that the spatial distribution of the values of the price of the real estate be very spatially structured: a panoramic view onto a lake or the proximity of a highly industrialized zone do not necessarily have the same effect on the value of the residences. Since these spatial amenities are shared between several residences, it is highly likely that the process of the creation of the value of real estate contains a non-negligible spatial component.

From these simple examples, it is possible to remark that spatial relations are potentially important in determining the processes that we

wish to study. Therefore, it is essential, if space plays a role in the decisions/reactions of the variables under study, to take them into account in analyses. It is for all these reasons that it becomes essential to take an interest in spatial econometrics.

The main particularity of spatial data is that the source of variation is richer. Not only does a spatial database contain information on the heterogeneity (the characteristics) of behavior of the individuals, but it also provides information on the relative location of observations in space. This relative localization usually enables the identification of the fact that two relatively close (spatially or even socially) individuals are likely to have behaviors that resemble each other's than if we were comparing with the case where individuals that are relatively far from each other. Therefore, spatial analysis aims to include this other source of variation in the quantitative analyses and is largely caused not only by the progress made in the domain of quantitative geography, but also by recent developments in econometrics.

Space introduces a relatively complex form of the influences that can occur between the realizations of a random variable. This characteristic is fundamental and differentiates the spatial data from non-spatial (aspatial) data, usually labeled as a cross-sectional data base. Spatial data is, in a sense, a natural extension of the cross-section data since they include additional information that is likely to bring another explanation regarding the source of the variation of the data.

1.3.1. *A picture is worth a thousand words*

To properly visualize the presentation of the previous examples, we will complete these cases with an image allowing a better understanding of the importance and the form of the spatial links between the observations.

By taking a simple example based on a variable y, for four observations (points) 1, 2, 3 and 4, whose realizations are given as y_1, y_2, y_3 and y_4 (Figure 1.1), respectively, the spatiality implies that the variables are potentially all linked together. In other words, the value of

y_1 can depend in part on the values taken by the other observations of the variable y (y_2, y_3 and y_4). In the same manner, the values taken by y_2 can depend on the other values taken by the other observations of variable y, and so on and so forth.

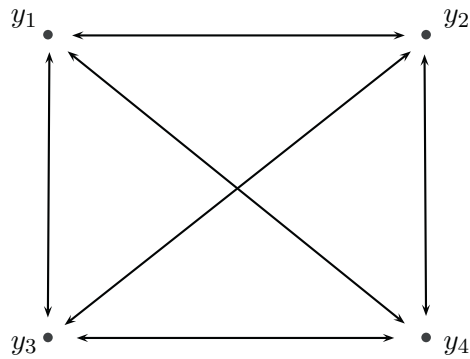

Figure 1.1. *Representation of the spatial relations between variables for four observations*

In this case, the variation of the variable for observation 1, y_1, in turn influences the response of the other observation which, in turn, influences the response of observation 1 and so on and so forth. LeSage and Pace [LES 09] qualify these effects of retroaction as indirect effects, while Abreu *et al.* [ABR 04] speak of induced effects. The decomposition of the marginal effects then takes on a much different form from the classical way. The responses all being linked to each other, the movement of a variable for a given observation indubitably causes a movement in the same variable for the other observations.

This particularity also marks the complexity of the spatial relations in the quantitative analysis that must be taken into account when the observations are spatialized. This type of relation must be integrated in the statistical models since it influences the DGP. Therefore, the challenge is to mathematically incorporate this effect.

It goes without saying that one of the central elements in spatial analysis and in spatial econometrics is the specification and the structuring of the spatial links between the variables. This

formalization of the possible spatial links is usually done through the construction of a spatial weights matrix (Chapter 2). As we will see throughout all the chapters, the spatial weights matrix plays a fundamental role in spatial econometrics. Therefore, it is essential to properly define the construction of this matrix before going any further into the presentation of spatial descriptive statistics (Chapter 3) or even in statistical modeling (Chapters 4 and 5).

