This book puts special emphasis on spatial data compilation and the structuring of connections between the observations. Descriptive analysis methods of spatial data are presented in order to identify and measure the global and local spatial autocorrelation. The authors then move on to incorporate this spatial component into spatial autoregressive models. These models allow us to control the problem of spatial autocorrelation among residuals of the linear statistical model, thereby contravening one of the basic hypotheses of the ordinary least squares approach.

This book can be used as a reference for those studying towards a bachelor’s or master’s degree in regional science or economic geography, looking to work with geolocalized (micro) data, but without possessing advanced statistical theoretical basics. The authors also address the application of the spatial analysis methods in the context where spatial data are pooled over time (spatio-temporal data), focusing on the recent developments in the field.

Jean Dubé is Professor in regional development at Laval University, Canada.
Diëgo Legros is a lecturer in economics and management at the University of Burgundy, France.
Spatial Econometrics Using Microdata
To the memory of Gilles Dubé.
For Mélanie, Karine, Philippe, Vincent and Mathieu.
Spatial Econometrics
Using Microdata

Jean Dubé
Diègo Legros
Contents

ACKNOWLEDGMENTS ........................................... ix

PREFACE ....................................................... xi

CHAPTER 1. ECONOMETRICS AND SPATIAL DIMENSIONS 1
  1.1. Introduction ........................................... 1
  1.2. The types of data ....................................... 6
    1.2.1. Cross-sectional data ............................... 7
    1.2.2. Time series ........................................ 8
    1.2.3. Spatio-temporal data ............................... 9
  1.3. Spatial econometrics ................................... 11
    1.3.1. A picture is worth a thousand words ............... 13
    1.3.2. The structure of the databases of spatial microdata . 15
  1.4. History of spatial econometrics ....................... 16
  1.5. Conclusion ............................................ 21

CHAPTER 2. STRUCTURING SPATIAL RELATIONS ............ 29
  2.1. Introduction ........................................... 29
  2.2. The spatial representation of data .................... 30
  2.3. The distance matrix ................................... 34
  2.4. Spatial weights matrices ............................... 37
    2.4.1. Connectivity relations ............................ 40
    2.4.2. Relations of inverse distance .................... 42
2.4.3. Relations based on the inverse (or negative) exponential .................................................. 45
2.4.4. Relations based on Gaussian transformation .......................... 47
2.4.5. The other spatial relation .......................................................... 47
2.4.6. One choice in particular? .......................................................... 48
2.4.7. To start .......................................................... 49
2.5. Standardization of the spatial weights matrix ......................... 50
2.6. Some examples .......................................................... 51
2.7. Advantages/disadvantages of micro-data .............................. 55
2.8. Conclusion .......................................................... 56

CHAPTER 3. SPATIAL AUTOCORRELATION ........................................ 59
3.1. Introduction .......................................................... 59
3.2. Statistics of global spatial autocorrelation ............................ 65
  3.2.1. Moran’s $I$ statistic ...................................................... 68
  3.2.2. Another way of testing significance ................................ 72
  3.2.3. Advantages of Moran’s $I$ statistic in modeling .................. 74
  3.2.4. Moran’s $I$ for determining the optimal form of $W$ .......... 75
3.3. Local spatial autocorrelation .............................................. 77
  3.3.1. The LISA indices ...................................................... 79
3.4. Some numerical examples of the detection tests .................... 86
3.5. Conclusion .......................................................... 89

CHAPTER 4. SPATIAL ECONOMETRIC MODELS ................................... 93
4.1. Introduction .......................................................... 93
4.2. Linear regression models .............................................. 95
  4.2.1. The different multiple linear regression model types .......... 99
4.3. Link between spatial and temporal models ......................... 102
  4.3.1. Temporal autoregressive models .............................. 103
  4.3.2. Spatial autoregressive models ................................ 110
4.4. Spatial autocorrelation sources ..................................... 115
  4.4.1. Spatial externalities .............................................. 117
  4.4.2. Spillover effect .............................................. 119
  4.4.3. Omission of variables or spatial heterogeneity ......... 123
4.4.4. Mixed effects .............................................. 127
4.5. Statistical tests ............................................. 129
  4.5.1. LM tests in spatial econometrics ................. 134
4.6. Conclusion .................................................. 140