1.3.2. *The structure of the databases of spatial microdata*

The structure of file of spatial data is not different from the structure of a conventional file of non-geolocalized data: a flat file in which a line represents an observation for which the detail of its description relies on a set of columns, each describing observable characteristics (variables).

An observation can be, for example, an individual characterized by age, academic degree level, income level, etc. In the same manner, a line can represent a real estate transaction whose characteristics, given by the detail of the columns, would enable the identification of the number of pieces, the area of the terrain (or the lot), the age of the building, etc.

If the structure of the database in table format is more or less the same, the possibility of locating the observations, in relative terms, makes all the difference. By resorting to the exact geographical coordinates (longitude, latitude or Cartesian coordinates) of the observation, it is possible to relativize its position in relation to the other observations. This particularity gives the researcher the possibility of taking into account possible relations that could intervene between a given observation (an individual or a region) and its neighbors.

The role of localization is twofold. First, the location gives rise, for an observation, to the construction of new characteristics linked to the description of the surroundings, notably the proximity to particular infrastructures or services. Second, localization identifies the neighboring observations of a given observation. In both cases, geolocalization allows us to incorporate a set of relations relying on spatial distances. We will return to this in Chapter 2.

This book only briefly covers the possibilities of generating new variables that express the relative distancing of an observation from centers of interest. This domain is of great interest to users of geographical information systems (GISs), as well as geographers. In such a context, geographical information allows us to make the links between that which appears at one location and the characterization of this location. GISs contextualize events as a function of the spatial description of the surroundings, as well as visually processing the data by notably producing a set of maps that can describe a given situation. While they do play an important role in the processing of spatialized information and are increasingly used, we do not go into detail on the function of these tools. Any readers interested in becoming more familiar with this tool can consult the work by Longley *et al.* [LON 01].

1.4. History of spatial econometrics

Spatial econometrics is a relatively new branch and its influence has only really been felt in the last two decades. Two reasons explain the attraction for this new branch of statistical analysis in economics.

The first reason is the increasing development of statistical models and estimators that integrate the spatial dimensions as well as the establishment of their theoretical properties. The second reason is linked to the importance of technological progress: the appearance of high-performance computers and the variety of specialized software helping with the processing of geolocalized data. Like several domains in which calculation capacities are used extensively, the popularity of the field has not truly been able to take off since the accessibility of various numerical techniques has become public.

In a recent article, Griffith [GRI 13] retraces the history of the developments of spatial analysis and spatial econometrics. He notably stresses the importance of the first works that highlighted the hypothesis according to which spatialized data behaves in a particular way. Among these authors are famous statisticians such as Student, Yule, Stephan, Fischer and Yates. All have noted a certain tendency

toward the grouping of the values measured in space. The notion of spatial dependence was then slowly starting to appear. However, it was only later that particular attention began being paid to the consideration of spatial effects in data modeling.

In fact, spatial phenomena have long captured attention without being explicitly considered in quantitative analysis. The example that stands out the most is certainly the case in London, in 1854, where John Snow, a doctor, identified the source of a cholera epidemic: by mapping all of the reported cases, the doctor managed to identify the well from which the people were getting their water as the source of the propagation. Thus, it was possible to stop the progression of the epidemic by blocking the well. However, this story remains controversial.

Formally, the first statistic test developed to detect the presence of spatial links between a given variable goes back to the works by Moran, 60 years ago [MOR 48, MOR 50]. Following this, Geary [GEA 54] proposed another one with a detection measurement of the spatial schema based on a different similarity index. Thus, the term of spatial autocorrelation makes an appearance[5].

Despite these developments from the 1950s, the possible link between spatialized variables was only formally defined in 1970 by Tobler who pronounced the first law of geography. This stipulates that all phenomena are spatially linked together, but that the phenomena that are the closest are the most strongly linked [TOB 70].

Thus, the importance of the spatial dimension reveals the "problem" of spatial correlation of variables [CLI 69]. This concept is formally defined by Anselin and Bera [ANS 98] as the coincidence of value similarity with locational similarity.

---

5 The term of autocorrelation refers to an existing correlation between a given variable and the value of the same variable for a given neighboring space. We talk of spatial autocorrelation when the notion of neighboring is applied to spatial demarcation and temporal autocorrelation when the dimension of the neighboring is applied to time periods.

It was only a decade later that the term of spatial econometrics appears following works by Paelinck and Klassen [PAE 79]. Anselin retraces the timeline of 30 years of spatial econometrics in an article published in *Papers in Regional Science* [ANS 10]. Several articles propose a census of the main fields of application in spatial econometrics [ARB 08, ARB 10], a presentation of the new developments [ARB 10, ELH 10] or even an opening toward future perspectives [ANS 07, ANS 09, PIN 10].

Formally, it is only during the 1990s that spatial autocorrelation became a subject that is really considered in the literature [ANS 98, GRI 92, HEP 00]. Getis [GET 09, p. 299] is of the opinion that the concept of spatial autocorrelation is now a unavoidable: “no other concept in empirical spatial research is as central as spatial autocorrelation”. Currently, the ramifications of this subject are such that it is impossible to ignore this problem when working with geolocalized data [ANS 07].

In the last few years, the development of estimation routines and specialized software has largely facilitated the use of spatial econometrics methods [ANS 92b, ANS 06, BIV 06, LES 99]. The appearance of two pieces of software freely available for download on the internet (GeoDa and R) has also favored a certain democratization of the methods of spatial analysis. Reference works are now increasingly common [ANS 00, ANS 04, CRE 93, CRE 10, LES 09, MUR 04, PAE 09] while the techniques and the models are increasingly diverse.

Moreover, the creation of the Spatial Econometric Association (SEA) in 2007 can be seen as a sign of the increasing popularity of spatial econometrics, in particular, and spatial analysis, in general. In a text retracing the history of the evolution of works in spatial econometrics since the creation of the SEA, Arbia [ARB 11] suggests that the main developments in the field relied on the appearance of the notion of spatial autocorrelation following the works by Cliff and Ord [CLI 69] (see also [CLI 73]). The 40 years since the appearance of this

notion are underlined in a special publication of the review *Geographical Analysis* in 2009[6].

The field's popularity can be noted, notably by the number of articles published in relation to spatial econometrics. Judging by the strong increase in the number of works on the subject from the research conducted on the search engine *Scopus*, the fields of spatial analysis and spatial econometrics truly took off around the end of the 1990s and the beginning of the 2000s (Figure 1.2). In fact, in the five years alone covering the period of 2007–2012, more that 60% of all the publications with articles containing the key words "spatial analysis" and more than 75% of all the publications containing the key words "spatial econometrics" were listed. In the same period, a particular emphasis was placed on the development of models using spatial panel data (Figure 1.3).

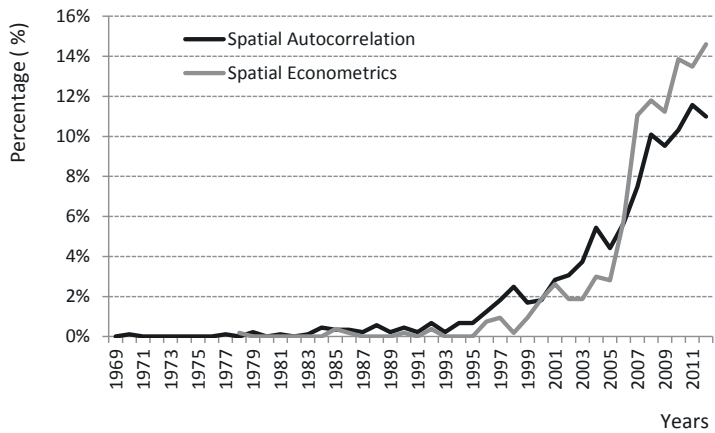

Figure 1.2. *Percentage of articles published in the most important journals, key words: spatial econometrics, spatial autocorrelation 1969–2011*

Currently, several reviews and scientific journals have been focused on several applications and developments of spatial econometrics and spatial statistics. Articles on the applications and theoretical

6 A set of articles is the object of the special publication in volume 41 (4).

developments are published in reviews of regional science (Tables 1.1, 1.2 and 1.3), of which most active are: *Geographical Analysis*, *Regional Science and Urban Economics*, *Papers in Regional Science*, *Journal of Regional Science*, *Spatial Economic Analysis*, *Economic Letters* and *Annals of Regional Science*. A particularly active journal in the domain of spatial econometrics since 2007, has been, without a doubt, the *Journal of Econometrics*. The review *Journal of Applied Econometrics* has also published a number of studies linking theoretical developments to empirical applications.