CHAPTER 5. SPATIO-TEMPORAL MODELING .................. 145
  5.1. Introduction .............................................. 145
  5.2. The impact of the two dimensions on the structure of the
       links: structuring of spatio-temporal links .......... 148
  5.3. Spatial representation of spatio-temporal data .... 150
  5.4. Graphic representation of the spatial data generating
       processes pooled over time ............................ 154
  5.5. Impacts on the shape of the weights matrix ....... 159
  5.6. The structuring of temporal links: a temporal weights
       matrix ...................................................... 162
  5.7. Creation of spatio-temporal weights matrices .. 167
  5.8. Applications of autocorrelation tests and of
       autoregressive models .................................. 170
  5.9. Some spatio-temporal applications .................. 172
  5.10. Conclusion ................................................ 173

CONCLUSION .................................................... 177

GLOSSARY ....................................................... 185

APPENDIX ....................................................... 189

BIBLIOGRAPHY .................................................. 215

INDEX .......................................................... 227
While producing a reference book does require a certain amount of time, it is also impossible without the support of partners. Without the help of the publisher, ISTE, it would have been impossible for us to share, on such a great scale, the fruit of our work and thoughts on spatial microdata.

Moreover, without the financial help of the Fonds de Recherche Québécois sur la Société et la Culture (FRQSC) and the Social Sciences and Humanities Research Council (SSHRC), the writing of this work would certainly not have been possible. Therefore, we thank these two financial partners.

The content of this work is largely the result of our thoughts and reflections on the processes that generate individual spatial data\(^1\) and the application of the various tests and models from the data available.

We thank the individuals who helped, whether closely or from afar, in the writing of this work by providing comments on some or all of the chapters: Nicolas Devaux (student in regional development), Cédric Brunelle (Professor at Memorial University), Sotirios Thanos (Reseracher Associate at University College London) and

\(^1\) Our first works largely focus on the data of real estate transactions: a data collection process that is neither strictly spatial, nor strictly temporal (see Chapter 5).
Philippe Trempe (masters student in regional development). Without the invaluable help of these people, the writing of this book would certainly have taken much longer and would have been far more difficult. Their comments helped us orientate the book towards an approach that would be more understandable by an audience that did not necessarily have a lot of experience in statistics.
P.1. Introduction

Before even bringing up the main subject, it would seem important to define the breadth that we wish to give this book. The title itself is quite evocative: it is an introduction to spatial econometrics when data consist of individual spatial units. The stress is on microdata: observations that are points on a geographical projection rather than geometrical forms that describe the limits (whatever they may be) of a geographical zone. Therefore, we propose to cover the methods of detection and descriptive spatial analysis, and spatial and spatio-temporal modeling.

In no case do we wish this work to substitute important references in the domain such as Anselin [ANS 88], Anselin and Florax [ANS 95], LeSage [LES 99], or even the more recent reference in this domain: LeSage and Pace [LES 09]. We consider these references to be essential for anyone wishing to become invested in this domain.

The objective of the book is to make a link between existing quantitative approaches (correlation analysis, bivariated analysis and linear regression) and the manner in which we can generalize these approaches to cases where the available data for analysis have a spatial dimension. While equations are presented, our approach is largely based on the description of the intuition behind each of the equations. The mathematical language is vital in statistical and quantitative
analyses. However, for many people, the acquisition of the knowledge necessary for a proper reading and understanding of the equations is often off-putting. For this reason, we try to establish the links between the intuition of the equations and the mathematical formalizations properly. In our opinion, too few introductory works place importance on this structure, which is nevertheless the cornerstone of quantitative analysis. After all, the goal of the quantitative approach is to provide a set of powerful tools that allow us to isolate some of the effects that we are looking to identify. However, the amplitude of these effects depends on the type of tool used to measure them.

The originality of the approach is, in our opinion, fourfold. First, the book presents simple fictional examples. These examples allow the readers to follow, for small samples, the detail of the calculations, for each of the steps of the construction of weighting matrices and descriptive statistics. The reader is also able to replicate the calculations in simple programs such as Excel, to make sure he/she understands all of the steps properly. In our opinion, this step allows non-specialist readers to integrate the particularities of the equations, the calculations and the spatial data.