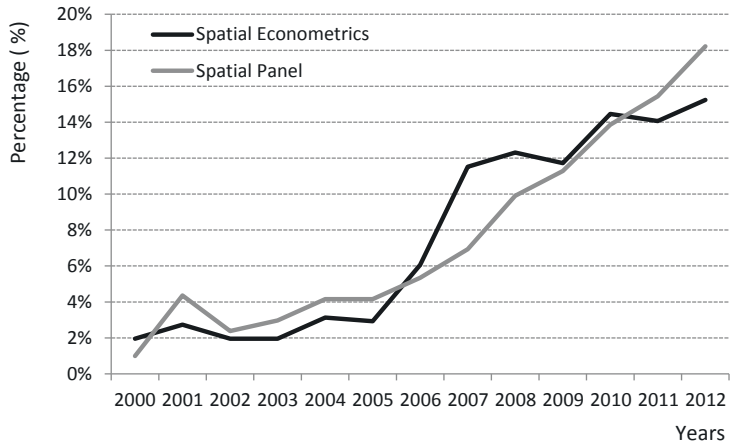

Figure 1.3. *Percentage of articles published in the most important journals, key words: spatial econometrics, spatial panel 2000–2012*

The most prolific authors in the domain are Griffith, previous editor of the review *Geographical Analysis*, Fingleton, editor of the review *Spatial Economic Analysis*, Anselin, one of the first to further spatial econometric analysis and conceiver of the GeoDa software, and also Baltagi, LeSage, Lee, Florax, Nijkamp, Kelejian, Elhorst, Getis and LeGallo (Tables 1.4, 1.5 and 1.6).

The classical way to approach the modeling of spatial autocorrelation relies on methods that attempt to control, for a certain form of spatial heterogeneity via, notably, the geographically weighted regression (GWR) – [FOT 98, FOT 02]), the locally weighted

regression (LWR) – [CLE 88, MCM 96]), and expansion of the coefficients from a previously established spatial segment [CAS 72, CAS 97], or even by an autoregressive specification of the error terms [LES 09]. We will return shortly, in slightly more detail, to the spatial autoregressive model in Chapters 4 and 5. Other models propose the isolation of the phenomenon of spatial autocorrelation generated by the omission of an important explicative variable or even from an autoregressive process on the dependent variable or on the independent variables (or explicative) [LES 09].

Rank	Review	# article	%	IF*
1	Regional Science and Urban Economics	38	7.1	1.008
2	Papers in Regional Science	32	6.0	1.430
3	Journal of Regional Science	22	4.1	2.000
4	Journal of Geographical Systems	18	3.4	1.171
5	American Journal of Agricultural Economics	15	2.8	1.169
6	Spatial Economic Analysis	14	2.6	1.200
7	Geographical Analysis	12	2.2	1.054
8	Economics Letters	11	2.1	0.447
9	Annals of Regional Science	11	2.1	1.026
10	Journal of Econometrics	11	2.1	1.349
11	Review of Regional Studies	10	1.9	0.696
12	Journal of Economic Geography	9	1.7	3.261
13	Regional Studies	7	1.3	1.187
* : *Impact factor* in 2012				

Table 1.1. *List of the most active reviews in spatial econometrics*

While there are certain geostatic approaches that allow us to consider latent variations, by developing explicative variables [DUB 12, KRI 66, TRI 67, WID 60], we will not formally deal with these models.

1.5. Conclusion

The particularity of the data requires the use of tools adapted to each of the cases: after all, quantitative analysis is only a tool that

allows dealing with research questions using a hypothetico-deductive approach. To ensure that we answer these questions properly, it is essential to use the right tools. The addition of a spatial dimension to the data in cross-sectional data makes it possible to exploit an additional source of variation. It means that relations between the observations (the points) can be made explicit, which is not possible with data in cross-sections that are not localized.

Rank	Reviews	# article	%	IF*
1	Geographical Analysis	43	4.9	1.054
2	Landscape Ecology	32	3.6	2.897
3	Journal of Biogeography	31	3.5	4.863
4	Journal of Geographical Systems	27	3.1	1.366
5	Acta Geographica Sinica	22	2.5	n.d.
6	Regional Science and Urban Economics	22	2.5	1.228
7	International Journal of Health Geographics	21	2.4	2.200
8	International Journal of Geographical Information Science	19	2.2	1.613
9	Social Science and Medicine	18	2.0	2.733
10	Proceedings of the National Academy of Sciences of the USA	16	1.8	9.737
11	Papers in Regional Science	12	1.4	1.541
12	American Journal of Physical Anthropology	12	1.4	2.481
13	Annals of the Association of American Geographers	12	1.4	2.110
* : *Impact factor* in 2012				

Table 1.2. *List of the most active reviews in spatial autocorrelation*

This book seeks to present the tools developed in spatial econometrics so as to cover the quantitative analysis of spatial data pooled over time. The second chapter looks to go even further to formalize the possible links between spatial observations through the construction, in an exogenous way, of the spatial weights matrix, an essential input in spatial econometrics. This matrix allows the formalization of the relations of spatial proximity between the observations, but can be easily transposed to other types of distances. This makes spatial econometrics more attractive for the formalization,

for example, of the effect of social, organizational, economic and other proximities on the various economical behaviors.

Rank	Reviews	# article	%	IF*
1	Regional Science and Urban Economics	20	4.0	1.228
2	Spatial Economic Analysis	18	3.6	1.375
3	Journal of Econometrics	14	2.8	1.710
4	Papers in Regional Science	11	2.2	1.541
5	Journal of Regional Science	11	2.2	2.279
6	Economics Letters	11	2.2	0.509
7	Journal of Urban Economics	10	2.0	1.910
8	Journal of Geographical Systems	9	1.8	1.366
9	Journal of Economic Geography	8	1.6	2.600
10	Economic Modeling	7	1.4	0.557
11	Proceedings of the National Academy of Sciences of the USA	7	1.4	n.d.
12	Annals of Regional Science	6	1.2	0.901
13	Empirical Economics	6	1.2	0.614
14	China Economic Review	6	1.2	1.390
15	Review of Regional Studies	6	1.2	0.785
* : *Impact factor* in 2012				

Table 1.3. *List of the most active reviews in spatial panels*

The third chapter is dedicated to the presentation of the main statistical tests that enable the detections of dependence patterns or patterns of spatial heterogeneity of the quantitative variables. We present, in a detailed manner, the indices of global and local detection of spatial autocorrelation at the base of a descriptive analysis of the spatial data, a step that often precedes a more advanced quantitative analysis.

The fourth chapter aims to present the different autoregressive models used in spatial econometrics to capture the spatial effects, either of dependence or of heterogeneity from the generalization of the standard model of linear regression. We establish the link that exists between the methods related to spatial data and those related to

temporal data. The econometric models developed, *a priori* in line with theories of the researchers, are presented, as are several statistical tests that allow us to select one spatial autoregressive model over another.