Second, this book aims to make the link between summation writing (see double summation) of statistics (or models) and matrix writing. Many people will have difficulties matching the transition from one to the other. In this work, we present for some spatial indices the two writings, stressing the transition from one writing to the other. The understanding of matrix writing is important since it is more compact than summation writing and makes the mathematical expressions containing double summation, such as detection indices of spatial correlation patterns, easier to read; this is particularly useful in the construction of statistics used for spatial detection of local patterns. The use of matrix calculations and simple examples allow the reader to generalize the calculations to greater datasets, helping their understanding of spatial econometrics. The matrix form also makes the calculations directly transposable into specialized software (such as MatLab and Mata (Stata)) allowing us to carry out calculations without having to use previously written programs, at least for the construction
of the spatial weighting matrices and for the calculation of spatial concentration indices. The presentation of matrix calculations step by step allows us to properly compute the calculation steps.

Third, in the appendix this work suggests programs that allow the simulation of spatial and spatio-temporal microdata. The programs then allow the transposing of the presentations of the chapters onto cases where the reality is known in advance. This approach, close to the Monte Carlo experiment, can be beneficial for some readers who would want to examine the behavior of test statistics as well as the behavior of estimators in some well-defined contexts. The advantages of this approach by simulation are numerous:

– it allows the intuitive establishment of the properties of statistical tools rather than a formal mathematical proof;
– it provides a better understanding of the data generating processes (DGP) and establishes links with the application of statistical models;
– it offers the possibility of testing the impact of omitting one dimension in particular (spatial or temporal) on the estimations and the results;
– it gives the reader the occasion to put into practice his/her own experiences, with some minor modifications.

Finally, the greatest particularity of this book is certainly the stress placed on the use of spatial microdata. Most of the works and applications in spatial econometrics rely on aggregate spatial data. This representation thus assumes that each observation takes the form of a polygon (a geometric shape) representing fixed limits of the geographical boundaries surrounding, for example, a country, a region, a town or a neighborhood. The data then represent an aggregate statistic of individual observations (average, median, proportion) rather than the detail of each of the individual observations. In our opinion, the applications relying on microdata are the future for not only putting into practice of spatial econometric methods, but also for a better understanding of several phenomena. Spatial microdata allow us to
avoid the classical problem of the ecological error \(^2\) [ROB 50] as well as directly replying to several critics saying that spatial aggregate data does not allow capturing some details that are only observable at a microscale. Moreover, while not exempt from the modifiable area unit problem (MAUP)\(^3\) [ARB 01, OPE 79], they do at least present the advantage of explicitly allowing for the possibility of testing the effect of spatial aggregation on the results of the analyses.

Thus, this book acts as an intermediatory for non-econometricians and non-statisticians to transition toward reference books in spatial econometrics. Therefore, the book is not a work of theoretical econometrics based on formal mathematical proofs\(^4\), but is rather an introductory document for spatial econometrics applied to microdata.

**P.2. Who is this work aimed at?**

Nevertheless, reading this book assumes a minimal amount of knowledge in statistics and econometrics. It does not require any particular knowledge of geographical information systems (GIS). Even if the work presents programs that allow for the simulation of data in the appendixes, it requires no particular experience or particular aptitudes in programming.

More particularly, this book is addressed especially to master’s and PhD students in the domains linked to regional sciences and economic geography. As the domain of regional sciences is rather large and multidisciplinary, we want to provide some context to those who would like to get into spatial quantitative analysis and go a bit further

---

2 The ecological error problem comes from the transposition of conclusions made with aggregate spatial units to individual spatial units that make up the spatial aggregation.

3 The concept of MAUP was proposed by Openshaw and Taylor in 1979 to designate the influence of spatial cutting (scale and zonage effects) on the results of statistical processing or modeling.