Rank	Authors	# article	%
1	Griffith, D.A.	30	3.4
2	Sokal, R.R.	9	1.0
3	Thomas, I.	8	0.9
4	Tiefelsdorf, M.	7	0.8
5	Le Gallo, J.	6	0.7
6	Getis, A.	6	0.7
7	Kelejian, H.H.	6	0.7
8	Anselin, L.	6	0.7
9	Oden, N.L.	6	0.7
10	Thomson, B.A.	5	0.6
11	Baltagi, B.H.	5	0.6
12	Khamis, F.G.	5	0.6
13	Elhorst, J.P.	5	0.6
14	Paez, A.	5	0.6
15	Novak, R.J.	5	0.6
16	Jacob, B.G.	5	0.6
17	Netrdova, P.	5	0.6
18	Martellosio, F.	5	0.6
19	Kockelman, K.M	4	0.5
20	Lu, Y.	4	0.5
Total		137	15.5
For the period 1978–2012			

Table 1.4. *List of the most active authors in spatial autocorrelation*

The fifth chapter deals with some applications of spatial models in the case where spatial microdata are gathered in a continuous manner over time without the individual observations necessarily being repeated (pooled cross-section data). Several case figures resemble this type of data collection, although currently few theoretical developments have been made in relation to this type of data. Most of

the developments relate to data in cross-sectional, or even spatial panel data. A presentation on the construction of spatio-temporal weights matrices enables the use of the outlines developed in Chapter 4, while taking into account the two dimensions of the data: spatial and temporal.

Rank	Authors	# article	%
1	Fingleton, B.	16	3.0
2	Anselin, L.	13	2.4
3	Pfaffermayr, M.	8	1.5
4	Egger, P.	7	1.3
5	LeSage, J.P.	7	1.3
6	Florax, R.J.G.M.	7	1.3
7	Griffith, D.A.	6	1.1
8	Lacombe, D.J.	6	1.1
9	Kosfeld, R.	6	1.1
10	Nijkamp, P.	6	1.1
11	Le Gallo, J.	5	0.9
12	Mur, J.	5	0.9
13	Piras, G.	5	0.9
14	Eckey, H.F.	5	0.9
15	Angulo, A.	5	0.9
16	Baltagi, B.H.	5	0.9
17	Pace, R.K.	5	0.9
18	Elhorst, J.P.	5	0.9
19	Turck, M.	3	0.6
20	Lewis, D.J.	3	0.6
Total		128	24.0
For the period 1978–2012			

Table 1.5. *List of the most active authors in spatial econometrics*

One of the main objectives of this book is especially to present a way in which to spatially link the observations among themselves and thus verify and test the presence of (spatial) links or spatial correlation (or autocorrelation) between the variables as suggested by the first law

of geography. This particularity of spatial relations can modify the statistical approaches normally used. The geographical coordinates of the observations allow us, in this context, to take into account the possible links that can exist between the observations, which is impossible with databases that do not contain any information on the geographical location. The particularity of spatial links relies on these links. The multidirectionality of the links stipulates that a given variable can influence the behavior of another neighboring variable, and that this very neighboring variable in turn influences the behavior (or the realization) of the variable considered.

Rank	Authors	# article	%
1	Baltagi, B.H.	14	2.8
2	Lee, L.F.	9	1.8
3	Yu, J.	9	1.8
4	Pfaffermayr, M.	8	1.6
5	Fingleton, B.	7	1.4
6	Crowder, K.	7	1.4
7	South, S.J.	7	1.4
8	Tosetti, E.	6	1.2
9	Moscone, F.	6	1.2
10	Laurisden, J.	5	1.0
11	Griffith, D.A.	4	0.8
12	Papalia, R.B.	4	0.8
13	Kelejian, H.H.	4	0.8
14	Egger, P.	4	0.8
15	Nijkamp, P.	4	0.8
16	Pirotte, A.	4	0.8
17	Rodriguez-Pose, A.	4	0.8
18	Tselios, V.	4	0.8
19	Kockelman, K.M.	4	0.8
20	Millimet, D.L.	4	0.8
Total		117	23.2
For the period 2000–2012			

Table 1.6. *List of the most active authors in panel spatial econometrics*

A second objective of the book is to provide an introduction to applied spatial econometric models. For this reason, we have deliberately decided to not go into detail in the calculations of the estimators and the mathematical proofs of the various properties of the estimators and statistical tests. We propose, instead, an approach based on the intuitive presentation of the main tools and models as well as a presentation where the behavior of the various statistical tools is numerically studied using programs, presented in the appendix. These programs can simulate spatial data according to the process that the reader is willing to provide them.