4 Any reader interested in a more formal presentation of spatial econometrics is invited to consult the recent work by LeSage and Pace (2009) [LES 09] that is considered by some researchers as a reference that marks a “big step forward” in “for spatial econometrics” [ELH 10, p. 9].
on this adventure. In our opinion, the application of statistics and statistical models can no longer be done without understanding the spatial reality of the observations. The spatial aspect provides a wealth of information that needs to be considered during quantitative empirical analyses.

The book is also aimed at undergraduate and postgraduate students in economics who wish to introduce the spatial dimension into their analyses. We believe that this book provides excellent context before formally dealing with theoretical aspects of econometrics aiming to develop the estimators, show the proofs of convergence as well develop the detection tests according to the classical approaches (likelihood ratio (LR) test, Lagrange multiplier (LM) test and Wald tests).

We also aim to reach researchers who are not econometricians or statisticians, but wish to learn a bit about the logic and the methods that allow the detection of the presence of spatial autocorrelation as well as the methods for the correction of eventual problems occurring in the presence of autocorrelation.

P.3. Structure of the book

The book is split into six chapters that follow a precise logic. Chapter 1 proposes an introduction to spatial analysis related to disaggregated or individual data (spatial microdata). Particular attention is placed on the structure of spatial databases as well as their particularities. It shows why it is essential to take account of the spatial dimension in econometrics if the researcher has data that is geolocalized; it presents a brief history of the development of the branch of spatial econometrics since its formation.

Chapter 2 is definitely the central piece of the work and spatial econometrics. It serves as an opening for the other chapters, which use weights matrices in their calculations. Therefore, it is crucial and it is the reason for which particular emphasis is placed on it with many examples. A fictional example is developed and taken up again in Chapter 3 to demonstrate the calculation of the detection indices of the spatial autocorrelation patterns.
Chapter 3 presents the most commonly used measurements to detect the presence of spatial patterns in the distribution of a given variable. These measurements prove to be particularly crucial to verify the assumption of the absence of spatial correlation between the residuals or error terms of the regression model. The presence of a spatial autocorrelation violates one of the assumptions that ensures the consistency of the estimator of the ordinary least squares (OLS) and can modify the conclusions coming from the statistical model. The detection of such a spatial pattern requires the correction of the regression model and the use of spatial and spatio-temporal regression models. Obviously, the detection indices can also be used as descriptive tools and this chapter is largely based on this fact.

Chapters 4 and 5 present the autoregressive models used in spatial econometrics. The spatial autoregressive models (Chapter 4) can easily be transposed to spatio-temporal applications (Chapter 5) by developing an adapted weights matrix to the analyzed reality. A particular emphasis is put on the intuition behind the use of one type of model rather than another: this is the fundamental idea behind the DGP. In function of the postulated model, the consequences of the spatial relation detected between the residuals of the regression model can be more or less important, going from an imprecision in the calculation of the estimated variance, to a bias in the estimations of the parameters. The appendixes linked to Chapters 4 (spatial modeling) and 5 (spatio-temporal modeling) are based on the simulation of a given DGP and the estimation of autoregressive models from the weights matrices built previously (see Chapter 2).

Finally, the Conclusion is proposed, underlying the central role of the construction of the spatial weights matrix in spatial econometrics and the different possible paths allowing the transposition of existing techniques and methods to different definitions of the “distance”.

We hope that this overview of the foundations of spatial econometrics will spike the interest of certain students and researchers, and encourage them to use spatial econometric modeling with the goal of getting as much as possible out of their databases and inspire some of them to propose new original approaches that will complete the
current methods developed. After all, the development of spatial methods notably allows the integration of notions of spatial proximity (and others). This aspect is particularly crucial for certain theoretical schools of thought linked to regional science and new geographical economics (NGE), largely inspired by the works of Krugman [FUJ 04, KRU 91a, KRU 91b, KRU 98], recipient of the 2008 Nobel prize in economics [BEH 09].

Figure P.1. Links between the chapters

Jean DUBÉ
and Diège LEGROS
August 2014
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:

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

[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 reinforce 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 \( \epsilon \). 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, \(\beta_k\), to the \(K\) \((k = 1, 2, ..., K)\) number of independent variables \(x_k\):

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

[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:

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

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

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

– 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:

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

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 $\beta$. 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 $\beta$ 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).